An investigation of the highly conserved long non-coding RNA *CEROX1*

KENNY ROBERTS

ORIEL COLLEGE

DEPARTMENT OF PHYSIOLOGY, ANATOMY AND GENETICS

UNIVERSITY OF OXFORD

MICHAELMAS TERM 2018

A thesis submitted for the degree of

Doctor of Philosophy

ABSTRACT

Amid growing catalogues of long non-coding RNAs (lncRNAs), experimentally characterised examples are few, especially for the recently identified subpopulation that are enriched in the cytoplasm. Amongst these are lncRNAs whose activity is proposed to be mediated by the competitive endogenous RNA (ceRNA) mechanism. De-repression of miRNA targets results in concerted gene expression changes following ceRNA modulation in numerous *in vitro* models, but evidence of *in vivo* physiological relevance is lacking.

In light of this, the novel human lncRNA *CEROX1* was subjected to a multifaceted characterisation to investigate its potential as a physiologically important candidate ceRNA. *In vitro* molecular biology and *in silico* bioinformatic approaches were used to investigate the expression, conservation, activity, and regulation of *CEROX1*. It was found that *CEROX1* is both extremely highly expressed and highly conserved, being expressed and enriched in the central nervous system of eutherian mammals. In particular, *CEROX1* is highly enriched in cells of the oligodendrocyte lineage, and putatively associated with the process of oligodendrocyte differentiation.

Modulation of *CEROX1* and its murine orthologue *Cerox1 in vitro* results in concerted regulation of complex I of the mitochondrial electron transport chain; these are therefore the first lncRNAs identified with such a role. In line with its putative activity in regulating energy metabolism in the developing CNS, *CEROX1* was found to be upregulated *in vitro* in response to calcium signalling, which contributes substantially to cell signalling pathways between neurons and glia including oligodendrocytes.

In addition to the conserved mitochondria-regulating role of *CEROX1*, evidence of translation of a 234 residue open reading frame was uncovered *in vitro* and *in vivo*, suggesting recent evolution of the locus towards a protogene state.

This work highly supports *CEROX1* as a promising *in vivo* ceRNA candidate, although future study is required to pursue its role in regulating energy metabolism in the CNS.

SUPERVISORS

Dr Tamara M Sirey

MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh

Professor Chris P Ponting

MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh

Professor Peter L Oliver

MRC Harwell Institute; Department of Physiology, Anatomy and Genetics, University of Oxford

Professor Dame Kay E Davies

Department of Physiology, Anatomy and Genetics, University of Oxford

All formerly of the MRC Functional Genomics Unit, University of Oxford

EXAMINERS

Professor Matthew J A Wood

Department of Physiology, Anatomy and Genetics, University of Oxford

Dr Keith W Vance

Department of Biology and Biochemistry, University of Bath

Examined viva voce Monday 25th February 2019.

DECLARATION OF AUTHORSHIP

The work presented henceforth in this thesis has been carried out during the course of this doctoral degree by me, the author, Kenny Roberts, unless otherwise specified or acknowledged. A list of other work that is incorporated is given below:

Characterisation of mouse *Cerox1* was carried out primarily by Dr Tamara Sirey, with assistance from me at a time prior to my doctoral studies. This work forms only an introductory basis to the investigation of human *CEROX1* in this thesis. This work is publically available as two manuscripts detailing the parallel characterisation of mouse and human *CEROX1* transcripts: one a bioRχiv pre-print, the other a peer-reviewed publication in eLife. The latter is included at the end of this thesis, for reference.

The initial preliminary characterisation of human *CEROX1* was carried out by me, during the course of a project submitted as part of my M.Biochem. degree within the Department of Biochemistry, University of Oxford. Where data generated during that degree is illustrated in this thesis, it will be clearly demarcated. This includes: RT-qPCR analysis of a human tissue panel to assess *CEROX1* expression; RT-qPCR analysis of fractionated HEK293T cells to assess *CEROX1* sub-cellular localisation; and the initial characterisation of the *CEROX1* overexpression phenotype by RT-qPCR and spectrophotometric enzyme assays.

Experimental data or resources generated by other researchers includes: preliminary analysis of GTEx transcriptomics data by Dr Wilfried Haerty; proteomics by Dr Svenja Hester and Dr Sabrina Liberatori of the Advanced Proteomics Service, Department of Biochemistry, University of Oxford; proteomics by Dr Roman Fischer and Dr Sarah Bonham of the Target Discovery Institute, University of Oxford; microRNA binding site predictions by Prof. Ana Marques, Dr Jennifer Tan, and Dr Oscar Bedoya-Reina; Seahorse respiratory measurements by Dr Tamara Sirey; Sanger sequencing was performed by Source Bioscience (Department of Biochemistry, University of Oxford); and a custom CEROX1-ORF1 antibody was generated by AMSbio, Abingdon, Oxfordshire.

Again, where data were generated by another researcher, it will be clearly acknowledged. Where appropriate, methodological descriptions of techniques used by these researchers or services are given in Chapter 2: Materials and Methods.

Kenny Roberts

ACKNOWLEDGEMENTS

To my family, especially my parents, my grandparents, and my great-granny, thank you for your encouragement throughout the many years of my long ongoing scientific education; for tolerating much mess generated by increasingly ambitious chemistry sets; and for helping support me financially through two degrees.

Thank you to my four supervisors: to Tam Sirey and Chris Ponting for starting the process of building my scientific skills, and for giving me the chance to continue working on this interesting project; and to Kay Davies and Pete Oliver for providing additional support, encouragement, lab space, and resources as the circumstances required it.

Mattéa Finelli, I am so grateful for all the support you've given me scientifically in the lab and otherwise. In a year of early mornings and late nights in the lab, your scientific advice, technical expertise, hard-working and determined philosophy have contributed monumentally to the consolidation of my project as well as the development of my scientific skills. Not to mention the occasional moment of shared hysteria and the not-so-occasional 'Heartbreak FM' disco.

Lauren Watson, you were such a welcome addition to Team Upstairs Office, and have been one of the few constants throughout my D.Phil. amid the many winds of change. Thank you for scientific help, different perspectives, and advice, but mostly for lots of food and many laughs.

I do not think that this would have been possible without Sarah Noujaim and her fierce determination in attempting to rectify the many logistical issues that have arisen during the past few years. Your job will be simpler once I finish!

A heartfelt thank you is due to Esther Becker for making me an honorary member of the Becker lab, as well as thought-provoking scientific discussion, and advice.

I am very grateful to Mary Board for giving me the chance to come to Oxford and start my scientific career. Thank you for nurturing my skills and confidence, and for keeping me a part of the St Hilda's College biochemistry family over the past 8 years.

For their work in helping me prepare for the D.Phil. scholarship interview without which I would not be here, I must deeply thank Ana Marques, Jennifer Tan, Wilfried Haerty, Andy Bassett, and Jenna Schwarz. Thank you for seeing the best in me and helping me to show it.

Kir Hopkins and Katie McElligott: my Glee Team.

A great many people have assisted me with scientific support, including providing reagents, samples, data, protocols, or suggestions: Wilfried Haerty, Mattéa Finelli, Lisa Heather, Ana Marques, Jennifer Tan, Oscar Bedoya-Reina, Esther Becker, Lauren Watson, Maggie Wong, Laurène Vetterli, Tahnee Kennedy, James Cantley, Hannah Dugdale, Claire Smillie, Katherine Brimblecombe, Steph Cragg, Sam Nayler, Kelly Perkins, Jenna Schwarz, Ben Edwards, Matthew Williamson, Dasha Svistunova, Matthew Kerr, Julia Grasegger, Nesrine Ramadan, Christoffer Nellåker, Liam Argent, Stuart Grice, Zoltan Molnar, Jacqueline Robbins, Amanda Anderson, and my Transfer of Status examiners Matthew Wood and Richard Wade-Martins.

A special acknowledgement in this vein is due to: a) the many researchers who assisted me in finding a very elusive gorilla brain sample, notably the enthusiastic Gordon Hull, as well as Twycross Zoo, and Mark Stidworthy of the International Zoo Veterinary Group pathology lab; b) Erwin Galinski of the University of Bonn, Germany, who kindly provided me with homoectoine.

Several services have provided me with crucial samples or technical support. Thank you to: Olaf Ansorge and Carolyn Sloan of the Oxford Brain Bank, supported by the UK Medical Research Council and Alzheimer's Brain Bank UK, for providing precious human brain samples; Roman Fischer and Sarah Bonham of the Target Discovery Institute for their proteomics expertise; Andrew Jefferson of the Micron Oxford Advanced Bioimaging Unit for emergency confocal use.

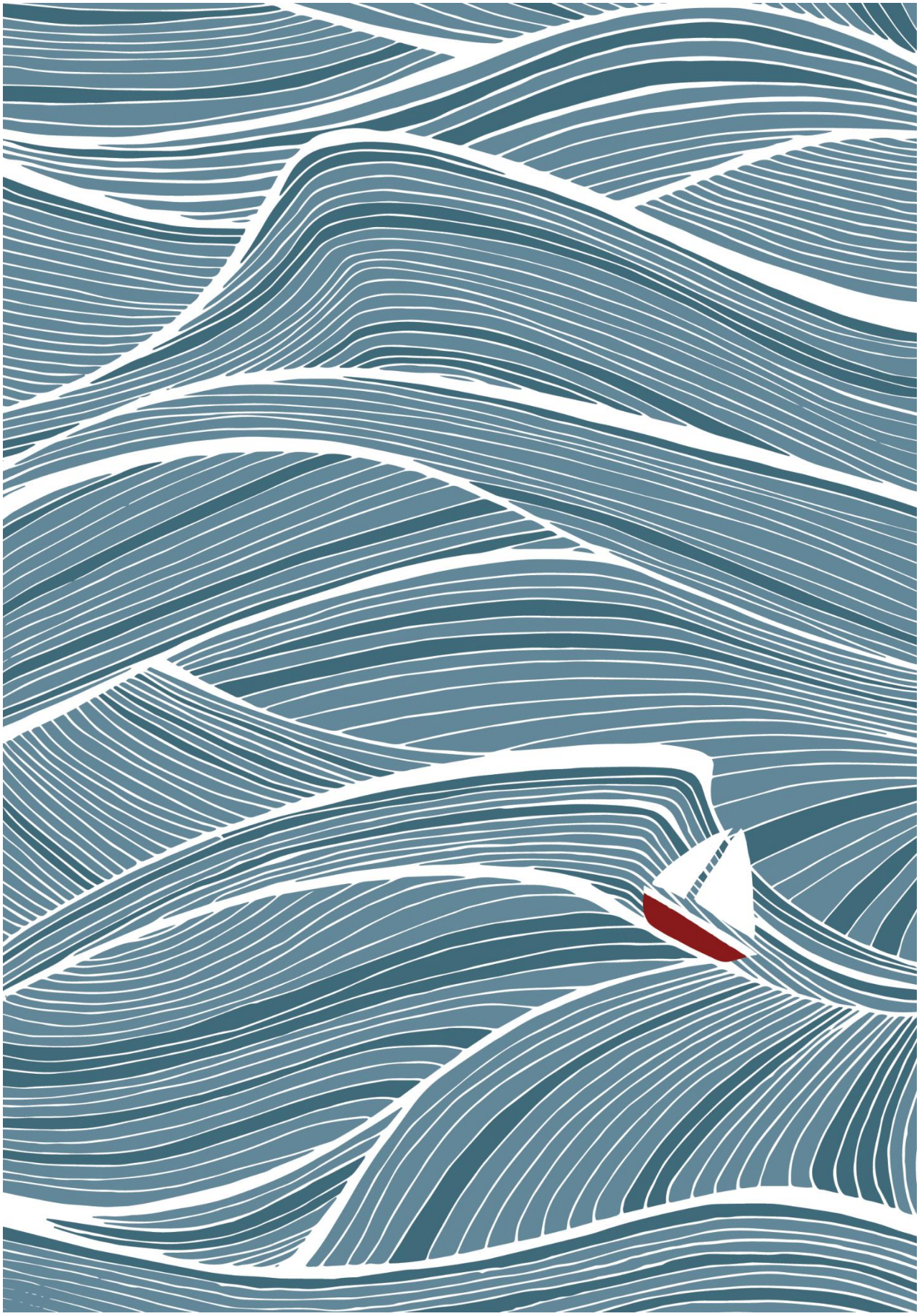
As well as scientific support, I've been very appreciative of other day-to-day help: ever-jolly IT support from Simon Middlecoate, as well as logistic support from the DPAG porters, Gary Mort, Andy Jordan, and Lee Brown, and Cathy Jones.

This D.Phil. has been a great challenge, and I'm very grateful for the many people not already mentioned who have provided counsel, distractions, confidence, encouragement, moral support, and cake both inside and outside of the lab: especially Filipa Landeiro, Jenna Schwarz, Libby Bamber, Charlotte Tibbit, Dini Shah, Ömür Yilmaz Tastan, and Emily Frape. The best thing that came out of my work in Edinburgh was meeting Gordon Tatnell, whose support helped propel me through the final throes of this project.

Also: Claudia Wehrspaun, Lee Moir, Arran Babbs, Sarah Squire, Viv Collins, Heu Chang, Megan Roper, Jethro Johnson, Charlie George, Kathy Fawcett, Mike Morgan, Tom Smith, Sebastian Luna Valero, Katie Brown, Will Pembroke, and Chris Rands.

Finally, I would like to acknowledge each of the funding sources that have enabled me to carry out this project, primarily: a PhD scholarship from the Medical Research Council and a Goodger and Schorstein research scholarship from the Medical Sciences Division. Also intermittent funding from the Department of Physiology, Anatomy and Genetics, and Oriel College.

"A smooth sea never made a skilled sailor."



© Rachel Seago Illustration; all rights reserved, and reproduced by kind permission of Rachel Seago.

TABLE OF CONTENTS

ABSTRACT ..... **i**
SUPERVISORS ..... **ii**
EXAMINERS ..... **ii**
DECLARATION OF AUTHORSHIP ..... **iii**
ACKNOWLEDGEMENTS ..... **iv**
TABLE OF CONTENTS ..... **vii**
TABLE OF FIGURES ..... **xiii**
ABBREVIATIONS ..... **xvii**
CHAPTER 1: INTRODUCTION **1**
1.1 Pervasive transcription in the human genome produces long non-coding RNAs 1
1.1.1 Non-coding RNAs are numerous and ubiquitous 1
1.1.2 Poor sequence conservation of lncRNAs as a group does not preclude functionality 2
1.1.3 The extent of the functional relevance of all lncRNAs is not clear 2
1.2 Many characterised lncRNAs play roles in co-transcriptional regulation 3
1.2.1 LncRNAs co-transcriptionally regulate gene expression via diverse interactions 3
1.2.2 At some loci, lncRNAs may be side-products of functional DNA elements 6
1.3 Cytoplasmic lncRNAs 7
1.3.1 RNA-miRNA-RNA crosstalk mediates the function of a subset of lncRNAs: ceRNAs 7
1.3.2 The widespread importance of the ceRNA mechanism remains contested 9
1.3.3 Other cytoplasmic functions of characterised lncRNAs 19
1.3.4 Multiple lncRNAs may co-participate in complex regulatory networks 23
1.3.5 Organellar lncRNAs 24
AIMS OF THIS THESIS 25
Characterisation of further individual lncRNAs and especially ceRNAs is essential 25
Aims and thesis structure 25
REFERENCES 27
CHAPTER 2: MATERIALS AND METHODS **38**
TECHNIQUES 38
2.1 Techniques for cell work 38
2.1.1 Mammalian cell culture 38
2.1.2 Cell counting 38
2.1.3 Transfection of exogenous DNA 39
2.1.4 Drug treatments 39
2.2 Techniques for human and animal tissue work 39
2.2.1 Human samples 40
2.2.2 Animal samples 40
2.2.3 Sectioning 40
2.3 Techniques for RNA extraction and experimentation 40

2.3.1 RNA extraction 41
2.3.2 cDNA synthesis 41
2.3.3 Rapid amplification of cDNA ends (RACE) 42
2.3.4 *In situ* hybridisation (ISH) 44
2.4 Polymerase chain reaction (PCR) 45
2.4.1 PCR 45
2.5 Cloning and DNA construct preparation 47
2.5.1 Restriction digestion 47
2.5.2 DNA analysis and purification 47
2.5.3 TA cloning 48
2.5.4 DNA ligation 48
2.5.5 Bacterial amplification of plasmids 48
2.5.6 Sanger sequencing 49
2.5.7 Construct preparation 49
2.6 Techniques for protein extraction and experimentation 52
2.6.1 Protein Extraction 52
2.6.2 Mitochondrial isolation 53
2.6.3 Electron transport chain complex enzyme activity assays 53
2.6.4 SDS polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting 55
2.6.5 Blue native polyacrylamide gel electrophoresis (BN-PAGE) 56
2.6.6 Densitometry 57
2.6.7 Immunoprecipitation 57
2.6.8 Immunocytochemistry 58
2.6.9 Immunohistochemistry 59
2.6.10 Antibodies 59
2.6.11 Mass spectrometry 61
2.7 Assays 62
2.7.1 Nano-Glo® luciferase assay 62
2.7.2 Seahorse extracellular flux analysis 63
2.8 Imaging 64
2.8.1 Confocal microscopy 64
2.8.2 Low-magnification imaging 64
2.9 Bioinformatics 64
2.9.1 Genome browsing 64
2.9.2 Alignments 65
2.9.3 Open reading frame annotations 65
2.9.4 MicroRNA response element predictions 65
2.9.5 Transcription factor predictions 65
2.9.6 Expression data 66
2.9.7 Statistical analysis 66
2.9.8 Other 67
2.10 Solutions 68

REFERENCES 69
CHAPTER 3: CHARACTERISATION OF *CEROX1* AS A CONSERVED LONG NON-CODING RNA 72
INTRODUCTION 72
3.1 Many lncRNAs are poorly conserved 72
3.1.1 *CEROX1* in context: *CEROX1* neighbours two protein-coding genes, *LMF1* and *SOX8* 73
3.1.2 Examination of the expression and genomic history of *CEROX1* will inform hypotheses of its functional relevance 78
RESULTS 79
3.2 Transcription at the *CEROX1* locus produces a number of isoforms 79
3.2.1 Transcriptional activity at the *CEROX1* locus is supported by epigenetic evidence 79
3.2.2 Retrotransposon insertion might have contributed to *CEROX1* evolution 82
3.2.3 At least twelve isoforms of *CEROX1* are annotated at present 83
3.2.4 Isoforms were investigated *in vitro* using rapid amplification of cDNA ends (RACE) 85
3.2.5 3' RACE supports the existence of at least two transcriptional end sites for *CEROX1* 88
3.3 Human *CEROX1* and mouse *Cerox1* demonstrate a high level of conservation 90
3.3.1 The human and mouse *CEROX1* loci are syntenic and partially conserved 90
3.3.2 The overall expression profile of human *CEROX1* mirrors that of mouse *Cerox1* 94
3.4 The expression of *CEROX1* is varied, dynamic, and conserved, to the cellular level 97
3.4.1 *CEROX1* is enriched in cells of the oligodendrocyte lineage, in both mouse and human 97
3.4.2 Even in the mostly highly expressing tissues, *CEROX1* expression is varied 100
3.4.3 *CEROX1* expression is dynamic throughout development, but tissue specificity is retained 103
3.5 Conservation of *CEROX1* is older than the divergence between human and mouse 106
3.5.1 The DNA sequence of the *CEROX1* locus is conserved throughout Eutheria 106
3.5.2 The conserved *CEROX1* locus is transcribed throughout Eutheria 109
3.5.3 Expression at the *CEROX1* locus has always been enriched in the central nervous system 115
DISCUSSION 117
REFERENCES 124
CHAPTER 4: A HIDDEN PROTEIN – IS *CEROX1* TRULY A NON-CODING RNA? 132
INTRODUCTION 132
4.1 Hiding in plain sight: not all 'lncRNAs' are entirely non-coding 132
4.1.1 Some 'non-coding' RNAs have been shown to produce microproteins from small ORFs 132
4.1.2 'Coding non-coding RNAs' illustrate that coding potential need not diminish the functional importance of non-coding RNAs 133
4.1.3 Translation is also associated with regulation of RNA stability, and possibly antigen production 136
4.1.4 Thorough case-by-case examination is needed to classify any RNA as non-coding 138
RESULTS 139
4.2 Is *CEROX1* simply a long-non coding RNA? 139
4.2.1 *CEROX1* contains a number of potential ORFs, including one of substantial length 139
4.2.2 Existing bioinformatic data inconclusively support a non-coding classification of *CEROX1* 139
4.2.3 CEROX1-ORF1 is exceptional for a lncRNA-ORF 144
4.2.4 Experimental approaches to determining protein-coding capacity *in vitro* and *in vivo* 145
4.2.5 CEROX1-ORF1 is a candidate for further investigation 148

4.3 A strategy was devised to investigate the putative CEROX1-ORFs in vitro 149
4.3.1 Detection of very small putative proteins requires additional considerations in detection by western blotting 149
4.4 At least two ORFs in *CEROX1* are subject to translation *in vitro* 150
4.4.1 *CEROX1*-ORF1 induces translation of polypeptides when overexpressed in HEK293T cells 150
4.4.2 Translation of CEROX1-ORF2, but not CEROX1-ORF3, is detectable under degradation-inhibiting conditions 152
4.4.3 Expression of CEROX1-ORF1 is visible by immunocytochemistry following overexpression 156
4.5 Does *CEROX1* encode a protein endogenously *in vitro* and *in vivo*? 157
4.5.1 CEROX1-ORF1 translation was not detected endogenously *in vitro* 157
4.5.2 Probing for CEROX1-ORF1 *in vivo* 160
DISCUSSION 167
REFERENCES 171
CHAPTER 5: *CEROX1* IS A PUTATIVE REGULATOR OF THE ELECTRON TRANSPORT CHAIN 175
INTRODUCTION 175
5.1 The mitochondrial electron transport chain is at the centre of eukaryotic life 175
5.1.2 The ETC is also subject to higher organisation into super-complexes 177
5.1.3 Post-transcriptional regulation of the mitochondrial ETC 177
5.2 Previous work by the Ponting lab has characterised mouse *Cerox1* as a ceRNA regulating mitochondrial activity 181
5.2.1 *Cerox1* regulates the mitochondrial electron transport chain 181
5.2.2 *Cerox1* acts by interacting with mmu-miR-488-3p 182
PRELIMINARY RESULTS 183
5.3 Initial studies of human *CEROX1* indicated that it is a conserved regulator of electron transport chain complexes 183
5.3.1 *CEROX1* upregulated the activities of complexes I and III of the human ETC *in vitro* 183
5.3.2 Preliminary gene expression data supported increased activity of complexes I and III 184
5.3.3 The molecular mechanism of complex I upregulation appeared to be conserved between mouse *Cerox1* and human *CEROX1* 186
5.3.4 *CEROX1* was hypothesised to be a putative conserved ceRNA 188
5.3.5 Difficulty in *CEROX1* knock-down impeded further phenotypic studies 189
5.3.6 Further investigation of the role of *CEROX1* and the underlying mechanism 190
RESULTS 193
5.4 *CEROX1* does not act via modulation of *SOX8* in this system 193
5.5 Human *CEROX1* contains many predicted miRNA binding sites 194
5.6 Experiments designed to further interrogate the *CEROX1* overexpression phenotype have shown it to be poorly reproducible 197
5.6.1 A *CEROX1* mutant transcript was produced to identify functionally-relevant elements 197
5.6.2 Thorough examination of experimental procedures did not identify the reason for the loss of phenotype 198
5.6.3 Native page was used as a secondary assay to confirm the alteration of phenotype 202
5.6.4 The results of a third assay of respiratory activity are more consistent with the original *CEROX1* overexpression phenotype 204
5.6.5 RNA and protein expression of complex I subunits also supports an altered phenotype 205

DISCUSSION 207
REFERENCES 213
CHAPTER 6: THE *CEROX1-SOX8* PROMOTER 219
INTRODUCTION 219
6.1 *CEROX1* is surely regulated, as well as being a putative regulator 219
6.1.1 Regulation would likely reflect *CEROX1*'s expression pattern and putative function 219
6.1.2 Combined approaches facilitate the identification of functionally relevant regulatory transcriptional mechanisms 220
6.1.3 Previous investigation of the regulation of *Sox8* produced an incomplete model 222
6.1.4 Addressing the *CEROX1* promoter 223
RESULTS 224
6.2 *CEROX1* and *SOX8* share a bidirectional promoter 224
6.2.1 Characteristics of the *CEROX1-SOX8* promoter are consistent with bidirectionality 224
6.2.2 Expression of *CEROX1* and *SOX8* is highly correlated across human tissues 224
6.3 Identification of the *CEROX1-SOX8* promoter and analysis *in silico* 226
6.3.1 The core promoter was defined so as to include all major transcriptional start sites 226
6.3.2 Bioinformatic approaches identify conserved transcription factor binding sites 226
6.4 The *CEROX1-SOX8* promoter was investigated experimentally using *in vitro* luciferase reporter assays 230
6.4.1 Luciferase assays provide a high-throughput measure of promoter-driven transcription 230
6.4.2 The putative promoter demonstrates bidirectional transcriptional activity *in vitro* 230
6.5 An initial screen of stimuli revealed several agents with putative *CEROX1-SOX8* promoter-regulating activity 233
6.5.1 Choosing an *in vitro* system for studying the *CEROX1* promoter 233
6.5.2 Two statistical approaches identified significant changes in promoter activity 233
6.5.3 Five compounds exhibited significant modulation of the *CEROX1-SOX8* promoter 235
6.6 Calcium signalling may regulate transcription of *CEROX1* 236
6.6.1 Two significant hits from the preliminary screen act on converging targets 236
DISCUSSION 239
REFERENCES 247
CHAPTER 7: DISCUSSION 252
7.1 Summary of work: a holistic approach to lncRNA investigation 252
7.2 Moving forward with *CEROX1* 255
7.2.1 The ceRNA character of *CEROX1* remains unclarified experimentally 255
7.2.2 *CEROX1* remains an excellent candidate for physiological lncRNA functionality 255
7.2.3 CEROX1-ORF1: moving beyond reasonable doubt? 257
7.3 Challenges of future ceRNA identification 258
REFERENCES 261
APPENDICES 263
SUPPLEMENTARY LITERATURE 263
A1 Supplementary literature describing *CEROX1* 263
A1.1 The long non-coding RNA *Cerox1* is a post transcriptional regulator of mitochondrial complex I catalytic activity 263

A1.2 *Cerox1* and microRNA-488-3p noncoding RNAs jointly regulate mitochondrial complex I catalytic activity 264
SUPPLEMENTARY MATERIALS AND METHODS 265
A2 Primer sequences 265
A3 Performing RACE upon human *CEROX1* was extremely challenging 269
A4 Species and loci used for analysis of *CEROX1* orthologous sequences 271
A4.1 Identification of transcribed *CEROX1* orthologues 271
A4.2 *CEROX1* orthologue tissue expression profiling 273
A5 Drugs used in screening activity of the CEROX1-SOX8 promoter 276
SUPPLEMENTARY DATA 277
A6 Tissue-specific transcriptome analyses of *CEROX1* expression 277
A6.1 *CEROX1* transcriptomic rank correlates with expression level 277
A6.2 Expression rank relative to mitochondrial genes is also tissue-specific 280
A7 Further investigation of *Cerox1* in the striatum 281
A8 Optimised western blotting permits detection of very small proteins 283
A9 CEROX1-ORF1 may have cryptic start codons 284
A10 Integrity of gorilla cortex protein 285
A11 *CEROX1* knock-down was achieved 286
SUPPLEMENTARY REFERENCES 287

TABLE OF FIGURES

CHAPTER 1: INTRODUCTION 1

Figure 1.1: The biogenesis of microRNAs and their role in gene regulation via the ceRNA mechanism. 8

Table 1.1: Examples of ceRNAs characterised largely *in vitro* and their different approaches to indicating *in vivo* relevance, part 1 15

CHAPTER 2: MATERIALS AND METHODS 38

Table 2.1: Cycling conditions used for PCR amplification of DNA targets 46

Table 2.2: Cycling and data collection conditions used for real time quantitative PCR. 46

Table 2.3: Cycling conditions used for touchdown RACE-PCR 47

Table 2.4: Features of the pCAGGS overexpression vector 49

Table 2.5: Flanking restriction enzymes used to produce tagged CEROX1-ORF expression constructs 50

Figure 2.1: Generation of an HA-tagged ORF construct by PCR-driven overlap assembly 51

Equation 2.1: Calculation of specific enzyme activity (A) using a spectrophotometric assay 54

Table 2.6: Buffers and protocols used in mitochondrial enzyme assays 54

Table 2.7: Wavelengths of absorbance readings, and molar extinction coefficients (ϵ) used in calculations of enzyme activity for ETC complexes I to IV (CI-IV) and citrate synthase (CS). 55

Table 2.8: Assay buffers used in staining ETC complex activities following native PAGE. 57

Table 2.9: Assay buffers trialled for staining of complex III activity following native PAGE. 57

Table 2.10: Primary antibodies used in western blotting (WB) and immunocytochemistry (ICC). 60

Table 2.11: Horseradish peroxidase-conjugated secondary antibodies used in western blotting. 60

Table 2.12: Fluorophore-conjugated secondary antibodies used in fluorescent RNA *in situ* hybridisation, immunohistochemistry, and immunocytochemistry 60

Equation 2.2: Calculation of fold change in promoter activity following drug treatment 63

Equation 2.3: Conversion of treatment-induced promoter activity fold change into a population Z-score 63

Equation 2.4: Calculation of the population standard deviation, σ 66

Equation 2.5: Calculation of the weighted average $\bar{c}$, and combined standard error of the mean, ϵ 67

Table 2.13: Recipes for solutions 68

CHAPTER 3: CHARACTERISATION OF *CEROX1* AS A CONSERVED LONG NON-CODING RNA 72

Figure 3.1: The genomic locus and annotation of human *CEROX1* 75

Figure 3.2: Epigenetic features of the *CEROX1* locus support its transcriptional activity 80

Figure 3.3: The 5' end of *CEROX1* contains two retrotransposed elements 82

Figure 3.4: Data from the CAGE-seq FANTOM5 project and the GTEx RNA-seq project offer differing results as to the relative activities of the transcriptional start sites of the *CEROX1* locus 84

Figure 3.5: The high GC content of *CEROX1* prompted the use of DNA denaturation-promoting additives 87

Figure 3.6: Rapid amplification of cDNA ends (RACE) confirms 3' and 5' ends of multiple *CEROX1* isoforms 89

Figure 3.7: The loci of human *CEROX1* and mouse *Cerox1* are syntenic 91

Figure 3.8: Genomic deletion in the *Cerox1* locus of some rodent species in has altered the architecture of the lncRNA 93

Figure 3.9: Expression patterns of human *CEROX1* and mouse *Cerox1* are highly similar 95

Figure 3.10: *CEROX1* expression is most elevated in the basal ganglia, where it outranks that of mitochondrial genes....96
Figure 3.11: Both human *CEROX1* and mouse *Cerox1* are enriched in oligodendrocytes....98
Figure 3.12: Occupancy of the oligodendrocyte transcription factor Olig2 around the rodent *Cerox1* loci 100
Figure 3.13: *In situ* hybridisation (ISH) indicates heterogeneity in *CEROX1* and *SOX8* expression throughout the striatum....102
Figure 3.14: Expression of *Cerox1* and *Sox8* in the developing mouse brain and liver are dynamic, and consistent with adult expression data....103
Figure 3.15: Human *CEROX1* is expressed throughout the developing human brain....105
Figure 3.16: Evolutionary conservation of the DNA sequence of, and surrounding, the *CEROX1* genomic locus....107
Figure 3.17: The synteny of *CEROX1*'s neighbouring protein-coding genes is highly conserved....108
Figure 3.18: Conservation of the *CEROX1* 3' end in Euarchontoglires....110
Figure 3.19: PLAR analyses identified *CEROX1*-orthologous transcripts in sheep (*Ovis aries*) and dog (*Canis lupus familiaris*)....112
Figure 3.20: Phylogenetic relationships of *CEROX1*-orthologous sequences in 18 eutherian mammals 114
Figure 3.21: Tissue expression profile of *CEROX1* orthologues is conserved throughout Eutheria....116
CHAPTER 4: A HIDDEN PROTEIN – IS *CEROX1* TRULY A NON-CODING RNA?.... 132
Figure 4.1: The three largest open reading frames in *CEROX1* are not well conserved....140
Figure 4.2: The CEROX1-ORF1 sequence is partially conserved within recent primate ancestors....142
Table 4.1: Estimations of the coding potential of *CEROX1* as produced by coding potential programs .143
Figure 4.3: *CEROX1* is bound by ribosomes....147
Figure 4.4: The distribution of length in RNase-protected reads following ribosome profiling of human brain tissue[81] indicates that reads ascribed to *CEROX1* are indeed associated with ribosomes....148
Figure 4.5: CEROX1-ORF1 may be translated *in vitro*. 151
Figure 4.6: Expression of CEROX1-ORF protein products following inhibition of cellular degradation pathways....153
Figure 4.7: Mass spectrometry sequencing of CEROX1-ORF1$^{C\text{-}HA}$ translation products....155
Figure 4.8: Expression of CEROX1-ORF1$^{C\text{-}HA}$ protein in HEK293T cells....156
Figure 4.9: A custom rabbit antibody successfully detects exogenous, but not endogenous, CEROX1-ORF1 *in vitr*....158
Figure 4.10: SK-N-SH neuroblastoma cells also translate exogenous CEROX1-ORF1 *in vitro*....159
Figure 4.11: Possible conserved translation of CEROX1-ORF1 *in vivo* was investigated in parallel in human and gorilla brain....162
Figure 4.12: MS/MS reveals the presence of a CEROX1-ORF1 peptide in gorilla cortex....163
Figure 4.13: Immunostaining of CEROX1-ORF1 *in vivo* reveals cell type specificity....165
Figure 4.14: Specificity of binding of the CEROX1-ORF1 antibody was not clarified by a peptide competition assay....166
CHAPTER 5: *CEROX1* IS A PUTATIVE REGULATOR OF THE ELECTRON TRANSPORT CHAIN..... 175
Figure 5.1: The mitochondrial electron transport chain is a complex system. 176
Figure 5.2: Overexpression of human *CEROX1* upregulated activities of ETC complexes I and III *in vitro* 184
Figure 5.3: Gene expression changes underlie the changes in enzymatic activities of complexes I and III 185

Figure 5.4: ETC complex-regulating activity of human *CEROX1* and mouse *Cerox1* in one another's respective cellular backgrounds suggests putative conservation of molecular mechanism 187

Figure 5.5: *CEROX1* is enriched in the cytoplasm of HEK293T cells 188

Figure 5.6: Inefficacious *CEROX1* knockdown produced no complex I or complex III phenotype 189

Figure 5.7: The hypothetical *CEROX1* regulatory model at present, and further routes of investigation. 192

Figure 5.8: *CEROX1* overexpression does not affect SOX8 expression in the HEK293T system used here 193

Figure 5.9: A dense network of putative interactions connects *CEROX1* to the transcripts encoding complex I subunits 195

Figure 5.10: Intersection of *CEROX1*-*Cerox1* and *CEROX1*-complex I shared MREs provides an approach to narrow down the investigation into causative miRNAs 196

Figure 5.11: Complex I assays designed to ascertain the functional importance of a deleted region revealed non-reproducibility of the *CEROX1* overexpression phenotype 197

Figure 5.12: *CEROX1* overexpression is potent and consistent 199

Figure 5.13: Repeated overexpression of human *CEROX1* and mouse *Cerox1* in HEK293T and N2A cells did not recapitulate the original results of these experiments 201

Figure 5.14: *CEROX1* overexpression resulted in no ETC phenotype in two other neuroblastoma human cell lines 201

Figure 5.15: The activities of four ETC complexes were assessed by BN-PAGE following *CEROX1* overexpression in HEK293T cells 203

Figure 5.16: Analysis of live HEK293 cells supports upregulation of ETC activity following overexpression of *CEROX1* 204

Figure 5.17: Altered RNA and protein changes in response to *CEROX1* overexpression 206

CHAPTER 6: THE *CEROX1-SOX8* PROMOTER 219

Figure 6.1: Expression of *CEROX1* and *SOX8* is highly correlated across human tissues 225

Figure 6.2: The *CEROX1-SOX8* promoter region used in this study 226

Figure 6.3: Conservation of predicted transcription factor binding sites within the *CEROX1-SOX8* promoter 228

Figure 6.4: The *CEROX1-SOX8* promoter shows bidirectional transcription-promoting activity *in vitro*. 231

Figure 6.5: Activity of the *CEROX1-SOX8* promoter was modulated by several exogenous agents, chiefly PMA 234

Figure 6.6: Calcium signalling mediated via PKC is capable of regulating the *CEROX1-SOX8* promoter . 237

Figure 6.7: Transcription regulation identified by the luciferase reporter model reflects steady-state gene expression changes 238

Table 6.1: Putative targets and downstream signalling of PKC, and strategies for further investigation of their relevance in the regulation of the *CEROX1-SOX8* promoter. 241

Figure 6.8: Examples of potential regulatory pathways via which calcium and PKC might act upon *CEROX1* transcription *in vitro*, part I 245

Figure 6.9: Examples of potential regulatory pathways via which calcium and PKC might act upon *CEROX1* transcription *in vitro*, part II 246

CHAPTER 7: DISCUSSION 252

Figure 7.1: A *hypothetical* model of the physiological role of *CEROX1* in oligodendrocytes 254

APPENDICES 263

Supplemental Figure S1: Primers used in various experiments throughout this project, including those for cloning, colony PCR, RT-qPCR, RACE, generation of probes for ISH, and sequencing 268

Supplemental Figure S2: Osmolytes used in *CEROX1* 5' RACE 270
Supplemental Figure S3: Species whose transcriptomes were analysed in search of *CEROX1* orthologues 272
Supplemental Figure S4: RNA-seq data sets used in tissue expression profiling of mammalian *CEROX1*-orthologous transcripts 275
Supplemental Figure S5: Drugs and hormones used in the SH-SY5Y *CEROX1-SOX8* Nluc promoter screen 276
Supplemental Figure S6: *CEROX1* expression within a given tissue is strongly correlated with expression rank in that tissue 277
Supplemental Figure S7: Expression distributions for all 53 GTEx tissues 278
Supplemental Figure S8: *CEROX1* expression within a given tissue is strongly correlated with expression rank in that tissue 279
Supplemental Figure S9: *CEROX1* is more highly expressed than a majority of complex I-associated genes in the CNS 280
Supplemental Figure S10: Investigating the *Cerox1*-striosome relationship 282
Supplemental Figure S11: Small proteins are retained and detectable during the western blotting protocol used in the experiments of Chapter 4 283
Supplemental Figure S12: Ablation of the CEROX1-ORF1 start codon does not remove all signal associated with the ORF 284
Supplemental Figure S13: High integrity protein was extracted from frozen gorilla cortex for use in IP-MS experiments 285
Supplemental Figure S14: *CEROX1* knockdown was produced using LNA™ GapmeRs 286

ABBREVIATIONS

Full forms of underlined abbreviations may also be found in this section.

aa – amino acid(s)
ADP – adenosine diphosphate
AGO – Argonaute
AMP – adenosine monophosphate
asRNA – antisense RNA
ATP – adenosine triphosphate
BCA – bicinchoninic acid
BCIG – 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal)
BLAST – basic local alignment search tool
BLAT – BLAST-like alignment tool
BN-PAGE – blue native PAGE
bp – base pair(s)
BSA – bovine serum albumin
CAGE-seq – cap analysis of gene expression sequencing
carbachol – carbamylcholine chloride
cDNA – complementary DNA
Cdr1as - cerebellar degeneration-related protein 1 antisense transcript
ceRNA – competing/competitive endogenous RNA
ceRNET – competing/competitive endogenous RNA network
CEROX1 – competitive endogenous regulator of oxidative phosphorylation 1
CEROX1-ORF – *CEROX1* open reading frame
CHAT – choline acetyltransferase
ChIP – chromatin immunoprecipitation
ChIP-seq – chromatin immunoprecipitation sequencing
CI – complex I, NADH:ubiquinone oxidoreductase
CII – complex II, succinate dehydrogenase
CIII – complex III, ubiquinol:cytochrome c oxidoreductase
CIV – complex IV, cytochrome c oxidase
CLASH – cross-linking ligation and sequencing of hybrids
CLIP – cross-linking immunoprecipitation
CNS – central nervous system
CoA – coenzyme A
CoQ – coenzyme Q, ubiquinone
CPAT – Coding Potential Assessment Tool
CPC – Coding Potential Calculator
CpG – cytosine-guanine dinucleotide
CRISPR – clustered regularly interspaced short palindromic repeats
CRISPRi – CRISPR interference
CS – citrate synthase
CV – complex V, ATP synthase
cyt – cytochrome

Da – dalton(s)
DAG – diacylglycerol
DAPI – 4’,6-diamidino-2-phenylindole
DCPI – 2,6-dichlorophenolindophenol
DEPC – diethylpyrocarbonate
DIG – digoxigenin
DMEM – Dulbecco’s modified Eagle’s medium
DMSO – dimethyl sulphoxide
DNA – deoxyribonucleic acid
DTNB – 5’,5’-dithiobis-2-nitrobenzoic acid
DTT – dithiothreitol
E – embryonic stage
ECL – enhanced chemiluminescence
EDTA – ethylenediaminetetraacetic acid
EGTA – ethylene glycol tetraacetic acid
ENCODE – Encyclopaedia of DNA Elements
eRNA – enhancer RNA
ESC – embryonic stem cell
ETC – electron transport chain
EYFP – enhanced yellow fluorescent protein
FAD – flavin adenine dinucleotide
$FADH_2$ – reduced flavin adenine dinucleotide
FANTOM – Functional Annotation of the Mammalian Genome, RIKEN, Japan
FBS – foetal bovine serum
FISH – fluorescent ISH
Fluc – firefly luciferase
g – gravitational force
GENCODE – ENCODE gene annotations
GFAP – glial fibrillary acidic protein
GFP – green fluorescent protein
GTEx project – Genome-Tissue Expression project
H3K27 – histone 3 lysine 27
H3K4 – histone 3 lysine 4
H3K9 – histone 3 lysine 9
HA – haemagglutinin
HDBR – Human Developmental Biology Resource
HEK293(T) – human embryonic kidney cells
HEPES – 4-(2-hydroxyethyl)-1-piperazineethanesulphonic acid
hiPSC – human induced pluripotent stem cell
IAA – iodoacetamide
IBA1 – ionized calcium binding adaptor molecule 1
ICC – immunocytochemistry
IHC – immunohistochemistry
INS-1 – rat pancreatic β insulinoma cells
IP – immunoprecipitation
IP_3 – inositol-3,4,5-trisphosphate
iPSC – induced pluripotent stem cell

ISH – in situ hybridisation
kb – kilobase(s)
KD – knockdown
KME buffer – potassium-MOPS-EGTA buffer
lacZ – β-galactosidase
LB – lysogeny broth
LDS – lithium dodecyl sulphate
Linc-MD1 – muscle differentiation lncRNA 1
lincRNA – long intergenic non-coding RNA
LMF1 – lipase maturation factor 1
LNA – locked nucleic acid
lncRNA – long non-coding RNA
mAChR – muscarinic acetylcholine receptor
MEM – minimal essential medium
miRISC – microRNA induced silencing complex
miRNA – microRNA
MOPS – 3-(N-morpholino)propanesulphonic acid
MPTP – 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
MRE – microRNA response element
mRNA – messenger RNA
MS – mass spectrometry
mtDNA – mitochondrial DNA/genome
MTT – 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
MUSCLE – multiple sequence comparison by log-expectation
My – million years
Mya – million years ago
N2A – mouse neuroblastoma 2A cells
NAD^+ – nicotinamide adenine dinucleotide
NADH – reduced nicotinamide adenine dinucleotide
ncRNA – non-coding RNA
N.D. – not detected
NDUF – NADH:ubiquinone oxidoreductase gene
Nluc – Nano-luciferase
NMD – nonsense-mediated decay
NT – non-transfected
nt – nucleotide(s)
OLIG1 – oligodendrocyte transcription factor 1
OLIG2 – oligodendrocyte transcription factor 2
OPC – oligodendrocyte precursor cell
ORF – open reading frame
OXPHOS – oxidative phosphorylation
P – post-natal stage
PAGE – polyacrylamide gel electrophoresis
PBS – phosphate-buffered saline
PCR – polymerase chain reaction
PDK1 – 3-phosphoinositide-dependent protein kinase 1
PFA – paraformaldehyde

PIP_3 – phosphatidylinositol-3,4,5-trisphosphate
PKC – protein kinase C
PLC – phospholipase C
PMA – phorbol 12-myristate 13-acetate (TPA)
PolII – RNA polymerase II
PolIII – RNA polymerase III
POLR2A – RNA polymerase II subunit A
PTEN – phosphatase and tensin homologue
PTENpg1 – PTEN pseudogene 1
PVDF – polyvinylidene difluoride
qPCR – quantitative PCR
RACE – rapid amplification of cDNA ends
RIPA – radioimmunoprecipitation assay
RLM-RACE – RNA ligation-mediated RACE
RNA – ribonucleic acid
RNA-seq – RNA sequencing
ROS – reactive oxygen species
rpm – revolutions per minute
rRNA – ribosomal RNA
RT-qPCR – reverse transcription quantitative PCR
SDS – sodium dodecyl sulphate
SDS-PAGE – SDS PAGE
secNluc – secreted Nano-luciferase
shRNA – short hairpin RNA
SH-SY5Y – human neuroblastoma cells (sub-clone of SK-N-SH)
SK-N-SH – human neuroblastoma cells
smORF – small open reading frame
SOX8 – sex determining region (SRY)-related high motility group (HMG)-box protein 8
SRA – Sequence Read Archive
SSC buffer – saline-sodium citrate buffer
TBS – Tris-buffered saline
TDMD – target-directed miRNA degradation
TPA – 12-O-tetradecanoylphorbol-13-acetate (PMA)
Tris – tris(hydroxymethyl)aminomethane
tRNA – transfer RNA
TSAP – thermosensitive alkaline phosphatase
TSS – transcriptional start site
TTS – transcriptional termination site
TUBA1A – α-tubulin 1a
uORF – upstream ORF
UQCR – ubiquinol:cytochrome *c* oxidoreductase gene
UTP – uridine triphosphate
UTR – untranslated region
X-gal – 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (BCIG)

CHAPTER 1: INTRODUCTION

1.1 Pervasive transcription in the human genome produces long non-coding RNAs

It has recently been estimated that around 8% of the human genome is evolutionarily constrained[1-4] and thus functional in a sequence-specific manner. This is substantially greater than the 1.22% of the genome that comprises protein-coding exons[5], even allowing for elements such as promoters. Over the course of the past 25 years, thousands of studies have advanced our understanding of the diversity of the non-coding genome, and shown that RNA, independently of its protein template role, is an important instigator of function in all life on Earth.

1.1.1 Non-coding RNAs are numerous and ubiquitous

Long non-coding RNAs (lncRNAs) are classified as RNA transcripts that have no predicted coding capacity and are greater than 200 nucleotides in length. This distinguishes them from the 'small' non-coding transcriptome, which includes microRNAs (miRNAs), transfer RNAs (tRNAs), Piwi-interacting RNAs (piRNAs), and small nuclear and nucleolar RNAs (snRNAs and snoRNAs). Many lncRNAs are processed analogously to protein-coding mRNAs, being transcribed by RNA PolII[6], spliced at canonical splice sites[7], 5' capped[6], and polyadenylated[8]. Despite uncertainty in their exact number[9,10], with over 16,000 genes and nearly 30,000 transcripts in the stringently annotated GENCODE set (version 29, November 2018; gencodegenes.org)[7,11,12], it is clear that lncRNAs comprise a substantial portion of the human transcriptome. They represent by far the majority of non-coding RNA[13], and are of a magnitude comparable to the ~20,000 protein-coding genes. Functional lncRNAs have been identified in all species, including plants[14], yeast[15], bacteria[16], and viruses[17].

Typically, in the absence of functional data with which to further categorise lncRNAs, several criteria have been used: genomic architecture and respective relationships to neighbouring genes; differential epigenetic modifications associated with their transcription[18]; and sub-cellular localisation. LncRNAs may be intergenic (lincRNAs), thereby being distinct from neighbouring genes. Many, however, overlap other genes, with respect to which these may be sense or antisense; they may overlap introns or exons or both, or even encompass the entire gene within an intron of their own[19].

1.1.2 Poor sequence conservation of lncRNAs as a group does not preclude functionality

It has been suggested that the correlation between the proportion of non-coding genomic sequence and the developmental complexity of metazoan organisms[20], in spite of the apparent lack of correlation with genome size or protein-coding gene number in eukaryotes, provides a compelling hint that these sequences are at least partly functional and have contributed to ongoing evolution, including human speciation[21-23].

Studies of mouse lncRNAs reveal modest levels of purifying selection[24] that have reduced mutation rates within their sequences, splice sites, and promoters, and which rebut the question of whether lncRNAs solely represent 'transcriptional noise'[25]. Although the evolutionary constraint observed upon lncRNAs is modest compared to protein coding exons[24,26] and ~80% of lncRNAs identified in humans are primate lineage-specific[27], it is clear that lncRNAs may been conserved throughout evolution[28,29], with at least 2,500 originating prior to the Euarchonta-Glires divergence (~90 Mya), and over 400 being common to all tetrapods, having been retained over 300 My of evolution[27].

However, estimates of functional sequence within lncRNAs are modest (4.1-5.5%, Ponjavic et al. (2007)[24]) and it may be that functionality is conferred by a relatively small proportion of sequence[30,31]. However, despite these modest levels of sequence constraint under purifying selection[31], conservation of high GC content and of exonic splicing enhancer elements supports higher selection for efficient transcription and splicing of human lncRNAs[32,33]. A study of lncRNAs across vertebrate species supports the hypothesis that function is conveyed by short stretches of sequence conservation that moreover can tolerate substantial genomic architectural rearrangements[34].

1.1.3 The extent of the functional relevance of all lncRNAs is not clear

These large-scale analyses have illustrated that: a) pervasive transcription produces an extensive transcriptome beyond simply protein-coding exons and introns; b) many lncRNAs are poorly conserved; and c) a few lncRNAs are highly evolutionarily conserved, but this is indicative but by no means absolute proof of important functionality. Experimentally characterised case studies have now validated the significance of lncRNAs as a biologically important class of genes. However, this does not preclude the notion that some or many lncRNAs may represent functionless products of transcriptional stochasticity, or that singly characterised examples of particular mechanisms may not represent widespread mechanisms of action[35].

That said, the biochemical versatility of RNA enables myriad possible functional interactions with all other categories of molecule in the cell, including DNA, RNA, and protein[36]; combined with its relative freedom within the cell, the theoretical mechanisms by which lncRNAs may influence biology are numerous.

Ultimately, determining the functional importance of any single lncRNA *in vivo* requires carefully designed experimental investigation[37,38], and large studies to date have been marred by poor experimental design; using large-scale genomic deletions in the absence of RNA-based rescue, Sauvageau et al.[39] cannot truly conclude that the lncRNAs *Peril* and *Mdgt* have an RNA-mediated function "required for life". The same problem is presented by the conclusions of a second similar study in association with the same laboratory[40].

There are many hundreds of lncRNAs to which roles have been ascribed *in silico*, *in vitro* and, more rarely, *in vivo*. This introduction will briefly visit exemplar lncRNAs representative of our current understanding of molecular mechanisms, focusing on the more recently characterised cytoplasmic lncRNAs, as well as the current state of the field in identifying physiological relevance of lncRNAs in biology and in disease.

1.2 Many characterised lncRNAs play roles in co-transcriptional regulation

Amongst the first mammalian lncRNAs to be characterised was *XIST*, which is crucial in the process of X-inactivation within female eutherian mammals[41-46]. Along with the very first eukaryotic lncRNA *H19*[47], an imprinted lncRNA whose role is still poorly understood[48], *XIST* set a precedent for many years of focus within the field upon mechanisms by which lncRNAs may act as (co-)transcriptional regulators of gene expression. Though nuclear lncRNAs will not be the focus of this thesis, the following examples demonstrate a sample of the mechanisms ascribed to these genes and some of the pitfalls in investigating lncRNAs.

1.2.1 LncRNAs co-transcriptionally regulate gene expression via diverse interactions

In contrast to the ubiquitous activity of *Xist*, for example, most nuclear lncRNAs are associated with the regulation of a specific locus or set of genes, defined by their interactions with chromatin and proteins. The lncRNA *Mhrt* (*Myheart*) was shown to play a role in regulating cardiac function in response to stress and contribute to hypertrophic heart failure. A competitive bi-directional relationship exists between the master chromatin regulator Brg1 and *Mhrt*: under non-stressed conditions, when it is expressed abundantly, *Mhrt* RNA binds the helicase domain of Brg1 and prevents it targeting expression of *Mhrt*, *Myh6*, and *Myh7*[49,50].

Only when Brg1 is induced by severe stress do levels overcome the inhibition of Mhrt and allow it to act as a chromatin modifier. This elicits a positive feedback loop whereby further expression of *Mhrt* is inhibited, and the Myh6 to Myh7 myosin isoform switch characteristic of hypertrophy occurs[49]. Transgenic restoration of *Mhrt* maintains the repression of Brg1 and prevents the pathological cascade; the competitive *Mhrt*-Brg1 feedback interaction thereby acts as a switch[49]. Human *MHRT* is transcribed from a syntenic locus and is also downregulated in ischaemia and hypertrophy, suggesting a conserved mechanism in cardiomyopathy[49]. Whether exogenous upregulation of *MHRT* would protect against hypertrophy in humans is unknown, especially since humans maintain MYH7 levels in the ventricles throughout development and adult life[51].

Fendrr (foetal-lethal non-coding developmental regulatory RNA) was one of the first mammalian lncRNAs whose function *in vivo* was characterised using genomic deletion as well as, importantly, RNA rescue[52,53]. During mid-gestation, *Fendrr* is localised to the lateral mesoderm that will form the heart and body wall, alongside the transcription factor gene Foxf1, from which it is divergently transcribed. Genetic deletion of *Fendrr* was found to be lethal at ~E13.75, with embryos frequently exhibiting omphalocele. Crucially, implication of the lncRNA transcript specifically was confirmed by rescue: an additional inserted copy of the endogenous locus restored ~50% expression and was sufficient to restore normal non-lethal development[52]. *Fendrr*-null embryos revealed heart and body wall defects, and dysregulated methylation of several key developmental regulators supported a hypothesis of epigenetic mediation suggested by the delay between the time of normal *Fendrr* expression and the appearance of a deletion phenotype: epigenetic signatures may be carried through generations of cells during development, propagating the effects of dysregulation into daughter lineages[52]. *Fendrr* was found to interact with PRC2 subunits, responsible for H3K27 methylation, occupancy of which was depleted at differentially expressed loci following *Fendrr* deletion[52,54]. The lncRNA binds via a putative DNA:DNA:RNA triplex to both the *Foxf1* (in *cis*) and *Pitx* (in *trans*) promoters and thereon recruits PRC2 to silence expression[52]. Interestingly, *Fendrr* was also subject to mutation as part of a large lncRNA deletion programme by Sauvageau et al., who also concluded perinatal lethality of deletion of the lncRNA, but demonstrated different underlying physiological defects, with respiratory failure being the cause[39]. This illustrates the importance of a carefully designed ablation strategy[37]: since rescue of lncRNA levels was not desmonstrated[39], perturbation of another genetic element, such as the neighbouring *Foxf1* gene, might underlie the alternative pathology. A second study again demonstrated perinatal respiratory lethality using a different ablation strategy, but still omitted an RNA rescue[40]. A genetic study of

the fatal neonatal pulmonary disorder ACD/MPV was recently associated with mutation of *FENDRR*, supporting its conserved physiological importance[55].

The structural scaffolding activity of some lncRNAs is more extensive. *HOTAIR* (HOX antisense intergenic RNA) was first identified as an antisense lncRNA transcribed from within the HOXC cluster of homeotic genes[56]. Expression of human *HOTAIR* was found to be localised to the posterior and distal tissue, where it silences HOXD genes in *trans* by recruiting PRC2[56] via a short sequence at its 5' end[57-59]. The 3' end, in addition, is capable of binding the LSD1/CoREST/REST complex, which acts as an H3K4 demethylase[60]. *In vitro* human data therefore suggests that *HOTAIR* bi-functionally tethers multiple chromatin modifiers to repress the HOXD locus. Interestingly, while a transcript sharing some homology to human HOTAIR is present in mouse, it exhibits substantial differences, lacking several of the identified binding domains[61]. Deletion of the entire HoxC locus has little effect upon the methylation status or expression of HoxD genes[61,62], but compensatory mechanisms due to the unspecific nature of the ablation cannot be excluded[63]. Indeed, targeted disruption of murine *Hotair* was shown to elicit dysregulation of gene expression, including de-repression of HoxD genes, and subsequent homeotic transformation and skeletal malformation phenotypes *in vivo*[63], some of which was corroborated by two additional disruption models[40]. A third study indicated contrary results, with little evidence of a highly penetrant homeotic phenotype and discordant differential gene expression[64]. Experimental differences, including mouse strains, which are known to display homeotic variability[65,66], and samples analysed, with heterogeneous cell populations within samples perhaps masking phenotypes[67], might underlie these conflicting data. None of these models assessed whether phenotypes were rescued by RNA replacement, suggesting that ablation strategy contributes to the observed phenotype. Furthermore, the substantial dissimilarity between human *HOTAIR* and mouse *Hotair*[61], the PRC2- and LSD1/CoREST/REST-binding properties of which have not been assayed, leaves the relevance of these models to human *HOTAIR* unclear.

The lncRNA *NEAT1* (nuclear-enriched abundant transcript) seeds the formation[68] of paraspeckles[69,70], nuclear sub-structures of RNA and RNA-binding proteins to which it is exclusively localised[71]. *Neat1* deletion in mice leads to loss of paraspeckles in the tissues in which *Neat1* is normally highly expressed[72], and highly specific phenotypes, including defective mammary epithelial cell proliferation, leading to impaired lactation and low pup survival[73], and impaired female fertility, owing to an unusually stochastic sometimes-failure to

form the corpus luteum and generate progesterone[74]. However, it remains to be seen whether *Neat1* function is "strictly and invariably associated with paraspeckle formation"[73].

These examples illustrate, as well as the diverse range of interactions, the varied scale of action for nuclear lncRNAs. From the single-locus activity of *Mhrt* to the extensive *trans* interactions of *HOTAIR* and the power of *Xist* in silencing an entire chromosome, lncRNAs appear to implement evolutionarily conserved control at all levels of gene expression regulation. They also demonstrate some of the difficulties faced in identifying lncRNA function *in vivo*: genetic manipulation strategy may profoundly affect phenotypes; disentangling DNA and RNA function is critical, especially since many lncRNA loci overlap or abut other genomic elements; effects of lncRNA disruption may be propagated throughout development, obscuring direct effects; the poorer conservation of lncRNAs may limit the choice of model and relevance of findings between species; and in the case of unusual lncRNAs such as *NEAT1*, disentangling additional functionality from a previously characterised function, e.g. paraspeckle formation, may be challenging.

1.2.2 At some loci, lncRNAs may be side-products of functional DNA elements

Enhancer lncRNAs (eRNAs)[75] are non-coding transcripts transcribed from enhancer sequences, regions of DNA that act as transcription factor binding regions to regulate RNA polymerase activity at nearby promoters. Traditionally, eRNAs were considered to be short, short-lived, and unlike mRNAs and promoter-associated lncRNAs[18] in that they are rarely spliced or polyadenylated[75], though these criteria have since been contended[76]. *Lockd* is a murine eRNA that is transcribed from a proximal enhancer element that regulates transcription of the neighbouring gene *Cdkn1b*[77]. It was originally considered a lncRNA due to being multi-exonic and polyadenylated, and the presence of typical promoter hallmarks, including $H3K4Me_3$ and DNase I hypersensitivity at its 5' end[78,79]. However, deletion of the *Lockd* gene locus was found to strongly influence expression at the neighbouring *Cdkn1b* promoter, while truncation of *Lockd* to only 80 nt using a premature polyadenylation signal had no such effect[77]. It was therefore concluded that the presence of the transcription site, acting as an enhancer, rather than the transcription of the *Lockd* gene, was responsible for *Cdkn1b* regulation. Indeed, next-generation Capture-C[80,81] demonstrated a physical interaction between the *Lockd* enhancer and the *Cdkn1b* promoter[77].

1.3 Cytoplasmic lncRNAs

It has in recent years become apparent that many lncRNAs are enriched in the cytoplasm[82-84]. Indeed Ulitsky and Bartel suggest that the assumption of nuclear localisation of lncRNAs is "perhaps [their] most common misperception"[82]. Efforts to characterise the functions of these are more recent, and attention has largely focused on one particular mechanism: the ceRNA hypothesis.

1.3.1 RNA-miRNA-RNA crosstalk mediates the function of a subset of lncRNAs: ceRNAs

Following the characterisation of several RNAs, foremost including *PTENpg1*[85] and *HULC*[86] in mammals, IPS1[87] in *Arabidopsis thaliana*, and the *Herpes* virus snRNA *HSUR1*[88], which were found to elicit post-transcriptional control over gene expression via the binding of miRNAs, a seminal publication by Salmena et al.[89] described the competitive/competing endogenous RNA (ceRNA) hypothesis, which lies at the interface between short and long regulatory RNAs. MicroRNAs are small non-coding RNAs that act as (typically repressive) post-transcriptional regulators of gene expression. The biogenesis of a miRNA and its subsequent incorporation into a miRNA-induced silencing complex (miRISC) that regulates gene expression is illustrated in Figure 1.1A-C. Via the sharing of miRNA response element (MREs) sequences, transcripts are placed under competition for the binding of free miRNAs. MicroRNA-mediated repression is thereby shared between those transcripts, and perturbation of the levels of one transcript may result in titration of miRNA levels that result in concerted gene expression changes in the other targets of that miRNA (Figure 1.1D), as had been predicted computationally by Hervé Seitz[90]. In the characterisation of the first ceRNA, *IPS1* (*INDUCED BY PHOSPHATE STARVATION*) in *A. thaliana*, this was referred to as "target mimicry", under the assumption that a coding transcript, in that case *PHOSPHATE 2* mRNA, is the 'true' target of a miRNA, here being the phosphate starvation-induced miR-399. However, in light of the extension of a simple RNA-miRNA-RNA ceRNA hypothesis into one which describes putative networks (ceRNETs)[91] of cross-talk between potentially the entire transcriptome[89], the definition of a 'true' miRNA target has become much more diffuse.

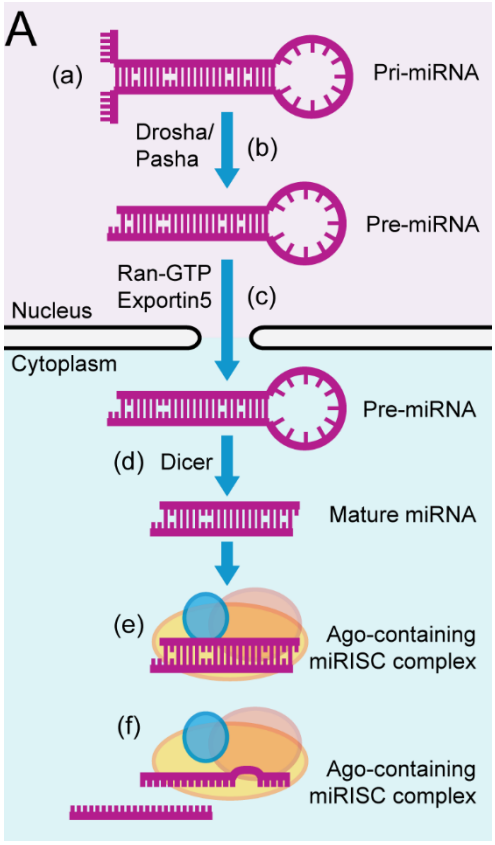
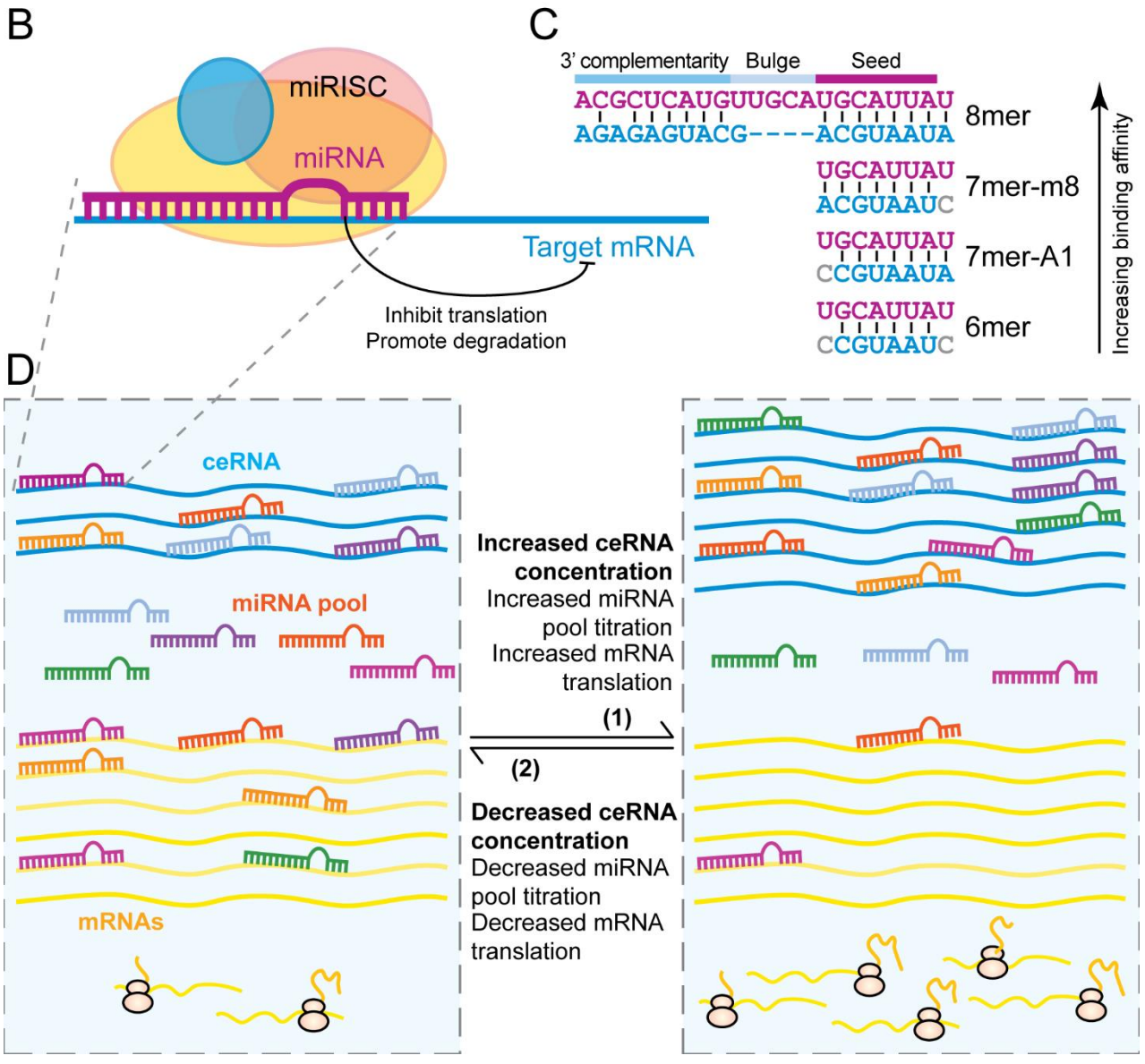


Figure 1.1: The biogenesis of microRNAs and their role in gene regulation via the ceRNA mechanism. (A) (adapted from Bartel, 2004[92]) (**a**) miRNAs are transcribed by RNA Pol II, and 5' capped and polyadenylated[93] to form a primary-miRNA (pri-miRNA). (**b**) In animals, the pri-miRNA is processed by the nuclear Microprocessor complex (containing Drosha[94] and its cofactor Pasha/DGCR8[95]) to form a pre-miRNA. Following export by Ran-GTP and Exportin5 (**c**), Dicer[96,97] in the cytoplasm cleaves the terminal stem loop to produce a mature miRNA of ~22 bp[92], which associates with a miRNA induced silencing complex (miRISC) alongside the endonuclease Argonaute (Ago)[98,99] (**e**). (**f**) Duplex unwinding releases the passenger strand, leaving the miRISC able to **(B)** bind miRNA response elements (MREs). **(C)** miRNA binding is mediated primarily by the seed region, with four types of seed complementarity[100], from 8mer matching including seven miRNA:RNA matches and the preferred adenosine anchor[100,101], to minimal 6mer. 3' complementarity supplements binding and may influence the likelihood of RNA cleavage or miRNA degradation. **(D)** In a ceRNA system, increased ceRNA concentration (**1**) leads to increased titration of miRNAs from the free pool, reducing their repression of mRNAs and increasing mRNA expression. Conversely, depletion of the ceRNA (**2**) leads to concomitant reduction of mRNA expression due to greater miRNA availability.



1.3.2 The widespread importance of the ceRNA mechanism remains contested

Since the transcriptome-wide relevance of the ceRNA mechanism was first postulated by Salmena et al.[89], myriad studies have sought to establish its significance, using *in silico*, *in vitro*, and *in vivo* models, as well as establish criteria for the necessary parameters of a ceRNA system. Unfortunately, most studies to date have been unsatisfactory in their approach, oversimplifying the numerous parameters that may influence ceRNA efficacy and significance[102].

A genome-wide analysis from the laboratory of Hervé Seitz, the original devisor of what became the ceRNA mechanism[90], showed that for many miRNAs, abundance of mRNA with MREs for that miRNA correlates with conservation of those MREs across human tissues[103]. This supports the hypothesis that evolutionarily beneficial titrating transcripts should be highly expressed and their interacting MREs conserved. It was noted that correlation was high for miRNAs that are specifically enriched within another tissue, such as the liver-enriched miR-122 in muscle, suggesting a selective pressure to titrate miRNAs that are already lowly expressed. Interestingly, an exception was found in the brain, where brain-enriched miRNAs exerted the greatest selective influence upon conserved sites. This is possibly due to the heterogeneity of the brain, where different miRNAs would be expected to be differentially active across cell populations.

Owing to the enrichment of shared MREs for brain-specific miRNAs within their 3'-UTRs, Wehrspaun et al. suggested that the observation that the brain expresses longer isoforms of genes encoding ion channels or transporters may be explained by the selective advantage of ceRNA co-regulation[104].

1.3.2.1 Models of miRNA titration have yielded differing views upon ceRNA validity

Initial mathematical models, in which one or several ceRNAs were in competition for the binding of one or several miRNAs, produced general hypotheses about interaction stoichiometry. However, the simplifying assumptions made by these studies, which offered support to the potential of early ceRNA studies, reduce their impact in modelling physiological environments[102]. Ala et al.[91], using an *in silico* ceRNET model subsequently validated *in vitro* using *PTEN* and its ceRNA *VAPA*, found that the most potent ceRNA cross-regulation was achieved when the expression of competing transcripts was equimolar, and miRNA abundance was intermediate. Interestingly, the study also suggested intersection of ceRNETs and transcriptional networks[91]. Similar conclusions were reached by Figliuzzi et al., who noted that the structure

of a ceRNET may alter following perturbations in miRNA levels or ceRNA levels with different transcripts and mediators poised in different states of the same system[105]. It was also suggested that interactions may be asymmetric, with some transcripts more prone to perturbation, and also limited by noise, which could be amplified or reduced depending on relative ceRNA correlations[105]. Interestingly, a later study by Bosia et al. also concluded that miRNA titration may process noise, and showed that the amplification of noise may result in concerted expression changes or a bimodal switch effect[106].

The strongest opposition to the physiological relevance of the ceRNA mechanism is based upon incompatibility of stoichiometry: Denzler et al.[107] (commentary[108]) suggested that ceRNAs are rare, if they exist at all, in normal cells, following a series of experiments in which they showed that an artificial ceRNA-mimicking system involving intracellular expression of a miR-122 target site failed to de-repress miR-122 targets in hepatocytes, requiring supra-physiological concentrations of the ceRNA. Additionally, the apparent reduction of the free miR-122 pool using an inhibitor did not alter the threshold of ceRNA required for de-repression. This suggested a non-hierarchical model of miRNA targeting across the genome, whereby miRNA binding is equally likely at each of the large number of sites, reducing the potential of a ceRNA to non-negligibly alter repression[107]. A number of issues limit the conclusions of this model, including the disingenuous choice of miRNA to model in the form of miR-122, expression of which in the liver is amongst the highest of any mammalian miRNA anywhere[107,109-111]. It is unsurprising that it would seem to be a poor ceRNA-mediating candidate; indeed, no model pre-existing nor derived since would have suggested so. Furthermore, it was assumed that: a) choosing an endogenous transcript at random (in this case *AldoA*) would accurately model a ceRNA; b) that modulating the abundance of miRNAs using an exogenous miRNA inhibitor will alter the abundance of miRISCs, and the level of repression, in a linear fashion. Assuming that ceRNA networks have been optimised by evolution[103], potentially with contextually optimised binding sites[112], and given that models suggest that different transcripts and miRNAs may be active in titration across different expression profile contexts[105,113,114], AldoA may simply lack the optimised ability to act as a ceRNA in the model chosen. Mayya and Duchaine[115] showed a substantial disparity between free miRNA levels and miRISC incorporation for individual miRNA species, as well as differential qualitative and quantitative behaviours of targets following miRNA inhibition, noting that only some miRNAs would have the potential to mediate a ceRNA effect[115]. Interestingly, Luna et al.[116] have since shown that during infection of hepatocytes, hepatitis C virus utilises an RNA to sequester miR-122 in a ceRNA-like mechanism

to de-repress miR-122 targets, with "biological consequence" that Denzler et al.[107] had deemed impossible. Given that some *in vitro* studies of ceRNA candidates has indicated an absence of phenotype upon ceRNA perturbation until the system was challenged[117,118], studies ought ideally to assess effects upon specific targets and downstream markers rather than simply a predicted target pool[102]. Especially since even complete miRNA mutants in multiple species typically exhibit limited phenotypes that are not overtly related to a majority of their targets[119-121], being typically related to one or two[103]. The "mixed-affinity model" favoured by Denzler et al.[107] was the subject of a second study by the same authors in which the findings regarding miR-122 were generalised to other miRNAs[122].

A study similar in design but different in its conclusions by Bosson et al.[113] instead took the assumption that miRNA binding is hierarchical: that is, that 8mer MREs, for example, are more likely to be bound by miRNAs and that 6mers, for example, contribute a lower expected binding fraction to the target pool. Using genome-wide CLIP-derived data to garner an accurate picture of miRISC binding at different sites based upon miRISC coverage, and validation of the subsequent model using single cell reporter assays, it was confirmed that miRNA:target ratio informs the potential activity of a ceRNA. Like Mayya and Duchaine[115], and Mullokandov et al.[123], Bosson et al. showed that miRISC occupancy is not simply a function of miRNA abundance, with broad differences between the behaviour of similarly expressed miRNAs[113]. Examining a single model[107] is therefore insufficient to assess the relevance of all ceRNA systems. With lower estimates of the physiologically relevant miRNA:target ratio, Bosson et al. concluded that the ceRNA mechanism is viable within physiological contexts, depending upon the relative levels of participating transcripts and miRNAs[113]. They noted that the refractive nature of some miRNAs with respect to titration may reflect their importance in enforcing cell identity[118,124,125], reminiscent of the observation by Pinzón et al. that levels of miRNA-titrating transcripts are enriched in tissues in which that miRNA is not highly expressed[103] and therefore presumably not strongly influencing cell identity.

1.3.2.2 The impact of multiple important parameters upon ceRNA regulation remain unexplored

Models to date, both experimental and in silico, assume a homogeneous distribution of miRNAs and targets throughout the cells, thereby precluding any contribution by spatiotemporal enrichment of miRNAs, their targets, and possible ceRNAs, particularly in spatially diverse cells such as neurons. Sub-cellular enrichment of miRNAs to organelles such as mitochondria[126-132], or to synapses[133,134], for example, has been observed,

as has the localised repression of an mRNA target (CamKIIα) by synapse activity-stimulated maturation of a miRNA (miR-181a)[135]. These trends may partially underlie the observation that levels of miRNA repression cannot be predicted by abundance alone[123]. Furthermore, while miRISC levels better indicate the repressive power of a miRNA within a system[136], the correlation of miRNA abundance and miRISC levels differ highly between individual miRNA species[115]. Crucially, La Rocca et al. showed that *in vitro* models poorly resemble physiological systems of *in vivo* tissues, where miRISCs accumulate in large complexes and their involvement in repression is low[137].

Other factors that likely influence contextual ceRNA activity but that most models have not sought to incorporate include: cooperativity between neighbouring MREs; the role that local sub-seed regions play in directing miRISC targeting; and other cellular regulation of AGO and the miRISC. Cooperativity between neighbouring MREs binding the same or different miRNA families increases repression relative to disparate sites[138-140], and has also been shown to reduce the level of titration required for de-repression of targets by up to 50%[122]. Initial binding of the human miRISC is directed by the sub-seed region (nucleotides 2 to 4-5)[141-143], prior to a conformational change that permits further complementarity-directed binding[141] and repressive action[143]. Since the miRISC appears to utilise lateral one-dimensional scanning of RNA transcripts to find full seed MREs, and that sub-seed regions will partially stabilise its interaction[142], the number of sub-seeds within a transcript may modulate the efficacy with which the transcript is targeted and may act to titrate the miRNA pool as a ceRNA[102]. The activity of miRISCs may also be modulated by AGO, phosphorylation of which appears to differentially affect binding to individual MREs and subsequent repression[144,145], and may contribute to a cellular environment that potentiates ceRNA activity. A limited abundance of AGO, which is altered following intracellular signalling[146], has also been suggested to potentiate competition for miRNA incorporation into the RISC[147], reflecting the observation of Figliuzzi et al. that miRNA titration may also be formulated from a perspective of miRNAs competing for their targets[105]. Complementary to the contribution of sub-seed sites and seed cooperativity is the possibility of titration of multiple miRNAs with common targets, which a comprehensive model by Chiu et al.[114] suggests produces more robust and likely consequential ceRNA effects.

1.3.2.3 Processes such as differentiation and oncogenesis highlight the potential for ceRNA regulation

It is true, as Denzler et al.[107,122] stated, that large changes in gene expression are not common in healthy cells that are terminally differentiated. During the dramatic reprogramming of expression during differentiation, however, abundance of a transcript may change wildly: ultra-high-resolution RNA-seq in the developing *Xenopus tropicalis* embryo reveal that some transcripts are subject to incredibly dynamic changes, increasing by factors up to >10^6 transcripts *per hour* throughout the embryo[148]. Perhaps it is in these situations that the ceRNA mechanism is most physiologically relevant.

A number of ceRNAs characterisation studies *in vitro* have been undertaken in embryonic or induced pluripotent stem cells (ESCs, iPSCs) during differentiation, including *linc-MD1* (see also Table 1.1), the expression of which changes from negligible to ~1000 copies per cell during myoblast differentiation[149,150], and *linc-ROR*[118]. Indeed, following the observation that *linc-RoR* promotes reprogramming of fibroblasts into iPSCs[151], Wang et al. identified *linc-ROR* (regulator of reprogramming) as a key pluripotency factor in ESCs[118]. Its expression is high in self-renewing ESCs, where it binds miR-145, which acts repressively upon the core pluripotency transcription factors *Oct4*, *Sox2*, and *Nanog*; furthermore these activate transcription of *linc-ROR* forming a positive feedback loop that stabilises the self-renewing state[118]. During stimulated differentiation, the ~90% loss of *linc-ROR* expression permits the rise of miR-145 levels, repressing the pluripotency factors and permitting expression of differentiated lineage genes. Ectopic knock-down of *linc-ROR* was sufficient to induce differentiation, and this was antagonised by inhibition of miR-145 or knock-down of *Oct4*. Likewise, ectopic overexpression of *linc-ROR*, but not a miR-145-MRE mutant transcript, delayed differentiation and reduced pluripotency repression; this was antagonised by miR-145 mimics[118]. These data strongly suggest that *linc-ROR* is crucial for the maintenance of self-renewal in ESCs in vitro, and that it acts as a ceRNA to stabilise pluripotency gene expression programmes by de-repressing *Oct4*, *Sox8*, and *Nanog*.

Pinzón et al. used a system of differentiating myoblasts to assess the contribution of the changing transcriptome to miRNA levels and the potential for titration[103]. Differentiating cells were treated with inhibitors of two muscle-elevated miRNA families, miR-1a/206 and miR-133, and *bona fide* targets were identified by Ago immunoprecipitation followed by RNA-seq. miRNA and target concentrations were then used to calculate free/bound miRNA equilibria, from which the titration contribution of each target binding site to the overall binding equilibrium was established. Several sites were deemed to have substantial

titrating effects, with seven transcripts titrating at least 10% of one miRNA species, and one transcript *Tmsb4x* titrating ~50% of the miR-1a/206 pool in each time point. This was validated by mutation of the endogenous MRE by CRISPR/Cas9 technology and assessment of miRNA activity using an encoded reporter, which was significantly more highly repressed following loss of *Tmsb4x* titration activity. The results of this study, as well as the relationship between *Tmsb4x* and miR-1a/206 are both in line with the conclusions of Bosson et al.[113] and others[91,105,106,114], with low-moderate miRNA levels and relatively higher ceRNA levels.

Another situation in which expression patterns are liable to dramatic change is during cancer: with mutation of key transcription factor and DNA protection pathways, and subsequent loss of constraint, expression of a gene may be removed or amplified to magnitudes not tolerated in normal cells. The prevalence of cancer cell models in ceRNA literature is therefore perhaps neither coincidence nor entirely a product of their ubiquity as *in vitro* research tools. By modelling ceRNA interactions within four cancer gene expression datasets and examining the interactions common to each, Chiu et al. produced a "pan-cancer interactome" (PCI) of ~165,000 ceRNA interactions between 3800 genes[114]. This overlap was only 13% of the total interactions modelled, with only 11 of 247 miRNAs mediating an interaction in all four cancers, suggesting that each cancer context was largely independent. The PCI was characterised by the large number of mediators capable of coupling interacting genes (3-fold higher than each individual context) and thus its functionality within even diverse miRNA expression profiles[114]. Functional regulation within the PCI was demonstrated using two sub-networks of oncogenes and tumour suppressors: in eight different cancer cell lines (three from the model, five additional), overexpression of 3'-UTRs was found to elicit upregulation (average 1.4-fold, up to 2-fold) of the interacting ceRNAs in 80% of pairs. ceRNA 3'-UTR overexpression antagonised siRNA depletion of target genes, and MRE mutants demonstrated the miRNA-dependence of the phenotypes[114]. Expression correlations predicted by the PCI could accurately predict the expression of genes within 12 additional cancer and non-cancer datasets, with the highest predictive power for genes involved in more interactions, owing to the accumulated correlation in expression of those genes[114].

1.3.2.4 Addressing the ceRNA hypothesis *in vivo*

As with many lncRNAs to which functions have been ascribed *in vitro*, the biological significance of ceRNAs is hotly contested[107,108,122], with no studies conclusively illustrating the importance of exemplar ceRNA transcripts *in vivo*[102]. Examples of incompletely-characterised ceRNAs may be found in Table 1.1, which reviews the evidence presented to support the functional relevance of the transcripts and further experiments required to support this relevance *in vivo*.

Investigation undertaken and evidence presented	Further work required
***linc-MD1*[149]**	
Cesana et al. identified *linc-MD1* as a conserved lncRNA that is restricted to skeletal muscle, where it is expressed during differentiation of myoblasts and dysregulated in Duchenne muscular dystrophy. Luciferase assays were used to show that *linc-MD1* binds both miR-133 and miR-135, which are also bound, respectively, by the transcripts of two major muscle transcription factors MAML1 and MEF2C. LNA miRNA inhibitors ablated the downregulation of the reporters by endogenous miRNAs, and when co-expressed, upregulation of linc-MD1 stabilised the target reporters, suggesting direct competition of miRNA binding. Importantly, *linc-MD1* modulation altered MAML1 and MEF2C RNA and protein levels in a concerted manner, and miRNA inhibition rescued *linc-MD1* knockdown. These in vitro studies were recapitulated in human primary myoblasts, where *linc-MD1* modulated the levels of the MEF2C/MAML1 downstream target MCK in a miRNA level-dependent manner, and partially rescued the delayed differentiation seen in Duchenne myoblasts.	As one of the earliest ceRNA studies, the work of Cesana et al. importantly identified a clear miRNA-mediated interplay between *linc-MD1* and two targets, and showed that this ceRNA regulation also influenced downstream processes via the transcription regulating activity of its direct ceRNA targets. No *in vivo* work was presented, however, leaving the significance of endogenous levels of *linc-MD1* expression unclear. Knock-out of *linc-MD1* in mice, which would perhaps result in a dystrophic-like delay or absence of myogenic differentiation, would confirm the relevance of the lncRNA in development *in vivo*. Likewise, mutation of the putative causative MRE elements for miR-133 and miR-135 would indicate the specific relevance of the ceRNA relationship characterised so clearly *in vitro*. Denzler et al. posited that the high 3' complementarity involved in the causative linc-MD1 MREs implicates target-directed miRNA degradation, not titration[122]. However perturbation of linc-MD1 does not alter miRNA stability (Professor Irene Bozzoni, personal communication).
***Braf-rs1* / *BRAFP1*[152]**	
Following their *in vitro* studies of pseudogenes and ceRNAs in cancer[117], Karreth et al. sought to further investigate the roles of such transcripts *in vivo*. Preliminary *in vitro* work showed that modulation of the pseudogenes of *BRAF* in both mouse (*Braf-rs1*) and human (*BRAFP1*) led to concerted changes in the RNA and protein expression of *Braf*/*BRAF*, and that the subsequent effect upon cell proliferation was mediated by a miRNA-dependent effect upon expression of the parent gene and subsequent MEK/ERK signalling. Further to this, miRNAs for which the transcripts shared MREs were identified, and their ability to mediate crosstalk between *BRAF*/*Braf* and *BRAFP1*/*Braf-rs1* was validated using luciferase reporters. *Braf-rs1* was then investigated *in vivo* using a Dox-inducible overexpression mouse model. Induction of *Braf-rs1* led to the extensive formation of lymphoma in spleens and other organs; this was dependent upon *Braf-rs1* overexpression since withdrawal of Dox led to reduction of *Braf* expression and of activated ERK, leading to tumour regression. A MEK inhibitor also antagonised the proliferation phenotype. Interestingly, inducible expression of only the 3' UTR of *Braf-rs1* largely recapitulated the phenotype following full length *Braf-rs1* overexpression, while the CDS had a limited effect.	The relevance of this study to normal physiology is unclear, but in satisfying the aim of the study to investigate the role of pseudogenes in disease progression, the overexpression of *Braf-rs1 in vivo* was shown to be sufficient in eliciting a lymphomatic phenotype that appeared to be dependent upon MEK/ERK signalling. This is relevant *in vivo* owing to the observation that *BRAFP1* is upregulated in some human cancers, including by genomic amplification, suggesting that this has a cancer-driving effect in at least some cases. However, the *in vivo* phenotype was not extensively proven to be ceRNA mediated, and the overexpression of an MRE mutant transcript would add much in this respect. With regards to non-cancer physiology, knockdown of *Braf-rs1* and associated ceRNA mechanistic work would indicate whether the transcript plays a role in promoting normal cell proliferation.

Table 1.1: Examples of ceRNAs characterised largely *in vitro* and their different approaches to indicating *in vivo* relevance, part 1. Evidence presented to support the functional role of the ceRNA (left); and comments on the applicability of the study to *in vivo* physiology, and further work required to evidence a functional role *in vivo*.

Investigation undertaken and evidence presented	Further work required
***LncND*[153]**	
As shown by Rani et al., the primate-specific lncRNA *LncND* has undergone a genomic insertion in the hominoid lineage that contains 13-18 MREs for four miRNAs. Repressive binding of two of these miRNAs to the *LncND* transcript as well as the putative targets *NOTCH-1* and *NOTCH-2* was confirmed by luciferase assay. Reciprocal relationships between the levels of *LncND*, *NOTCH-1*, and *NOTCH-2* and the level of miR-143-3p were demonstrated, as was the associated destabilisation of luciferase reporters, and the ability of *LncND* modulation to affect levels of downstream Notch signalling targets *HES1* and *HEY1*. *LncND* expression was found to be elevated during neural differentiation, before dropping in neurons, indicating a role in regulating this process, which was consistent with its putative regulation of Notch signalling. Knock-down of *LncND* was found to induce differentiation of SH-SY5Y cells, and this was reproduced by knock-down of *NOTCH-1* and by overexpression of miR-143-3p. RNA-seq analysis of SH-SY5Y cells following *LncND* knockdown or miR-143-3p overexpression indicated a strong and significant association between the respectively differentially expressed genes, which were enriched for genes involved in Notch signalling and cell cycle regulation. Single-cell RNA-seq of cerebral organoid cultures demonstrated that *LncND* is expressed within a small population whose expression profile is consistent with radial glial cells, and this was further supported by ISH in human embryonic neocortical layers. Finally, L*ncND* was overexpressed in radial glial cells of the embryonic mouse by *in utero* electroporation, and was found to stimulate the expansion of this cell population while repressing proliferation of cell identities antagonised by Notch signalling.	This expansive study provides thorough evidence for a miRNA-dependent regulatory relationship between *LncND* and Notch signalling, via miR-143-3p, possibly amongst other miRNAs. The expression pattern and putative function of the lncRNA correlate with the known functions of Notch signalling, and modulated output of this important signalling pathway was demonstrated following *LncND* modulation. Unfortunately, the potential for *in vivo* study is complicated by the primate lineage-only nature of this gene: it is not expressed in mouse. Furthermore, the genomic expansion associated with the putative causative MREs is not present even in rhesus macaque. However, the approach taken, in exogenously applying *LncND* to mouse neocortex *in utero* and taking advantage of the evolutionary conservation of other signalling circuitry, is still informative, and strongly supports the conclusion that *LncND* modulates Notch signalling-mediated regulation of neural differentiation pathways. Use of MRE-mutant transcripts, or combinatorial treatment with miRNA mimics and inhibitors would further indicate the relevance of the ceRNA mechanism that is highly evidenced *in vitro*. Owing to the evolutionary history of the gene, a true *in vivo* genome editing study would require exceptional invasive studies in a hominoid primate. However, identification of other putative miR-143-3p-mediated ceRNAs that are more extensively conserved may more readily permit *in vivo* work.
***lnc-mg*[154]**	
Zhu et al. used conditional knock-out mice to investigate the lncRNA *lnc-mg*, which is endogenously expressed in skeletal muscle. The absence of *lnc-mg in vivo* leads to muscle atrophy and weakened performance, while hypertrophy results from overexpression, leading to the hypothesis that it might regulate differentiation of muscle stem cells, which was demonstrated *in vitro*. A ceRNA mechanism was supported by *in vitro* data that indicated a reciprocal relationship between the expression of *lnc-mg* and miR-125b as well as physical interaction between the two RNAs, demonstrated using luciferase reporter assays alongside miRNA mimics and inhibitors as well as biotinylated miRNA pulldown. Furthermore the expression of Igf2, a key myogenic regulator that was previously shown to mediate the repression of myoblast differentiation by miR-125b[155], was modulated in a manner dependent upon the *lnc-mg*:miR-125b interaction.	Frustratingly, however, given this promising evidence of physiologically relevant ceRNA activity, and the construction of the knock-out mouse lines, only gene expression relationships were assayed *in vivo*, with no direct mechanistic work. Application of either exogenous wildtype and MRE-mutant transcripts into the knock-out mouse system, or ideally genome editing of the putative causative MRE *in vivo*, would have illustrated that the ceRNA mechanism is directly crucial for the proposed function, and that the RNA is therefore able to mediate rescue of the phenotype in an MRE-dependent manner.

Table 1.1: Examples of ceRNAs characterised largely *in vitro* and their different approaches to indicating *in vivo* relevance, part 2. Evidence presented to support the functional role of the ceRNA (left); and comments on the applicability of the study to *in vivo* physiology, and further work required to evidence a functional role *in vivo*.

A study of the circular RNA (circRNA) *Cdr1as* (cerebellar degeneration-related protein 1 antisense transcript) by Piwecka et al.[156] originally seemed a promising line of evidence for *in vivo* miRNA decoy-mediated functionality. The sequence and expression of *Cdr1as* are highly conserved in mammals, including its enrichment in the central nervous system (CNS). Work *in vitro* identified that the circular ~3 kb sequence of *CDR1AS*/*Cdr1as* contains over 70 partial binding sites for miR-7, as well as one site of nearly complete complementarity to miR-671, and is therefore capable of acting as a ceRNA 'sponge' for miR-7[157,158]. Using AGO CLIP-seq data[159,160], Piwecka et al.[156] identified *Cdr1as* as the transcript most associated with miR-7 in the human and mouse brain, followed in the mouse brain by the previously characterised miR-7-binding lncRNA *Cyrano*[161]. FISH indicated that *Cdr1as* is predominantly highly expressed in excitatory neurons. Following careful analysis of the genomic locus to exclude the possibility of perturbing other transcripts, CRISPR/Cas9 was used to delete *Cdr1as* in mice. This resulted in downregulation of both miR-7 species (miR-7a-5p and miR-7b-5p), and upregulation of miR-671. The highly specific post-transcriptional nature of the *Cdr1as* knockout effect was confirmed by unchanged levels of passenger strands for these miRNAs, with the few other changes appearing to be transcriptional downstream effects, and was localised to the tissues where *Cdr1as* expression was highest in the CNS. MiR-7 and miR-671 are therefore highly specifically dysregulated by *Cdr1as* deletion[156]. This was accompanied by perturbation of known miR-7 targets including immediate-early genes such as *Fos*, which are involved in rapid response to stimuli and act as neuronal activity markers, importantly at both the RNA and protein level. This suggested that neuronal activity is elevated in *Cdr1as* knock-out mice, and this was confirmed by electrophysiology and behavioural studies that identified dysfunctional excitatory synapse transmission and strongly reduced pre-pulse inhibition of the startle response reminiscent of the sensorimotor gating deficit associated with neuropsychiatric disorders[156]. Given that the post-transcriptional expression of miR-7 was *decreased* upon loss of Cdr1as, and levels of targets such as *Fos increased*, the relationship *in vivo* is not that of a canonical ceRNA, in which the long RNA would be expected to antagonise miRNA levels or activity and produce concerted changes in other targets. The authors proposed a model whereby *Cdr1as* – which author Rajewsky likens to a "boat"[162] – binds and stably sequesters miR-7, preventing its degradation until binding of miR-671 results in cleavage of the circular RNA transcript and release of the miR-7 "cargo".

However, a recent work continuing the investigation into the relationship between *Cdr1as* and miR-7 has re-addressed these conclusions. It was shown by Kleaveland et al.[163] that levels of miR-7 in the brain are

tightly regulated by the lncRNA *Cyrano*, which was originally noted as having a highly conserved and complementary binding site for miR-7[161]. Knock-down of *cyrano* in zebrafish using morpholinos (phosphorodiamidate oligomer mimics) was previously deemed to impede correct embryogenesis[161], but this phenotype has since been attributed to off-target effects[164], and indeed the CRISPR/Cas9 deletion of *Cyrano* in mice by Kleaveland et al. did not result in an overt phenotype[163]. It did, however, result in a ~45-fold post-transcriptional upregulation of miR-7 levels, shown to be due to the loss of *Cyrano*'s highly efficient target-directed miRNA degradation (TDMD)[165] of miR-7. This resulted in a ~10-fold downregulation of *Cdr1as*, mediated by the increased miR-7 levels, as shown by the insensitivity of *Cdr1as* levels to *Cyrano* deficiency in a miR-7 knock-out background. While miR-7 is not expected to mediate cleavage of *Cdr1as* itself, a miR-671 site mutation illustrated that it does stimulate 3- to 5-fold increases in *Cdr1as* cleavage following miR-671 binding (as well as a further 2- to 3-fold decreased expression through another unidentified miR-7-dependent, miR-671-independent mechanism)[163]. Kleaveland et al. questioned the importance of the small fold-changes observed in miR-7 targets by Piwecka et al. following *Cdr1as* knock-out (<2-fold RNA levels)[156] given that miR-7 knock-down (>5-fold) elicited no de-repression of targets in the regions of the brain where *Cdr1as* was deemed active (e.g. cerebellum)[163]. However, they did not examine protein levels: Piwecka et al. showed substantially larger effects on Fos protein expression than on RNA levels[156]. Indeed previous putative ceRNA studies have identified amplified changes at the level of translation despite unaltered miRNA and mRNA expression[153,166]. The low level of general target perturbation following miRNA loss led the authors to extend the previous hypothesis that *Cdr1as* acts to protect or carry miR-7, suggesting that it might act to deliver miR-7 to target specific mRNAs at, for example, synapses: both *Cdr1as*[163] and miR-7[133,134] are enriched within synaptosomes, and miR-7 has multiple conserved targets involved in vesicle trafficking[167].

These *in vivo* data do not fully support a role for *Cdr1as* as a traditional ceRNA as proposed by Salmena et al.[89], but the notion that binding of miRNAs to 'decoy' transcripts may have a role other than inhibiting their activity, such as spatiotemporal transportation, may be interpreted as an extension of the ceRNA hypothesis: competition for binding in one sub-cellular compartment and subsequent re-release in another would layer further complexity upon the regulation of miRNAs' canonical mRNA targets *in vivo*. Deeper study of ceRNAs *in vivo* may identify more examples in which a competing transcript's role is more than globally reducing miRNA activity.

The conserved regulation of miR-7 by *Cyrano* is supported by a recent study of zebrafish brain morphogenesis, in which exogenous *cyrano* was shown to rescue defective development caused by miR-7 mimic overexpression in an MRE-dependent manner[168]. However, the authors perhaps erroneously concluded that the phenotype of miR-7 overexpression was mediated by loss of *cyrano*, hence the *cyrano*-mediated rescue[168]. It seems more likely that the phenotype resulted from increased repression of a miR-7 target, which was relieved by the additional *cyrano*-mediated binding and degradation of miR-7.

While most circRNAs do not contain more MREs than expected[158,169-171], several other miRNA-binding circRNAs ceRNA candidates have been identified: *ZNF91* (zinc finger 91) contains 39 MREs for miR-296 and 24 MREs for miR-23[170]; and a circRNA form of murine *Sry* specific to the testis contains 19 miR-138 sites[157], though only one is conserved in humans[170]. Careful *in vivo* dissection of these circRNAs will be required to assess the validity of these transcripts' proposed ceRNA activity.

1.3.3 Other cytoplasmic functions of characterised lncRNAs

1.3.3.1 Target-directed miRNA degradation (TDMD)

Genome-wide miRNA stability ranges widely, with multiple studies illustrating concordant ranges of half-lives between <1 hour and >48 hours[172-174], and the individual regulation of each species remains largely unknown. Another mechanism, like the ceRNA mechanism, modelling the regulation of miRNAs by their targets is the target-directed miRNA degradation (TDMD) mechanism[165] as described above for the highly conserved lncRNA *Cyrano*[163]. The binding of miRNAs with high complementarity to a target promotes them for degradation[175] by trimming and 3' tailing[176], and AGO unloading[177]. Ghini et al.[178] identified the highly complementary binding of miR-30b/c to the mRNA *Serpine1* as a model in which to study the role of TDMD in modulating levels of these miRNAs. CRISPR/Cas9 mutation of the predicted miR-30b/c-binding MRE at the endogenous *Serpine1* locus resulted in upregulation of miR-30b/c in a post-transcriptional manner consistent with reduced 3' tailing and dysregulation of miR-30b/c targets[178]. This interaction was found to be dependent upon the substantial 3' complementarity found in the MRE in addition to the seed, suggesting that only highly complementary matches promote TDMD. While 3' complementarity was found in ~15% of TargetScan predicted MREs, only ~0.8% possessed levels comparable to the *Serpine1*:miR-30c interaction[178],

suggesting that this mechanism, while potent, occurs rarely, though cell type-specific differences were noted with primary neurons being especially active in TDMD[175]. This is especially so due to the additional requirement of stoichiometry: like the ceRNA mechanism, TDMD was proposed to occur in a target:miRNA ratio-dependent manner, here with approximately equal stoichiometry necessary and an elevated ratio increasing favourability[178]. The sharply reduced TDMD activity observed in a 200 nt conserved region of murine *Cyrano* relative to the entire transcript[163] suggests that, as with ceRNAs' MREs, context is highly important. The zebrafish lncRNA *linc-epb4.1l4* was highlighted as containing areas of strict conservation in the same manner as *cyrano*[161]. Renamed *libra*, it has recently been shown to direct degradation of miR-29b via a highly complementary binding site[179,180]. Interestingly, *libra* shares homology with the 3'-UTR of the mammalian conserved protein-coding gene *Nrep*. *libra*[179] and *Nrep*[181] are highly enriched in the cerebellum, specifically the granule cell layer in the case of mouse. Within the cerebellum, miR-29b, on the other hand, is expressed in Purkinje cells, mutually exclusive of *libra*/*Nrep*[179]. The dependence of these mutually exclusive patterns upon *Nrep* expression in the mouse is strikingly shown by the mis-localisation of miR-29b expression upon mutation of the miRNA binding site in the endogenous *Nrep* transcript, which was shown to induce 3' trimming of the miRNA when present[179]. Mutation of *libra*/*Nrep* was associated with behavioural and motor coordination dysfunction in both zebrafish and mouse[179], reflecting the crucial roles of the cerebellum in both movement and behavioural processing.

Though Ghini et al.[178] make a clear distinction between the TDMD and ceRNA mechanisms, and it may be that each is relatively rare *in vivo* owing to the need for specific stoichiometric/binding relationships, interaction between targets or mediators of each mechanism is possible, as demonstrated by the *Cyrano*-miR-7-*Cdr1as* network[163]. Akin to the way in which different miRNAs might mediate ceRNA interactions during different states involving altered ranges of ceRNA or miRNA levels[105], such concentration differences may direct differential longevity of miRNAs as well as long RNAs.

1.3.3.2 Regulation of mRNA stability and translation

LncRNAs may interact directly with protein-coding mRNA transcripts. *BACE1-AS*[182] is transcribed antisense to *BACE1* (β-site APP-cleaving enzyme), which is responsible for the production of amyloid β fragments 1-40 and 1-42 (aβ 1-40, aβ 1-42) from amyloid precursor protein (APP). The involvement of aβ 1-42, particularly, in the pathophysiology of Alzheimer's disease have implicated BACE1 as a possible aetiological

component. Ectopic perturbation of *BACE1-AS in vitro* results in concerted concentration-dependent changes in *BACE1* mRNA and protein levels; furthermore, the loss of BACE1 expression following *BACE1-AS* knockdown resulted in reduced production of aβ 1-40 and aβ 1-42[182]. A stabilising duplex between a ~100 nt region of complementarity increases the longevity of *BACE1* mRNA in the cytoplasm[182]. Interestingly, this was partly due to occlusion of a miR-485-5p binding site that represses *BACE1* expression[183]. It was proposed that the duplex-mediated regulation of *BACE1* by *BACE1-AS* might represent a general mechanism of action for antisense RNA-mediated regulation, due to the significant depletion of predicted MREs within overlapping regions for 391 sense-antisense pairs with at least 25 nt of overlap[183]. Ectopic application of aβ 1-42 oligomers increased expression of both *BACE1-AS* and *BACE1*, implicating a positive feedback loop; it is tempting to speculate that this contributes to Alzheimer's pathogenesis. Indeed expression of *BACE1-AS*, and to a lesser extent *BACE1*, is elevated across multiple brain regions in Alzheimer's disease patients, as well as in aβ 1-42-overexpressing mice carrying mutant APP[182]. Additionally, levels of miR-485 are reduced, suggesting a multifaceted mechanism underlying the increased pathophysiological expression of *BACE1*[183].

The lncRNA *NORAD* (non-coding RNA activated by DNA damage)[184] modulates genome stability by interaction with the Pumilio (PUM) family of proteins[184,185], RNA binding proteins that act as negative repressors of gene expression for which it contains ~17 binding sites. Depletion of NORAD *in vitro* leads to chromosomal instability, with erroneous events during anaphase leading to aneuploidy[184]. Binding of *NORAD* to PUM2 is mediated by and dependent upon another cell cycle-regulated RNA-binding protein SAM68[186]. The increased availability of PUM1/2 following *NORAD* loss leads to decreased expression of over 100 target genes, enriched for components of the cell cycle, mitosis, and DNA repair and replication[184]. It was shown in two human cell lines that the level of *NORAD*, which is upregulated ~2-fold following DNA damage, and its capacity for PUM binding would enable it to sequester a large fraction of the highly specific[187] PUM proteins within the cell[184,185]. In mice, *PUM1* haploinsufficiency leads to spinocerebellar ataxia-like neurodegeneration, owing to dysregulated expression of ATXN1[188], suggesting that maintaining correct levels of these proteins is essential[184]. NORAD is therefore the first example of a lncRNA required for correct regulation of targets of an RNA-binding protein at physiological levels[185].

One special class of lncRNAs that regulate translation do so via SINE (short interspersed nuclear element) retrotransposons. So-called "SINEUPs" (SINE elements upregulating translation)[189] are typified by the antisense lncRNA *AS-Uchl1*, which was shown to upregulate translation of Uchl1 protein via two domains:

a targeting domain that binds via complementarity to *Uchl1* mRNA, and an effector domain that contains an inverted SINE2B repeat[190]. Through an as-yet not understood mechanism, this targets the RNA to polyribosomes[190,191]. A number of other antisense transcripts that fit this general structure have been identified[192], including the SINEB2-containing *AS-Uxt*[190] and *PPP1R12A-AS1*, which instead contains a free right Alu monomer repeat element (FRAM), suggesting a more general mechanism of targeting by transposable elments[192]. Mutational analysis of the SINE2B element from *AS-Uchl1* indicates that both sequence and structural domains are required in effecting translational enhancement[193]. Interestingly from a translational perspective, SINEUPs based upon a miniaturised *AS-Uchl1* were simply designed to target other mRNAs including endogenous and exogenous genes, by alteration of the targeting domain[189-191,194], suggesting a novel effective method of upregulating gene expression to modulate pathological phenotypes. Indeed, a SINEUP designed to target *cox7B* in a morpholino-induced medaka fish model of MLS syndrome restored cox7B proteins in a dose-dependent manner, while rescuing normal levels of cell death[195]. Increased body weight following injection into mouse zygotes of a SINEUP (referred as an eRNA) targeting growth hormone is another tantalising hint that SINEUP technology might be used successfully *in vivo*[196]. Unfortunately, effects upon RNA levels and other possible side-effects of perturbation were not assayed, precluding the unequivocal identification of causal SINEUP activity[193]. Nevertheless, this technology represents a promising advancement in the development of molecular therapies and molecular biology tools such as antibodies[197].

1.3.3.3 Regulation of protein modification

LincRNA-p21 has been reported to modulate the degradation-promoting ubiquitination of the hypoxia master regulator HIF-1α by preventing its interaction with the ligase VHL[198]. *LincRNA-p21* is induced by hypoxia, partly through action of HIF-1α, creating a feedback loop that promotes hypoxic responses[198].

1.3.3.4 LncRNAs as precursors to other molecules

The lncRNA *H19* was originally identified as a developmentally regulated RNA[199-201], expressed exclusively from the maternal allele[202]. It was inferred from the dysregulation of the neighbouring paternally imprinted gene *IGF2* (insulin-like growth factor 2) in an early deletion mouse model that H19 was responsible for

establishing imprinting[203]. However, subsequent deletions revealed that the H19 locus and transcript were dispensable: a sequence upstream of *H19* is methylated on the paternal chromosome, preventing binding of the insulator CTCF that would prevent enhancement of Igf2 expression from a downstream enhancer whilst on the maternal chromosome this limits the effect of the enhancer to *H19*[204-206]. In the 20 years since, a large number of functions or interactions have been ascribed to the RNA[207-213], largely *in vitro* with little evidence of physiological relevance, including numerous putative ceRNA interactions[166,214-222]. One clear function, however, is that *H19* acts as the primary transcript from which the evolutionarily conserved miRNAs miR-675-3p/5p are synthesised[223,224]. Reintroduction of miR-675 rescues the impaired myogenic capacity of H19 knockout mice, suggesting a major function of H19 as a pri-miRNA[211]. The previously discussed ceRNA *linc-MD1* is also responsible for generation of miR-206 and miR-133b[149].

Finally, some lncRNAs may in fact encode small or sparsely translated proteins[225]. These will be addressed in detail in Chapter 4.

1.3.4 Multiple lncRNAs may co-participate in complex regulatory networks

Following the identification of *PTENpg1* as the first putative mammalian ceRNA[85], it was suggested that its low expression relative to PTEN would limit its activity as a ceRNA[226]. However, PTEN is a highly dose-sensitive haploinsufficient tumour suppressor, and a reduction of only 20% has been shown to strongly promote cancer formation[227]. However, as well as being the first mammalian ceRNA to be identified, *PTENpg1* (*PTEN* pseudogene 1, also known as *PTENP1*) has been shown to participate in a network of lncRNAs that regulate *PTEN*[228]. Two antisense transcripts, *PTENpg1-asRNA α* and *β*, were found to be divergently transcribed at the *PTENpg1* locus, and to influence PTEN expression through distinct mechanisms: *PTENpg1-asRNA α* acts directly at the *PTEN* locus, where it induces repressive H3K27 methylation via recruitment of the methyltransferases DNMT3a and EZH2[228]. Furthermore, this activity was later discovered to be dependent upon another lncRNA, a 5'-UTR-associated promoter-spanning transcript with which *PTENpg1-asRNA α* interacts in order to tether the repressive activity of DNMT3a[229]. *PTENpg1-asRNA β*, forms an RNA duplex with *PTENpg1*, stabilising it and permitting its accumulation in the cytoplasm[228]. It thereby augments the ceRNA activity of *PTENpg1*. The participation of multiple transcripts in concerted regulation likely explains the failure of exogenous expression of *PTENpg1* to fully restore

wildtype levels of *PTEN* in carcinoma cells[230], and suggests that future ceRNA investigations of the ceRNA potential of *PTENpg1* should carefully assess the contribution of all transcripts involved. Likewise, the architecture of novel ceRNA loci should be closely assessed prior to drawing supportive conclusions.

1.3.5 Organellar lncRNAs

Some lncRNAs are enriched within specific organelles such as mitochondria[231], including at least nine encoded by the mitochondrial genome. Three mitochondrial lncRNAs are produced directly antisense to the coding mRNAs *ND5*, *ND6*, and *CYTB*[128,232], at levels 1.4- to 6-fold lower than their coding partners, to which they bind to form duplexes, possibly regulating expression[232]. Indeed, *ND6* is the only non-polyadenylated mitochondrial mRNA[233], and the binding of *lncND6* might assuage this[232]. At least one other mitochondrial lncRNA incorporates the antisense sequence of a coding gene: *LIPCAR* (long intergenic noncoding RNA predicting cardiac remodelling)[234,235], identified as a serum biomarker following myocardial infarction[235,236], is a chimeric lncRNA containing a 5' sequence antisense to *CYTB* and a 3' end sequence antisense to *COX2*. Two lncRNAs are comprised of the region previously considered non-coding, the D-loop: *MDL1* and *MDL1AS*[237]. Three lncRNAs are formed by an as-yet unknown mechanism – possible trans-splicing or RNA ligation – that covalently links sense and antisense products of the mitochondrial rRNA gene[238-241]. All three (*SncmtRNA*, *ASncmtRNA-1, ASncmtRNA-2* – sense/a*ntisense* mitochondrial ncRNAs*)* contain degradation-resistant hairpins created by inverted repeats[239,240], possibly conferring stability owing to the lack of 5' caps in mitochondria[233]. Interestingly, all three are expressed only in proliferating cells[241], and the two antisense lncRNAs are strongly downregulated in cancer cells and tumours *in vivo*[241-243]. Owing to the observation that knockdown of these transcripts potentiates apoptosis in cancerous but not healthy cells[243], they have been proposed as unique mitochondrial tumour suppressors. One is also upregulated in the aortas of ageing mice and in senescent endothelial cells; overexpression *in vitro* induces G2-phase cell cycle arrest[244].

Noh et al. demonstrated that the nuclear-encoded *RMRP* lncRNA is exported to the mitochondria[245]. *RMRP* knockdown impairs mitochondrial respiration[245]. Accumulation in the mitochondrial matrix is dependent upon the mitochondrial RNA binding protein GRSF1, also required for the stability of mitochondrial mRNAs and lncRNAs[246] in mitochondrial RNA granules[247]. The transport of mitochondrial lncRNAs to[242] and from[245] the nucleus may represent anterograde/retrograde signalling between these major hubs of regulation[242,248].

AIMS OF THIS THESIS

Characterisation of further individual lncRNAs and especially ceRNAs is essential

As these diverse examples illustrate, lncRNAs utilise interactions with all other major macromolecules in the cell to exert influence over gene expression, cell growth, and development, in both healthy and disease states. Nonetheless, the relevance of many annotated lncRNAs remains unclear. Although large-scale high throughput approaches – such as a CRISPRi (interference) library used to screen over 16,000 lncRNAs for their effect upon cell growth[249] – may identify candidates with putative biological implications, the typical cell type- and tissue-specific expression[7,8,250-256], and more importantly, activity[249] of lncRNAs limits the scope of such *in vitro* screens. Likewise, bioinformatic tools that aim to predict lncRNA function[257-261] are limited by the current well-characterised examples with which to draw parallels. Highly informed, targeted investigations on a transcript-by-transcript basis are required to assess the diverse functionality of lncRNA candidates, especially *in vivo*. This could not be truer than for putative ceRNAs: the *bona fide* physiological evidence in favour of the ceRNA mechanism lags starkly behind over 700 primary research publications discussing ceRNAs largely *in vitro*, despite invocations of pathophysiological states from Parkinson's disease[262-264] to Paulownia Witches' Broom disease[265]. As hypothesised above, physiological ceRNAs might have the greatest potential for appreciable regulation under the circumstances of development, when gene expression programs are highly dynamic.

Aims and thesis structure

In this thesis, I characterise the human lncRNA *CEROX1* (competitive endogenous regulator of oxidative phosphorylation). Drawing upon existing work on its murine orthologue *Cerox1*, and utilising a multidisciplinary approach encompassing *in silico* and experimental analyses, I investigate many facets of this novel human lncRNA.

As will be described in greater detail in Chapter 5, and is communicated through the manuscript in which some of the work of this thesis is published (Sirey et al., 2019[266]), murine *Cerox1* has been demonstrated to act as a ceRNA *in vitro*: by interaction with mmu-miR-488-3p, *Cerox1* post-transcriptionally regulates complexes of the mitochondrial electron transport chain. This thesis therefore examines the hypothesis that *CEROX1* is a conserved and functional lncRNA; specifically that human *CEROX1* has conserved the

functionality of murine *Cerox1*, acting as a ceRNA to regulate the human mitochondrial electron transport chain in a conserved manner.

In **Chapter 3**, the genomic locus, expression, and evolutionary conservation of *CEROX1* are explored. The identification of a highly evolutionarily conserved expression pattern supports a role for the lncRNA in the CNS. The extremely high and spatiotemporally controlled expression of *CEROX1* is consistent with a developmental role in regulating genes of the electron transport chain, which are both highly and stably expressed.

Chapter 4 focuses upon the classification of *CEROX1* as a long non-coding RNA. The assumption that *CEROX1* is indeed non-coding is challenged, using bioinformatic tools and a targeted proteomics approach *in vivo*. Preliminary evidence of conserved translation of an open reading frame within recent primates is presented, informing future study of *CEROX1* of the need to consider its bifunctional molecular nature *in vitro* and putatively *in vivo*.

Chapter 5 provides an overview of the preliminary work undertaken into understanding the role of *CEROX1* in human cells *in vitro*, including initial evidence of its influence upon mitochondrial activity. Difficulties in replicating preliminary data are discussed, and additional experimental approaches are utilised to assess the phenotype of *CEROX1* overexpression.

Chapter 6 begins the consideration of the upstream cellular regulation of *CEROX1*. A transcriptional reporter screen confirms the bidirectional nature of the *CEROX1* promoter, and identifies a putative regulator of *CEROX1* expression *in vitro*.

Finally, in **Chapter 7**, the findings of the four results chapters are used to assemble a model of predicted *CEROX1* functionality. Appraisal of the investigation undertaken within this thesis is followed by perspectives upon future ceRNA research.

REFERENCES

1. Rands, C.M., Meader, S., Ponting, C.P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLOS Genetics* **10**, e1004525 (2014).
2. Gulko, B., Hubisz, M.J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics* **47**, 276 (2015).
3. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics* **49**, 618 (2017).
4. Gulko, B. & Siepel, A. How Much Information is Provided by Human Epigenomic Data? An Evolutionary View. (2018).
5. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
6. Guttman, M., Amit, I., Garber, M., French, C. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223 (2009).
7. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).
8. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-27 (2011).
9. Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199-208 (2015).
10. Hon, C.-C., Ramilowski, J.A., Harshbarger, J., Bertin, N. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199 (2017).
11. Harrow, J., Denoeud, F., Frankish, A., Reymond, A. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biology* **7**(2006).
12. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760-1774 (2012).
13. Mattick, J.S. & Makunin, I.V. Non-coding RNA. *Human Molecular Genetics* **15**, R17-R29 (2006).
14. Swiezewski, S., Liu, F., Magusin, A. & Dean, C. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* **462**, 799 (2009).
15. Houseley, J., Rubbi, L., Grunstein, M., Tollervey, D. & Vogelauer, M. A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster. *Molecular Cell* **32**, 685-695 (2008).
16. Bernstein, H.D., Zopf, D., Freymann, D.M. & Walter, P. Functional substitution of the signal recognition particle 54-kDa subunit by its Escherichia coli homolog. *Proceedings of the National Academy of Sciences* **90**, 5229-5233 (1993).
17. Reeves, M.B., Davies, A.A., McSharry, B.P., Wilkinson, G.W. & Sinclair, J.H. Complex I Binding by a Virally Encoded RNA Regulates Mitochondria-Induced Cell Death. *Science* **316**, 1345-1348 (2007).
18. Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S. *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biology* **14**, R131-R131 (2013).
19. Ma, L., Bajic, V.B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA Biology* **10**, 924-933 (2013).
20. Taft, R.J., Pheasant, M. & Mattick, J.S. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* **29**, 288-299 (2007).
21. Khaitovich, P., Kelso, J., Franz, H., Visagie, J. *et al.* Functionality of Intergenic Transcription: An Evolutionary Comparison. *PLOS Genetics* **2**, e171 (2006).
22. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167 (2006).
23. Prabhakar, S., Noonan, J.P., Pääbo, S. & Rubin, E.M. Accelerated Evolution of Conserved Noncoding Sequences in Humans. *Science* **314**, 786-786 (2006).
24. Ponjavic, J., Ponting, C.P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-65 (2007).
25. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology* **14**, 103 (2007).

26. Marques, A.C. & Ponting, C.P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**, R124 (2009).
27. Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635 (2014).
28. Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L. *et al.* Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biology* **11**, R72 (2010).
29. Nitsche, A., Rose, D., Fasold, M., Reiche, K. & Stadler, P.F. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. *Rna* **21**, 801-812 (2015).
30. Pang, K.C., Frith, M.C. & Mattick, J.S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**, 1-5 (2006).
31. Haerty, W. & Ponting, C.P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* **14**, R49 (2013).
32. Schüler, A., Ghanbarian, A.T. & Hurst, L.D. Purifying Selection on Splice-Related Motifs, Not Expression Level nor RNA Folding, Explains Nearly All Constraint on Human lincRNAs. *Molecular Biology and Evolution* **31**, 3164-3183 (2014).
33. Haerty, W. & Ponting, C.P. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**, 333-46 (2015).
34. Hezroni, H., Koppstein, D., Schwartz, Matthew G., Avrutin, A. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11**, 1110-1122 (2015).
35. Ponting, C.P., Oliver, P.L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629-41 (2009).
36. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol* **14**, 699-712 (2013).
37. Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P. *et al.* Considerations when investigating lncRNA function in vivo. *Elife* **3**, e03058 (2014).
38. Li, L. & Chang, H.Y. Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends in Cell Biology* **24**, 594-602 (2014).
39. Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**, e01749 (2013).
40. Lai, K.-M.V., Gong, G., Atanasio, A., Rojas, J. *et al.* Diverse Phenotypes and Specific Transcription Patterns in Twenty Mouse Lines with Ablated LincRNAs. *PLOS ONE* **10**, e0125522 (2015).
41. Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P. *et al.* Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351**, 329-31 (1991).
42. Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515-26 (1992).
43. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38-44 (1991).
44. Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527-42 (1992).
45. Brown, C.J., Lafreniere, R.G., Powers, V.E., Sebastio, G. *et al.* Localization of the X inactivation centre on the human X chromosome in Xq13. *Nature* **349**, 82-4 (1991).
46. Brown, S.D. XIST and the mapping of the X chromosome inactivation centre. *Bioessays* **13**, 607-12 (1991).
47. Brannan, C.I., Dees, E.C., Ingram, R.S. & Tilghman, S.M. The product of the H19 gene may function as an RNA. *Mol Cell Biol* **10**, 28-36 (1990).
48. Gabory, A., Ripoche, M.A., Yoshimizu, T. & Dandolo, L. The H19 gene: regulation and function of a non-coding RNA. *Cytogenet Genome Res* **113**, 188-93 (2006).
49. Han, P., Li, W., Lin, C.H., Yang, J. *et al.* A long noncoding RNA protects the heart from pathological hypertrophy. *Nature* **514**, 102-106 (2014).
50. Chang, C.-P. & Han, P. Epigenetic and lncRNA regulation of cardiac pathophysiology. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1863**, 1767-1771 (2016).
51. Morkin, E. Control of cardiac myosin heavy chain gene expression. *Microscopy Research and Technique* **50**, 522-531 (2000).

52. Grote, P., Wittler, L., Hendrix, D., Koch, F. *et al.* The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell* **24**, 206-14 (2013).
53. Grote, P. & Herrmann, B.G. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol* **10**, 1579-85 (2013).
54. Khalil, A.M., Guttman, M., Huarte, M., Garber, M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).
55. Szafranski, P., Dharmadhikari, A.V., Brosens, E., Gurha, P. *et al.* Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. *Genome Res* **23**, 23-33 (2013).
56. Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-23 (2007).
57. Cifuentes-Rojas, C., Hernandez, Alfredo J., Sarma, K. & Lee, Jeannie T. Regulatory Interactions between RNA and Polycomb Repressive Complex 2. *Molecular Cell* **55**, 171-185 (2014).
58. Davidovich, C., Zheng, L., Goodrich, K.J. & Cech, T.R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nature Structural & Molecular Biology* **20**, 1250 (2013).
59. Wu, L., Murat, P., Matak-Vinkovic, D., Murrell, A. & Balasubramanian, S. Binding Interactions between Long Noncoding RNA HOTAIR and PRC2 Proteins. *Biochemistry* **52**, 9519-9527 (2013).
60. Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N. *et al.* Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* **329**, 689-693 (2010).
61. Schorderet, P. & Duboule, D. Structural and Functional Differences in the Long Non-Coding RNA Hotair in Mouse and Human. *PLOS Genetics* **7**, e1002071 (2011).
62. Suemori, H. & Noguchi, S. Hox C Cluster Genes Are Dispensable for Overall Body Plan of Mouse Embryonic Development. *Developmental Biology* **220**, 333-342 (2000).
63. Li, L., Liu, B., Wapinski, Orly L., Tsai, M.-C. *et al.* Targeted Disruption of Hotair Leads to Homeotic Transformation and Gene Derepression. *Cell Reports* **5**, 3-12 (2013).
64. Amândio, A.R., Necsulea, A., Joye, E., Mascrez, B. & Duboule, D. Hotair Is Dispensible for Mouse Development. *PLOS Genetics* **12**, e1006232 (2016).
65. Green, E.L. Genetic and non-genetic factors which influence the type of skeleton in an inbred strain of mice. *Genetics* **26**, 192-222 (1941).
66. Threadgill, D., Dlugosz, A., Hansen, L., Tennenbaum, T. *et al.* Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype. *Science* **269**, 230-234 (1995).
67. Li, L., Helms, J.A. & Chang, H.Y. Comment on "Hotair Is Dispensable for Mouse Development". *PLOS Genetics* **12**, e1006406 (2016).
68. Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S. *et al.* Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Molecular Cell* **70**, 1038-1053.e7 (2018).
69. Fox, A.H., Lam, Y.W., Leung, A.K.L., Lyon, C.E. *et al.* Paraspeckles: A Novel Nuclear Domain. *Current Biology* **12**, 13-25 (2002).
70. Fox, A.H., Nakagawa, S., Hirose, T. & Bond, C.S. Paraspeckles: Where Long Noncoding RNA Meets Phase Separation. *Trends in Biochemical Sciences* **43**, 124-135 (2018).
71. Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39 (2007).
72. Nakagawa, S., Naganuma, T., Shioi, G. & Hirose, T. Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J Cell Biol* **193**, 31-39 (2011).
73. Standaert, L., Adriaens, C., Radaelli, E., Van Keymeulen, A. *et al.* The long noncoding RNA Neat1 is required for mammary gland development and lactation. *RNA* **20**, 1844-1849 (2014).
74. Nakagawa, S., Shimada, M., Yanaka, K., Mito, M. *et al.* The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a subpopulation of mice. *Development* **141**, 4618-4627 (2014).
75. Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182 (2010).
76. Liu, F. Enhancer-derived RNA: A Primer. *Genomics, Proteomics & Bioinformatics* **15**, 196-200 (2017).
77. Paralkar, Vikram R., Taborda, Cristian C., Huang, P., Yao, Y. *et al.* Unlinking an lncRNA from Its Associated cis Element. *Molecular Cell* **62**, 104-110 (2016).

78. Espinosa, Joaquín M. Revisiting lncRNAs: How Do You Know Yours Is Not an eRNA? *Molecular Cell* **62**, 1-2 (2016).
79. Paralkar, V.R., Mishra, T., Luan, J., Yao, Y. *et al.* Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood* **123**, 1927-37 (2014).
80. Hughes, J.R., Roberts, N., McGowan, S., Hay, D. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics* **46**, 205 (2014).
81. Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods* **13**, 74 (2015).
82. Ulitsky, I. & Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26-46 (2013).
83. van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S. *et al.* Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol* **15**, R6 (2014).
84. Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Biaesch, A. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biology* **16**, 20 (2015).
85. Poliseno, L., Salmena, L., Zhang, J., Carver, B. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8 (2010).
86. Wang, J., Liu, X., Wu, H., Ni, P. *et al.* CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* **38**, 5366-83 (2010).
87. Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics* **39**, 1033 (2007).
88. Cazalla, D., Yario, T. & Steitz, J.A. Down-Regulation of a Host MicroRNA by a Herpesvirus saimiri Noncoding RNA. *Science* **328**, 1563-1566 (2010).
89. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P.P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353-8 (2011).
90. Seitz, H. Redefining microRNA targets. *Curr Biol* **19**, 870-3 (2009).
91. Ala, U., Karreth, F.A., Bosia, C., Pagnani, A. *et al.* Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc Natl Acad Sci U S A* **110**, 7154-9 (2013).
92. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-97 (2004).
93. Kim, V.N. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* **6**, 376-85 (2005).
94. Lee, Y., Ahn, C., Han, J., Choi, H. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415-9 (2003).
95. Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. & Hannon, G.J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231-5 (2004).
96. Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-8 (2001).
97. Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T. *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev* **15**, 2654-9 (2001).
98. Behm-Ansmant, I., Rehwinkel, J. & Izaurralde, E. MicroRNAs silence gene expression by repressing protein expression and/or by promoting mRNA decay. *Cold Spring Harb Symp Quant Biol* **71**, 523-30 (2006).
99. Schmitter, D., Filkowski, J., Sewer, A., Pillai, R.S. *et al.* Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res* **34**, 4801-15 (2006).
100. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20 (2005).
101. Schirle, N.T., Sheu-Gruttadauria, J., Chandradoss, S.D., Joo, C. & MacRae, I.J. Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. *Elife* **4**(2015).
102. Smillie, C.L., Sirey, T. & Ponting, C.P. Complexities of post-transcriptional regulation and the modeling of ceRNA crosstalk. *Critical Reviews in Biochemistry and Molecular Biology*, 1-15 (2018).
103. Pinzón, N., Li, B., Martinez, L., Sergeeva, A. *et al.* microRNA target prediction programs predict many false positives. *Genome Res* **27**, 234-245 (2017).
104. Wehrspaun, C., Ponting, C. & Marques, A. Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs. *Front Genet* **5**, 41 (2014).
105. Figliuzzi, M., Marinari, E. & De Martino, A. MicroRNAs as a selective channel of communication between competing RNAs: a steady-state theory. *Biophys J* **104**, 1203-13 (2013).

106. Bosia, C., Sgrò, F., Conti, L., Baldassi, C. *et al.* RNAs competing for microRNAs mutually influence their fluctuations in a highly non-linear microRNA-dependent manner in single cells. *Genome Biology* **18**, 37 (2017).
107. Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P. & Stoffel, M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell* **54**, 766-76 (2014).
108. Broderick, J.A. & Zamore, P.D. Competitive endogenous RNAs cannot alter microRNA function in vivo. *Mol Cell* **54**, 711-3 (2014).
109. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J. *et al.* Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12**, 735-9 (2002).
110. Chang, J., Nicolas, E., Marks, D., Sander, C. *et al.* miR-122, a mammalian liver-specific microRNA, is processed from hcr mRNA and may downregulate the high affinity cationic amino acid transporter CAT-1. *RNA Biol* **1**, 106-13 (2004).
111. Haussecker, D. & Kay, M.A. miR-122 continues to blaze the trail for microRNA therapeutics. *Mol Ther* **18**, 240-2 (2010).
112. Werfel, S., Leierseder, S., Ruprecht, B., Kuster, B. & Engelhardt, S. Preferential microRNA targeting revealed by in vivo competitive binding and differential Argonaute immunoprecipitation. *Nucleic Acids Research* **45**, 10218-10228 (2017).
113. Bosson, A.D., Zamudio, J.R. & Sharp, P.A. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Mol Cell* **56**, 347-59 (2014).
114. Chiu, H.-S., Martínez, M.R., Komissarova, E.V., Llobet-Navas, D. *et al.* The number of titrated microRNA species dictates ceRNA regulation. *Nucleic Acids Research* **46**, 4354-4369 (2018).
115. Mayya, V.K. & Duchaine, T.F. On the availability of microRNA-induced silencing complexes, saturation of microRNA-binding sites and stoichiometry. *Nucleic Acids Research* **43**, 7556-7565 (2015).
116. Luna, J.M., Scheel, T.K., Danino, T., Shaw, K.S. *et al.* Hepatitis C virus RNA functionally sequesters miR-122. *Cell* **160**, 1099-110 (2015).
117. Karreth, F.A., Tay, Y., Perna, D., Ala, U. *et al.* In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147**, 382-95 (2011).
118. Wang, Y., Xu, Z., Jiang, J., Xu, C. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell* **25**, 69-80 (2013).
119. Alvarez-Saavedra, E. & Horvitz, H.R. Many Families of C. elegans MicroRNAs Are Not Essential for Development or Viability. *Current Biology* **20**, 367-373 (2010).
120. Park, C.Y., Choi, Y.S. & McManus, M.T. Analysis of microRNA knockouts in mice. *Human Molecular Genetics* **19**, R169-R175 (2010).
121. Chen, Y.-W., Song, S., Weng, R., Verma, P. *et al.* Systematic Study of Drosophila MicroRNA Functions Using a Collection of Targeted Knockout Mutations. *Developmental Cell* **31**, 784-800 (2014).
122. Denzler, R., McGeary, Sean E., Title, Alexandra C., Agarwal, V. *et al.* Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Molecular Cell* **64**, 565-579 (2016).
123. Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A.D. *et al.* High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nature Methods* **9**, 840 (2012).
124. Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M. *et al.* Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells. *Cell* **134**, 521-533 (2008).
125. Melton, C., Judson, R.L. & Blelloch, R. Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature* **463**, 621 (2010).
126. Kren, B.T., Wong, P.Y.P., Sarver, A., Zhang, X. *et al.* MicroRNAs identified in highly purified liver-derived mitochondria may play a role in apoptosis. *RNA Biology* **6**, 65-72 (2009).
127. Bian, Z., Li, L.-M., Tang, R., Hou, D.-X. *et al.* Identification of mouse liver mitochondria-associated miRNAs and their potential biological functions. *Cell Research* **20**, 1076 (2010).
128. Mercer, Tim R., Neph, S., Dinger, Marcel E., Crawford, J. *et al.* The Human Mitochondrial Transcriptome. *Cell* **146**, 645-658 (2011).
129. Barrey, E., Saint-Auret, G., Bonnamy, B., Damas, D. *et al.* Pre-microRNA and Mature microRNA in Human Mitochondria. *PLOS ONE* **6**, e20220 (2011).
130. Sripada, L., Tomar, D., Prajapati, P., Singh, R. *et al.* Systematic Analysis of Small RNAs Associated with Human Mitochondria by Deep Sequencing: Detailed Analysis of Mitochondrial Associated miRNA. *PLOS ONE* **7**, e44873 (2012).

131. Borralho, P.M., Rodrigues, C.M.P. & Steer, C.J. Mitochondrial MicroRNAs and Their Potential Role in Cell Function. *Current Pathobiology Reports* **2**, 123-132 (2014).
132. Das, S., Ferlito, M., Kent, O.A., Fox-Talbot, K. *et al.* Nuclear miRNA Regulates the Mitochondrial Genome in the Heart. *Circulation Research* **110**, 1596-1603 (2012).
133. Smalheiser, N.R., Lugli, G., Zhang, H., Rizavi, H. *et al.* Expression of microRNAs and Other Small RNAs in Prefrontal Cortex in Schizophrenia, Bipolar Disorder and Depressed Subjects. *PLOS ONE* **9**, e86469 (2014).
134. Siegel, G., Obernosterer, G., Fiore, R., Oehmen, M. *et al.* A functional screen implicates microRNA-138-dependent regulation of the depalmitoylation enzyme APT1 in dendritic spine morphogenesis. *Nature Cell Biology* **11**, 705 (2009).
135. Sambandan, S., Akbalik, G., Kochen, L., Rinne, J. *et al.* Activity-dependent spatially localized miRNA maturation in neuronal dendrites. *Science* **355**, 634-637 (2017).
136. Flores, O., Kennedy, E.M., Skalsky, R.L. & Cullen, B.R. Differential RISC association of endogenous human microRNAs predicts their inhibitory potential. *Nucleic Acids Research* **42**, 4629-4639 (2014).
137. La Rocca, G., Olejniczak, S.H., González, A.J., Briskin, D. *et al.* In vivo, Argonaute-bound microRNAs exist predominantly in a reservoir of low molecular weight complexes not associated with mRNA. *Proc Natl Acad Sci U S A* **112**, 767-772 (2015).
138. Smith, K.N., Starmer, J., Miller, S.C., Sethupathy, P. & Magnuson, T. Long Noncoding RNA Moderates MicroRNA Activity to Maintain Self-Renewal in Embryonic Stem Cells. *Stem Cell Reports* **9**, 108-121 (2017).
139. Broderick, J.A., Salomon, W.E., Ryder, S.P., Aronin, N. & Zamore, P.D. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA* **17**, 1858-1869 (2011).
140. Sætrom, P., Heale, B.S.E., Snøve, J.O., Aagaard, L. *et al.* Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research* **35**, 2333-2342 (2007).
141. Schirle, N.T., Sheu-Gruttadauria, J. & MacRae, I.J. Structural basis for microRNA targeting. *Science* **346**, 608-613 (2014).
142. Chandradoss, Stanley D., Schirle, Nicole T., Szczepaniak, M., MacRae, Ian J. & Joo, C. A Dynamic Search Process Underlies MicroRNA Targeting. *Cell* **162**, 96-107 (2015).
143. Salomon, William E., Jolly, Samson M., Moore, Melissa J., Zamore, Phillip D. & Serebrov, V. Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. *Cell* **162**, 84-95 (2015).
144. Golden, R.J., Chen, B., Li, T., Braun, J. *et al.* An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* **542**, 197 (2017).
145. Quévillon Huberdeau, M., Zeitler, D.M., Hauptmann, J., Bruckmann, A. *et al.* Phosphorylation of Argonaute proteins affects mRNA binding and is essential for microRNA-guided gene silencing in vivo. *EMBO J* (2017).
146. Jee, D. & Lai, E.C. Alteration of miRNA activity via context-specific modifications of Argonaute proteins. *Trends in cell biology* **24**, 546-553 (2014).
147. Loinger, A., Shemla, Y., Simon, I., Margalit, H. & Biham, O. Competition between Small RNAs: A Quantitative View. *Biophysical Journal* **102**, 1712-1721 (2012).
148. Owens, Nick D.L., Blitz, Ira L., Lane, Maura A., Patrushev, I. *et al.* Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Reports* **14**, 632-647 (2016).
149. Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358-69 (2011).
150. Jens, M. & Rajewsky, N. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nature Reviews Genetics* **16**, 113 (2014).
151. Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.-H. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genetics* **42**, 1113 (2010).
152. Karreth, Florian A., Reschke, M., Ruocco, A., Ng, C. *et al.* The BRAF Pseudogene Functions as a Competitive Endogenous RNA and Induces Lymphoma In Vivo. *Cell* **161**, 319-332 (2015).
153. Rani, N., Nowakowski, T.J., Zhou, H., Godshalk, S.E. *et al.* A Primate lncRNA Mediates Notch Signaling during Neuronal Development by Sequestering miRNA. *Neuron* **90**, 1174-88 (2016).
154. Zhu, M., Liu, J., Xiao, J., Yang, L. *et al.* Lnc-mg is a long non-coding RNA that promotes myogenesis. *Nature Communications* **8**, 14718 (2017).

155. Ge, Y., Sun, Y. & Chen, J. IGF-II is regulated by microRNA-125b in skeletal myogenesis. *J Cell Biol* **192**, 69-81 (2011).
156. Piwecka, M., Glažar, P., Hernandez-Miranda, L.R., Memczak, S. *et al.* Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science* **357**(2017).
157. Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384 (2013).
158. Memczak, S., Jens, M., Elefsinioti, A., Torti, F. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333 (2013).
159. Moore, M.J., Scheel, T.K.H., Luna, J.M., Park, C.Y. *et al.* miRNA–target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature Communications* **6**, 8864 (2015).
160. Boudreau, Ryan L., Jiang, P., Gilmore, Brian L., Spengler, Ryan M. *et al.* Transcriptome-wide Discovery of microRNA Binding Sites in Human Brain. *Neuron* **81**, 294-305 (2014).
161. Ulitsky, I., Shkumatava, A., Jan, Calvin H., Sive, H. & Bartel, David P. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* **147**, 1537-1550 (2011).
162. Circular RNA linked to brain function. (press release, Max Delbrück Center for Molecular Medicine, Berlin, 2017). https://www.mdc-berlin.de/news/press/circular-rna-linked-brain-function
163. Kleaveland, B., Shi, C.Y., Stefano, J. & Bartel, D.P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* **174**, 350-362.e17 (2018).
164. Goudarzi, M., Berg, K., Pieper, L.M. & Schier, A.F. Long non-coding RNAs are largely dispensable for zebrafish embryogenesis, viability and fertility. *BioRxiv* (2018).
165. Rüegger, S. & Großhans, H. MicroRNA turnover: when, how, and why. *Trends in Biochemical Sciences* **37**, 436-446 (2012).
166. Kallen, A.N., Zhou, X.B., Xu, J., Qiao, C. *et al.* The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell* **52**, 101-12 (2013).
167. Latreille, M., Hausser, J., Stützer, I., Zhang, Q. *et al.* MicroRNA-7a regulates pancreatic β cell function. *The Journal of Clinical Investigation* **124**, 2722-2735 (2014).
168. Sarangdhar, M.A., Chaubey, D., Srikakulam, N. & Pillai, B. Parentally inherited long non-coding RNA Cyrano is involved in zebrafish neurodevelopment. *Nucleic Acids Research* **46**, 9726-9735 (2018).
169. Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M. *et al.* Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular Cell* **58**, 870-885 (2015).
170. Guo, J.U., Agarwal, V., Guo, H. & Bartel, D.P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biology* **15**, 409 (2014).
171. You, X., Vlatkovic, I., Babic, A., Will, T. *et al.* Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nature neuroscience* **18**, 603-610 (2015).
172. Marzi, M.J., Ghini, F., Cerruti, B., de Pretis, S. *et al.* Degradation dynamics of microRNAs revealed by a novel pulse-chase approach. *Genome Res* **26**, 554-65 (2016).
173. Guo, Y., Liu, J., Elfenbein, S.J., Ma, Y. *et al.* Characterization of the mammalian miRNA turnover landscape. *Nucleic Acids Research* **43**, 2326-2341 (2015).
174. Duffy, Erin E., Rutenberg-Schoenberg, M., Stark, Catherine D., Kitchen, Robert R. *et al.* Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Molecular Cell* **59**, 858-866 (2015).
175. de la Mata, M., Gaidatzis, D., Vitanescu, M., Stadler, M.B. *et al.* Potent degradation of neuronal miRNAs induced by highly complementary targets. *EMBO Rep* **16**, 500-511 (2015).
176. Ameres, S.L., Horwich, M.D., Hung, J.-H., Xu, J. *et al.* Target RNA–Directed Trimming and Tailing of Small Silencing RNAs. *Science* **328**, 1534-1539 (2010).
177. De, N., Young, L., Lau, P.-W., Meisner, N.-C. *et al.* Highly Complementary Target RNAs Promote Release of Guide RNAs from Human Argonaute2. *Molecular Cell* **50**, 344-355 (2013).
178. Ghini, F., Rubolino, C., Climent, M., Simeone, I. *et al.* Endogenous transcripts control miRNA levels and activity in mammalian cells by target-directed miRNA degradation. *Nature Communications* **9**, 3119 (2018).
179. Bitetti, A., Mallory, A.C., Golini, E., Carrieri, C. *et al.* MicroRNA degradation by a conserved target RNA regulates animal behavior. *Nature Structural & Molecular Biology* **25**, 244-251 (2018).
180. de la Mata, M. & Großhans, H. Turning the table on miRNAs. *Nature Structural & Molecular Biology* **25**, 195-197 (2018).

181. Studler, J.-M., Glowinski, J. & Lévi-Strauss, M. An Abundant mRNA of the Embryonic Brain Persists at a High Level in Cerebellum, Hippocampus and Olfactory Bulb During Adulthood. *European Journal of Neuroscience* **5**, 614-623 (1993).
182. Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E. *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* **14**, 723-30 (2008).
183. Faghihi, M.A., Zhang, M., Huang, J., Modarresi, F. *et al.* Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol* **11**, R56 (2010).
184. Lee, S., Kopp, F., Chang, T.-C., Sataluri, A. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69-80 (2016).
185. Tichon, A., Gil, N., Lubelsky, Y., Havkin Solomon, T. *et al.* A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nature Communications* **7**, 12209 (2016).
186. Tichon, A., Perry, R.B.-T., Stojic, L. & Ulitsky, I. SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. *Genes Dev* **32**, 70-78 (2018).
187. Wang, X., McLachlan, J., Zamore, P.D. & Hall, T.M.T. Modular Recognition of RNA by a Human Pumilio-Homology Domain. *Cell* **110**, 501-512 (2002).
188. Gennarino, Vincenzo A., Singh, Ravi K., White, Joshua J., De Maio, A. *et al.* Pumilio1 Haploinsufficiency Leads to SCA1-like Neurodegeneration by Increasing Wild-Type Ataxin1 Levels. *Cell* **160**, 1087-1098 (2015).
189. Zucchelli, S., Fasolo, F., Russo, R., Cimatti, L. *et al.* SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. *Front Cell Neurosci* **9**, 174 (2015).
190. Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A. *et al.* Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**, 454 (2012).
191. Yao, Y., Jin, S., Long, H., Yu, Y. *et al.* RNAe: an effective method for targeted protein translation enhancement by artificial non-coding RNA with SINEB2 repeat. *Nucleic Acids Research* **43**, e58-e58 (2015).
192. Schein, A., Zucchelli, S., Kauppinen, S., Gustincich, S. & Carninci, P. Identification of antisense long noncoding RNAs that function as SINEUPs in human cells. *Scientific Reports* **6**, 33605 (2016).
193. Takahashi, H., Kozhuharova, A., Sharma, H., Hirose, M. *et al.* Identification of functional features of synthetic SINEUPs, antisense lncRNAs that specifically enhance protein translation. *PLOS ONE* **13**, e0183229 (2018).
194. Patrucco, L., Chiesa, A., Soluri, M.F., Fasolo, F. *et al.* Engineering mammalian cell factories with SINEUP noncoding RNAs to improve translation of secreted proteins. *Gene* **569**, 287-293 (2015).
195. Indrieri, A., Grimaldi, C., Zucchelli, S., Tammaro, R. *et al.* Synthetic long non-coding RNAs [SINEUPs] rescue defective gene expression in vivo. *Scientific Reports* **6**, 27315 (2016).
196. Long, H., Yao, Y., Jin, S., Yu, Y. *et al.* RNAe in a transgenic growth hormone mouse model shows potential for use in gene therapy. *Biotechnology Letters* **39**, 179-188 (2017).
197. Sasso, E., Latino, D., Froechlich, G., Succoio, M. *et al.* A long non-coding SINEUP RNA boosts semi-stable production of fully human monoclonal antibodies in HEK293E cells. *mAbs* **10**, 730-737 (2018).
198. Yang, F., Zhang, H., Mei, Y. & Wu, M. Reciprocal Regulation of HIF-1α and LincRNA-p21 Modulates the Warburg Effect. *Molecular Cell* **53**, 88-100 (2014).
199. Pachnis, V., Brannan, C.I. & Tilghman, S.M. The structure and expression of a novel gene activated in early mouse embryogenesis. *The EMBO journal* **7**, 673-681 (1988).
200. Pachnis, V., Belayew, A. & Tilghman, S.M. Locus unlinked to alpha-fetoprotein under the control of the murine raf and Rif genes. *Proc Natl Acad Sci U S A* **81**, 5523-5527 (1984).
201. Davis, R.L., Weintraub, H. & Lassar, A.B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987-1000 (1987).
202. Bartolomei, M.S., Zemel, S. & Tilghman, S.M. Parental imprinting of the mouse H19 gene. *Nature* **351**, 153 (1991).
203. Leighton, P.A., Ingram, R.S., Eggenschwiler, J., Efstratiadis, A. & Tilghman, S.M. Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* **375**, 34 (1995).
204. Ripoche, M.A., Kress, C., Poirier, F. & Dandolo, L. Deletion of the H19 transcription unit reveals the existence of a putative imprinting control element. *Genes Dev* **11**, 1596-1604 (1997).
205. Thorvaldsen, J.L., Duran, K.L. & Bartolomei, M.S. Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. *Genes Dev* **12**, 3693-3702 (1998).

206. Schmidt, J.V., Levorse, J.M. & Tilghman, S.M. Enhancer competition between H19 and Igf2 does not mediate their imprinting. *Proc Natl Acad Sci U S A* **96**, 9733-9738 (1999).
207. Runge, S., Nielsen, F.C., Nielsen, J., Lykke-Andersen, J. *et al.* H19 RNA Binds Four Molecules of Insulin-like Growth Factor II mRNA-binding Protein. *J Biol Chem* **275**, 29562-29569 (2000).
208. Nielsen, J., Kristensen, M.A., Willemoës, M., Nielsen, F.C. & Christiansen, J. Sequential dimerization of human zipcode-binding protein IMP1 on RNA: a cooperative mechanism providing RNP stability. *Nucleic Acids Research* **32**, 4368-4376 (2004).
209. Gabory, A., Jammes, H. & Dandolo, L. The H19 locus: Role of an imprinted non-coding RNA in growth and development. *BioEssays* **32**, 473-480 (2010).
210. Monnier, P., Martinet, C., Pontis, J., Stancheva, I. *et al.* H19 lncRNA controls gene expression of the Imprinted Gene Network by recruiting MBD1. *Proc Natl Acad Sci U S A* **110**, 20693-20698 (2013).
211. Dey, B.K., Pfeifer, K. & Dutta, A. The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. *Genes Dev* **28**, 491-501 (2014).
212. Martinet, C., Monnier, P., Louault, Y., Benard, M. *et al.* H19 controls reactivation of the imprinted gene network during muscle regeneration. *Development* **143**, 962-971 (2016).
213. Zhang, L., Zhou, Y.H., Huang, T.T., Cheng, A.S.L. *et al.* The Interplay of LncRNA-H19 and Its Binding Partners in Physiological Process and Gastric Carcinogenesis. *International Journal of Molecular Sciences* **18**(2017).
214. Li, M., Chen, H., Zhao, Y., Gao, S. & Cheng, C. H19 Functions as a ceRNA in Promoting Metastasis Through Decreasing miR-200s Activity in Osteosarcoma. *DNA and Cell Biology* **35**, 235-240 (2016).
215. Wang, S.-H., Ma, F., Tang, Z.-h., Wu, X.-C. *et al.* Long non-coding RNA H19 regulates FOXM1 expression by competitively binding endogenous miR-342-3p in gallbladder cancer. *Journal of Experimental & Clinical Cancer Research* **35**, 160 (2016).
216. Schultheiss, C.S., Laggai, S., Czepukojc, B., Hussein, U.K. *et al.* The long non-coding RNA H19 suppresses carcinogenesis and chemoresistance in hepatocellular carcinoma. *Cell Stress* **1**, 37-58 (2017).
217. Li, C.-x., Li, H.-g., Huang, L.-t., Kong, Y.-w. *et al.* H19 lncRNA regulates keratinocyte differentiation by targeting miR-130b-3p. *Cell Death &Amp; Disease* **8**, e3174 (2017).
218. Wu, J., Zhao, J., Sun, L., Pan, Y. *et al.* Long non-coding RNA H19 mediates mechanical tension-induced osteogenesis of bone marrow mesenchymal stem cells via FAK by sponging miR-138. *Bone* **108**, 62-70 (2018).
219. Hong, Y., He, H., Sui, W., Zhang, J. *et al.* Long non-coding RNA H19 promotes cell proliferation and invasion by acting as a ceRNA of miR138 and releasing EZH2 in oral squamous cell carcinoma. *Int J Oncol* **52**, 901-912 (2018).
220. Li, J., Huang, Y., Deng, X., Luo, M. *et al.* Long noncoding RNA H19 promotes transforming growth factor-beta-induced epithelial-mesenchymal transition by acting as a competing endogenous RNA of miR-370-3p in ovarian cancer cells. *Onco Targets Ther* **11**, 427-440 (2018).
221. Hu, Q., Yin, J., Zeng, A., Jin, X. *et al.* H19 Functions as a Competing Endogenous RNA to Regulate EMT by Sponging miR-130a-3p in Glioma. *Cell Physiol Biochem* **50**, 233-245 (2018).
222. Qian, B., Wang, D.M., Gu, X.S., Zhou, K. *et al.* LncRNA H19 serves as a ceRNA and participates in non-small cell lung cancer development by regulating microRNA-107. *Eur Rev Med Pharmacol Sci* **22**, 5946-5953 (2018).
223. Cai, X. & Cullen, B.R. The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA* **13**, 313-316 (2007).
224. Keniry, A., Oxley, D., Monnier, P., Kyba, M. *et al.* The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nature Cell Biology* **14**, 659 (2012).
225. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981-93 (2014).
226. Ebert, M.S. & Sharp, P.A. Emerging Roles for Natural MicroRNA Sponges. *Current Biology* **20**, R858-R861 (2010).
227. Alimonti, A., Carracedo, A., Clohessy, J.G., Trotman, L.C. *et al.* Subtle variations in Pten dose determine cancer susceptibility. *Nature Genetics* **42**, 454 (2010).
228. Johnsson, P., Ackley, A., Vidarsdottir, L., Lui, W.-O. *et al.* A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nature Structural & Molecular Biology* **20**, 440 (2013).

229. Lister, N., Shevchenko, G., Walshe, J.L., Groen, J. *et al.* The molecular dynamics of long noncoding RNA control of transcription in PTEN and its pseudogene. *Proc Natl Acad Sci U S A* **114**, 9942-9947 (2017).
230. Liu, J., Xing, Y., Xu, L., Chen, W. *et al.* Decreased expression of pseudogene PTENP1 promotes malignant behaviours and is associated with the poor survival of patients with HNSCC. *Scientific Reports* **7**, 41179 (2017).
231. Dietrich, A., Wallet, C., Iqbal, R.K., Gualberto, J.M. & Lotfi, F. Organellar non-coding RNAs: Emerging regulation mechanisms. *Biochimie* **117**, 48-62 (2015).
232. Rackham, O., Shearwood, A.-M.J., Mercer, T.R., Davies, S.M.K. *et al.* Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA (New York, N.Y.)* **17**, 2085-2093 (2011).
233. Temperley, R.J., Wydro, M., Lightowlers, R.N. & Chrzanowska-Lightowlers, Z.M. Human mitochondrial mRNAs—like members of all families, similar but different. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1797**, 1081-1085 (2010).
234. Dorn, G.W., 2nd. LIPCAR: a mitochondrial lnc in the noncoding RNA chain? *Circulation research* **114**, 1548-1550 (2014).
235. Kumarswamy, R., Bauters, C., Volkmann, I., Maury, F. *et al.* Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. *Circ Res* **114**, 1569-75 (2014).
236. Li, M., Wang, Y.-F., Yang, X.-C., Xu, L. *et al.* Circulating Long Noncoding RNA LIPCAR Acts as a Novel Biomarker in Patients with ST-Segment Elevation Myocardial Infarction. *Medical science monitor : international medical journal of experimental and clinical research* **24**, 5064-5070 (2018).
237. Gao, S., Tian, X., Chang, H., Sun, Y. *et al.* Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. *Mitochondrion* **38**, 41-47 (2018).
238. Villegas, J., Burzio, V., Villota, C., Landerer, E. *et al.* Expression of a novel non-coding mitochondrial RNA in human proliferating cells. *Nucleic Acids Research* **35**, 7336-7347 (2007).
239. Villegas, J., Zárraga, A.M., Muller, I., Montecinos, L. *et al.* A Novel Chimeric Mitochondrial RNA Localized in the Nucleus of Mouse Sperm. *DNA and Cell Biology* **19**, 579-588 (2000).
240. Villegas, J., Araya, P., Bustos-Obregon, E. & Burzio, L.O. Localization of the 16S mitochondrial rRNA in the nucleus of mammalian spermatogenic cells. *MHR: Basic science of reproductive medicine* **8**, 977-983 (2002).
241. Burzio, V.A., Villota, C., Villegas, J., Landerer, E. *et al.* Expression of a family of noncoding mitochondrial RNAs distinguishes normal from cancer cells. *Proc Natl Acad Sci U S A* **106**, 9430-9434 (2009).
242. Landerer, E., Villegas, J., Burzio, V.A., Oliveira, L. *et al.* Nuclear localization of the mitochondrial ncRNAs in normal and cancer cells. *Cellular Oncology* **34**, 297-305 (2011).
243. Vidaurre, S., Fitzpatrick, C., Burzio, V.A., Briones, M. *et al.* Down-regulation of the Antisense Mitochondrial Non-coding RNAs (ncRNAs) Is a Unique Vulnerability of Cancer Cells and a Potential Target for Cancer Therapy. *J Biol Chem* **289**, 27182-27198 (2014).
244. Bianchessi, V., Badi, I., Bertolotti, M., Nigro, P. *et al.* The mitochondrial lncRNA ASncmtRNA-2 is induced in aging and replicative senescence in Endothelial Cells. *Journal of Molecular and Cellular Cardiology* **81**, 62-70 (2015).
245. Noh, J.H., Kim, K.M., Abdelmohsen, K., Yoon, J.-H. *et al.* HuR and GRSF1 modulate the nuclear export and mitochondrial localization of the lncRNA RMRP. *Genes Dev* **30**, 1224-1239 (2016).
246. Antonicka, H., Sasarman, F., Nishimura, T., Paupe, V. & Shoubridge, Eric A. The Mitochondrial RNA-Binding Protein GRSF1 Localizes to RNA Granules and Is Required for Posttranscriptional Mitochondrial Gene Expression. *Cell Metabolism* **17**, 386-398 (2013).
247. Jourdain, Alexis A., Koppen, M., Wydro, M., Rodley, Chris D. *et al.* GRSF1 Regulates RNA Processing in Mitochondrial RNA Granules. *Cell Metabolism* **17**, 399-410 (2013).
248. Dong, Y., Yoshitomi, T., Hu, J.-F. & Cui, J. Long noncoding RNAs coordinate functions between mitochondria and the nucleus. *Epigenetics & Chromatin* **10**, 41 (2017).
249. Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**(2017).
250. Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18**, 1433-45 (2008).
251. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F. & Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**, 716-21 (2008).

252. Mercer, T.R., Dinger, M.E., Mariani, J., Kosik, K.S. *et al.* Noncoding RNAs in Long-Term Memory Formation. *Neuroscientist* **14**, 434-45 (2008).
253. Pang, K.C., Dinger, M.E., Mercer, T.R., Malquori, L. *et al.* Genome-wide identification of long noncoding RNAs in CD8+ T cells. *J Immunol* **182**, 7738-48 (2009).
254. Kaushik, K., Leonard, V.E., Kv, S., Lalwani, M.K. *et al.* Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PLoS One* **8**, e83616 (2013).
255. Tsoi, L.C., Iyer, M.K., Stuart, P.E., Swindell, W.R. *et al.* Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol* **16**, 24 (2015).
256. Clark, M.B., Mercer, T.R., Bussotti, G., Leonardi, T. *et al.* Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat Methods* **12**, 339-42 (2015).
257. Zheng, L.-L., Li, J.-H., Wu, J., Sun, W.-J. *et al.* deepBase v2.0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic acids research* **44**, D196-D202 (2016).
258. Perron, U., Provero, P. & Molineris, I. In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression. *BMC Bioinformatics* **18**, 144 (2017).
259. Zhou, J., Zhang, S., Wang, H. & Sun, H. LncFunNet: an integrated computational framework for identification of functional long noncoding RNAs in mouse skeletal muscle cells. *Nucleic acids research* **45**, e108-e108 (2017).
260. Yang, C., Yang, L., Zhou, M., Xie, H. *et al.* LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, bty428-bty428 (2018).
261. Kirk, J.M., Kim, S.O., Inoue, K., Smola, M.J. *et al.* Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics* **50**, 1474-1482 (2018).
262. Straniero, L., Rimoldi, V., Samarani, M., Goldwurm, S. *et al.* The GBAP1 pseudogene acts as a ceRNA for the glucocerebrosidase gene GBA by sponging miR-22-3p. *Sci Rep* **7**, 12702 (2017).
263. Lin, D., Liang, Y., Jing, X., Chen, Y. *et al.* Microarray analysis of an synthetic alpha-synuclein induced cellular model reveals the expression profile of long non-coding RNA in Parkinson's disease. *Brain Res* **1678**, 384-396 (2018).
264. Sang, Q., Liu, X., Wang, L., Qi, L. *et al.* CircSNCA downregulation by pramipexole treatment mediates cell apoptosis and autophagy in Parkinson's disease by targeting miR-7. *Aging (Albany NY)* **10**, 1281-1293 (2018).
265. Fan, G., Wang, Z., Zhai, X. & Cao, Y. ceRNA Cross-Talk in Paulownia Witches' Broom Disease. *Int J Mol Sci* **19**(2018).
266. Sirey, T.M., Roberts, K., Haerty, W., Bedoya-Reina, O. *et al.* The long non-coding RNA Cerox1 is a post transcriptional regulator of mitochondrial complex I catalytic activity. *eLife* **8**, e45051 (2019).

CHAPTER 2: MATERIALS AND METHODS

TECHNIQUES

2.1 Techniques for cell work

2.1.1 Mammalian cell culture

All mammalian cells were cultured at 37°C in a humidified atmosphere supplemented with 5% CO_2. Cells were cultured in canted-neck cell culture flasks of 25 cm^2, 75 cm^2, and 175 cm^2 area, as well as in multi-well plates (all Corning®).

2.1.1.1 Immortalised cell lines

Human embryonic kidney (HEK293T) cells, human neuroblastoma SK-N-SH cells, and mouse neuroblastoma N2A cells were cultured in high-glucose (22.5 mM) Dulbecco's Modified Eagle Medium (DMEM) with sodium pyruvate (0.9 mM), GlutaMAX™, and phenol red, supplemented with 9% foetal bovine serum (FBS) as well as penicillin (90 U ml^{-1}) and streptomycin (90 µg ml^{-1}).

Human neuroblastoma SH-SY5Y cells were cultured in 1:1 DMEM and Ham's F-12 (DMEM/F-12) with high-glucose (15.8 mM), sodium pyruvate (0.45 mM), GlutaMAX™, and phenol red, supplemented with 9% FBS as well as penicillin (90 U ml^{-1}) and streptomycin (90 µg ml^{-1}).

Samples of the clonal rat pancreatic beta cell line INS-1 were obtained from Dr Laurène Vetterli (Cantley laboratory, DPAG, University of Oxford), who in turn obtained them from their originating laboratory, that of Professor Pierre Maechler (University of Geneva)[1]. Cells were cultured in RPMI 1640 medium supplemented with 4.6% foetal bovine serum, penicillin-streptomycin (92 U ml^{-1}), L-glutamine (1.83 mM), HEPES (9.2 mM), 2-mercaptoethanol (45 µM), and sodium pyruvate (920 mM).

2.1.2 Cell counting

In preparation for seeding cells, cell suspensions were stained with Trypan blue (Thermo Fisher) to a final concentration of 0.2%, and quantified using a C-chip™ haemocytometer (Neubauer improved).

2.1.3 Transfection of exogenous DNA

2.1.3.1 Transient transfection

Transient transfection of cells was carried out using FuGENE®6 Transfection Reagent (Promega), diluted in Opti-MEM™ (Gibco®, Thermo Fisher), according to manufacturer's instructions. A FuGENE®:DNA ratio of 6 µl:1 µg was used when transfecting HEK293T and SH-SY5Y cells with plasmids.

For transfection of LNA™ GapmeRs (Exiqon), which were also transfected using FuGENE®6, 2 µl of 50 µM GapmeR suspension (made up in nuclease-free water) was added per well of a six-well plate, resulting in a concentration of 50 nM in 2 ml of medium.

Cells were plated 24 hours before transfection, and harvested or assayed 24- 72 hours post-transfection.

2.1.4 Drug treatments

2.1.4.1 Proteolysis inhibitors

In order to inhibit proteolysis pathways, proteasome and lysosome activities were inhibited by incubation with 50 µM MG132 or 100 nM bafilomycin, respectively, in normal growth medium for 6 hours.

2.1.4.2 Calcium signalling modulators

In order to investigate the effect of calcium signalling upon *CEROX1* RNA levels, SH-SY5Y cells were treated with 20 nM phorbol 12-myristate 13-acetate (PMA), or 10 µM carbamylcholine chloride (carbachol) in normal growth medium for up to 24 hours.

2.2 Techniques for human and animal tissue work

No animals were terminated for the purpose of this study. All animal samples were obtained commercially, from scientific collaborators, or via material transfer agreement.

Additional health and safety precautions were taken when working with human and primate samples; namely, all work was carried out in a Category 2 biological safety cabinet and all waste was sterilised with 2% Virkon.

2.2.1 Human samples

Samples of human striatum tissue and a diffuse glioma tumour were obtained from the Oxford Brain Bank (Dr Olaf Ansorge, Carolyn Sloan; REC reference 15/SC/0639).

2.2.2 Animal samples

2.2.2.1 Gorilla

A fragment of fresh-frozen tissue from the cerebral cortex of a western lowland gorilla (*Gorilla gorilla gorilla*) was obtained from the International Zoo Veterinary Group pathology laboratory (Dr Mark Stidworthy). Permission for third-party material transfer from the IZVG was kindly granted by Twycross Zoo where the sample originated.

2.2.2.2 Ferret

The brain of a ferret (*Mustela putorius furo*) was obtained from the Department of Pharmacology, University of Oxford, and dissected by Professor Peter Oliver. Fragments of frozen brain tissue were equilibrated in RNAlater®-ICE Frozen Tissue Transition Solution (Ambion™) at -70°C on dry ice, then incubated at -20°C for 24 hours to facilitate absorption prior to long term storage at -80°C.

2.2.3 Sectioning

Mouse tissues and embryos were fresh frozen in optimal cutting temperature compound (OCT) on dry ice. Frozen blocks were sectioned into 15 µm slices using a cryostat and then transferred onto SuperFrost™ Plus slides and stored at -80°C.

Fresh frozen human tissue sections from control striatum and glioma tumour were obtained from the Oxford Brain Bank (Dr Olaf Ansorge).

2.3 Techniques for RNA extraction and experimentation

During all RNA work, care was taken to eliminate RNase contamination, by cleaning all equipment and surfaces with RNaseZap™ RNase Decontamination Solution (Invitrogen, Thermo).

2.3.1 RNA extraction

Cells grown for RNA extraction were washed with PBS, following removal of media. Cells were then re-suspended and lysed in 700 µl TRIzol® Reagent. When RNA isolation was not carried out immediately, lysed samples were stored in TRIzol® Reagent at -80°C in the interim.

RNA was isolated using the miRNeasy Mini Kit (Qiagen), in order to preserve both long and shorter RNA species, including miRNAs. The kit was used according to the manufacturer's instructions, including the optional on-column DNase digest (supplementary manufacturer's protocol), and RNA was eluted using between 30 µl and 50 µl of RNase-free water.

In some cases, a supplementary manufacturer's protocol was used to isolate a miRNA-enriched fraction separate from long RNAs.

RNA concentrations were assayed using a NanoDrop ND-1000 UV-Vis spectrophotometer.

2.3.1.1 RNA extraction from animal tissues

Frozen tissue samples were rapidly homogenised in cold TRIzol® Reagent using a Precellys® Homogenizer with a CryoLys® cooling system (Bertin Instruments) to keep the equipment below 10°C, and RNA was subsequently isolated using the miRNeasy Mini Kit (Qiagen).

2.3.2 cDNA synthesis

cDNA was produced from extracted RNA samples using different reverse transcription kits, depending upon the RNA species being investigated. For general cDNA production, when only long RNAs were to be investigated, the QuantiTect® Reverse Transcription Kit (Qiagen) was used. In cases where miRNAs were to be quantified, the NCode™ VILO™ miRNA cDNA Synthesis Kit (Invitrogen) was used.

Typically, one microgram of total RNA was fed into each reverse transcription reaction. In preparation of quantification of many genes by RT-qPCR – for example, when analysing complex I gene expression changes following *CEROX1* overexpression – multiple cDNA synthesis reactions were set up, and the products pooled. Following cDNA synthesis, the cDNA produced was diluted two- to five-fold with nuclease-free water.

2.3.3 Rapid amplification of cDNA ends (RACE)

Total RNA samples from mammalian brains – human (*Homo sapiens*), rhesus monkey (*Mucaca mulatta*), house mouse (*Mus musculus*), common rat (*Rattus norvegicus*), and domestic horse (*Equus ferus caballus*) – were obtained from AMSbio (Abingdon, Oxfordshire, UK). RNA was also extracted as previously described from ferret and gorilla tissues, as well as various cell lines including HEK293T, SH-SY5Y, and SK-N-SH and primary human fibroblasts. Quality of all RNA samples used for RACE, both commercial and otherwise, was assessed using an Agilent RNA 6000 Nano Kit with the Agilent 2100 Bioanalyzer. RIN scores ranged from 8 to 10 in those samples deemed fit for RACE.

2.3.3.1 RACE with the GeneRacer™ Kit

Initially, RACE was carried out primarily according to the protocol for the GeneRacer™ Kit (Invitrogen).

For 3' RACE, 5 µg of RNA were reverse transcribed with Superscript® III reverse transcriptase (Thermo Fisher), using the GeneRacer™ Oligo-dT Primer (Invitrogen). The GeneRacer™ 3' and nested 3' primers were then used alongside gene-specific primers to amplify the 3' termini of transcripts (see RACE-PCR below for conditions).

For 5' RACE, 5 µg of RNA were processed according to the GeneRacer™ Kit protocol: briefly, total RNA was dephosphorylated with calf intestinal phosphatase (CIP), the 5' cap was removed with tobacco acid pyrophosphatase (TAP), and then the now-accessible end was ligated to an RNA oligonucleotide of known sequence using T4 RNA ligase. Reverse transcription with Superscript® III reverse transcriptase, using a gene-specific primer, was used to produce cDNA. The GeneRacer™ 5' and nested 5' primers were then used alongside gene-specific primers to amplify the 5' termini of transcripts (see RACE-PCR below for conditions).

RACE-PCR products were gel extracted or purified with the QIAquick Gel Extraction Kit or QIAquick PCR Purification Kit (Qiagen), respectively, and then cloned into pGEM-Teasy for Sanger sequencing with M13F/R.

2.3.3.2 Optimised 5' RACE

Upon the failure of 5' RACE using the GeneRacer™ Kit, the protocol of Shi et al.[2] was utilised. Five micrograms of total RNA was reverse transcribed using ThermoScript™ Reverse Transcriptase in the

presence of the synthetic compatible solute homoectoine[2,3] at 500 mM. Homoectoine, which is not commercially available, was kindly donated by Professor Erwin Galinski, of the University of Bonn, Germany. The reverse transcription reaction was primed with a single gene-specific primer at 5 µM or a cocktail of multiple gene-specific primers at 2-3 µM each. Following cDNA synthesis, RNA was degraded by treatment with 0.3M sodium hydroxide at 65°C for 20 minutes. cDNA was then precipitated by the addition of one-ninth volume of 5 M sodium chloride and three volumes of 100% ethanol, at -20°C overnight. Following centrifugation at 16,000 g for 20 minutes at 4°C and washing with 70% ethanol, the cDNA pellet was re-suspended in a solution of pre-annealed 5' adaptor. This was produced from two complementary oligonucleotides, which were heated to 100°C for five minutes at a concentration of 10 µM each in 50 mM sodium chloride, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, and then allowed to cool to room temperature. The adaptor-cDNA mixture was ligated using T4 DNA ligase (NEB) according to the manufacturer's instructions. The ligated mixture was diluted three- or four-fold with nuclease-free water, and then used in RACE-PCR.

2.3.3.3 mRNA enrichment

In order to produce an enriched mRNA fraction for 5' RACE, a protocol was devised to purify mRNAs and lncRNAs (including *CEROX1*) from a mixed total RNA sample via their polyadenylation tails. This protocol was adapted from that provided for the PolyATtract® mRNA Isolation System (Promega). Following denaturation of the RNA at 65°C for ten minutes, 500 pmol of 5' biotinylated dT_{25} was annealed to 2.5 mg of total RNA in 75 mM sodium chloride, 7.5 mM sodium citrate for 30 minutes at room temperature. Fifty microlitres of EZview red streptavidin affinity gel suspension were washed three times with 75 mM sodium chloride, 7.5 mM sodium citrate, and then incubated with the annealing reaction for ten minutes at room temperature. The mixture was washed four times with 15 mM sodium chloride, 1.5 mM sodium citrate, and then the RNA was eluted twice in 100 µl of RNase-free water each.

Due to the variability of the above protocol, mRNA was instead enriched using the GenElute mRNA Miniprep Kit (Sigma), according to the manufacturer's protocol, and then concentrated by precipitation with 20 µg of glycogen, 0.1 volumes of 3 M sodium acetate pH 5.2, and three volumes of ethanol.

2.3.4 *In situ* hybridisation (ISH)

2.3.4.1 Probe generation

In vitro transcription of labelled ribonucleic acid probes was carried out using the Roche DIG RNA Labelling Kit. Using fragments of target genes cloned into pCR™4-TOPO and linearised using NotI or SpeI as a template (1 µg of DNA per *in vitro* transcription reaction), RNA was transcribed by T7 or T3 RNA polymerases, and labelled via the incorporation of digoxigenin-UTP. Resulting RNA was purified using RNA Mini Quick Spin Columns (Roche) according to the manufacturer's protocol. One and a half microlitres of RNA extract was used per section in the subsequent hybridisation protocol.

2.3.4.2 Probe hybridisation

All solutions were made using diethylpyrocarbonate (DEPC)-treated water. Following air drying, sections were fixed in 4% paraformaldehyde for ten minutes, and then washed thrice in PBS for three minutes each. Sections were then subjected to acetylation in 96 mM triethanolamine, 20.8 mM hydrochloric acid, with the gradual addition of acetic anhydride to 23.5 mM, and washed as above. Sections were equilibrated with hybridisation solution (50% formamide, 1 mM EDTA pH 8.0, 100 mg ml^{-1} dextran sulphate, 3 mM sodium chloride, 8.3 mM sodium dodecyl sulphate, 0.02% Ficoll-400, 0.02% bovine serum albumin, 0.02% polyvinylpyrrolidone, 200 µg ml^{-1} *Escherichia coli* tRNA) for one hour at room temperature. Slides were then incubated with probe in hybridisation solution in a humidified box (50% formamide, 750 mM sodium chloride, 75 mM sodium citrate) at 60°C for 16 hours.

2.3.4.3 Anti-digoxigenin probing

Sections were rinsed in 750 mM sodium chloride, 75 mM sodium citrate at 70°C for five minutes, and then washed twice in 30 mM sodium chloride, 3 mM sodium citrate, first at 70°C for one hour, and second at room temperature for five minutes. Another wash for five minutes in 100 mM Tris-HCl pH 7.6, 150 mM sodium chloride was followed by blocking at room temperature for one hour with 10% (w/v) Blocking Reagent (Roche, Sigma) in the same solution. Sections were then incubated at 4°C overnight with sheep anti-digoxigenin-AP conjugate (Roche, Sigma) at 1:5,000 in blocking solution.

2.3.4.4 Chromogenic detection

Sections were washed thrice in 100 mM Tris-HCl pH 7.6, 150 mM sodium chloride for five minutes each, and then equilibrated in 100 mM Tris-HCl pH 9.5, 100 mM sodium chloride, 50 mM magnesium chloride for five minutes. Staining was developed by incubation with 375 µg ml^{-1} nitro blue tetrazolium chloride and 188 µg ml^{-1} 5-bromo-4-chloro-3-indolyl-phosphate in 100 mM Tris-HCl pH 9.5, 100 mM sodium chloride, 50 mM magnesium chloride for 48 to 60 hours. Staining was halted by several washes in PBS. Sections were counterstained with 0.1% Nuclear Fast Red in 5% aluminium sulphate (Sigma) for three minutes, and then washed thoroughly in water before mounting with 90% glycerol.

2.4 Polymerase chain reaction (PCR)

2.4.1 PCR

For routine PCRs, Platinum® DNA Polymerase (Thermo Fisher) was used, according to the manufacturer's instructions. When amplifying difficult amplicons involving *CEROX1*, Phusion® High-Fidelity Polymerase (NEB) was instead used, alongside its component GC buffer and with the addition of 5% DMSO.

Reaction cycling conditions for all polymerases are found below, in Table 2.1.

2.4.1.1 Primers

Primers for RT-qPCR were designed using Primer-BLAST[4], with an optimal melting temperature of 60°C. Other primers for amplification of genomic DNA, cDNA or plasmids, or for mutagenesis, were designed manually. All primer sequences may be found in Appendix A2.

2.4.1.2 Colony PCR

In order to assess bacterial transformants, colonies were screened by PCR using GoTaq® DNA Polymerase (Promega), according to the manufacturer's instructions.

2.4.1.3 Real-time quantitative PCR (RT-qPCR)

Quantification of amplicons was carried out using a StepOnePlus™ Real-Time PCR System, using SYBR Green Master Mix (Applied Biosystems™, Thermo Fisher). For precise quantitation of *CEROX1* expression,

amplicon-sensitive locked nucleic acid probes from the Universal ProbeLibrary (Roche Life Sciences) were used at 100 nM final concentration. Reaction cycling conditions for quantitative PCR are found below, in Table 2.2.

Platinum® DNA Polymerase			
Step	Temperature	Time	Cycles
Initial denaturation	94°C	2 minutes	1
Denaturation	94°C	30 seconds	35
Annealing	55°C – 68°C *	30 seconds	
Extension	68°C	30 seconds – 4 minutes **	
Final extension	68°C	10 minutes	1
Storage	4°C		
Phusion® High-Fidelity Polymerase			
Step	Temperature	Time	Cycles
Initial denaturation	98°C	30 seconds	1
Denaturation	98°C	15 seconds	35
Annealing	55°C – 72°C *	30 seconds	
Extension	72°C	15 seconds – 4 minutes **	
Final extension	72°C	10 minutes	1
Storage	4°C		
GoTaq® DNA Polymerase			
Step	Temperature	Time	Cycles
Initial denaturation	95°C	2 minutes	1
Denaturation	95°C	30 seconds	35
Annealing	55°C – 60°C *	30 seconds	
Extension	72°C	30 seconds – 2 minutes **	
Final extension	72°C	5 minutes	1
Storage	4°C		

Table 2.1: Cycling conditions used for PCR amplification of DNA targets. *Annealing temperatures were dependent upon primers. **Extension times were dependent upon amplicon length, being approximately 1 minute per kb.

SYBR Green Master Mix – AmpliTaq Gold® DNA Polymerase				
Step	Temperature	Time	Cycles	Data collection
Initial denaturation	95°C	10 minutes	1	-
Denaturation	95°C	15 seconds	40	-
Annealing and extension	60°C	1 minute		Amplification
Melt curve	95°C	15 seconds	1	-
	60°C	+ 0.3°C per		Melt curve
	95°C	15 seconds		
SYBR Green Master Mix – AmpliTaq Gold® DNA Polymerase with Universal ProbeLibrary Probes (Roche Life Science)				
Step	Temperature	Time	Cycles	Data collection
Initial denaturation	95°C	10 minutes	1	-
Denaturation	95°C	15 seconds	40	-
Annealing and extension	60°C	1 minute		Amplification

Table 2.2: Cycling and data collection conditions used for real time quantitative PCR.

2.4.1.4 RACE-PCR

As per the instructions of the GeneRacer™ Kit, Platinum® DNA Polymerase (Thermo Fisher) was used to amplify reverse transcribed transcript termini. High melting temperature gene-specific primers were designed and used with touchdown PCR in order to heighten amplification specificity (Table 2.3).

RACE-PCR with Platinum® DNA Polymerase			
Step	Temperature	Time	Cycles
Initial denaturation	94°C	2 minutes	1
Denaturation	94°C	30 seconds	5
Annealing and extension	72°C	1-2 minutes **	
Denaturation	94°C	30 seconds	5
Annealing and extension	70°C	1-2 minutes **	
Denaturation	94°C	30 seconds	25
Annealing	60°C – 70°C *	30 seconds	
Extension	68°C	1-2 minutes **	
Final extension	68°C	10 minutes	1
Storage	4°C		

Table 2.3: Cycling conditions used for touchdown RACE-PCR. *Annealing temperatures were dependent upon primers. **Extension times were dependent upon amplicon length, being approximately 1 minute per kb.

2.5 Cloning and DNA construct preparation

All restriction enzymes, DNA ligase and DNA polymerase I Klenow fragment were obtained from NEB.

2.5.1 Restriction digestion

When cutting DNA sequences, 1 µg of plasmid DNA was typically digested for 15-30 minutes using NEB restriction enzymes, according to the manufacturer's instructions. Where multiple enzyme digestions were required but not compatible in a common reaction, one digestion was followed by purification with the QIAquick PCR Purification Kit prior to the second digestion in an alternative buffer.

2.5.2 DNA analysis and purification

Following PCR, restriction enzyme digest, or cloning, DNA fragments were analysed by electrophoresis at 90-100 V in 1-2% agarose in 1× TAE buffer. Where loading dyes were not included in the reaction mixture, samples were first mixed with 6× Purple Gel Loading Dye (NEB).

Depending upon the purity of a reaction product when analysed by electrophoresis, DNA products were purified by excision from a gel followed by use of the QIAquick Gel Extraction Kit or MinElute Gel Extraction Kit, or by use of the QIAquick PCR Purification Kit or MinElute PCR Purification Kit (all Qiagen).

2.5.3 TA cloning

Following amplification by PCR of a fragment for cloning or sequencing, such as RACE products, amplicons were cloned into pGEM-Teasy (Promega), according to manufacturer's instructions. Where fragments had been prepared with Phusion® High-Fidelity Polymerase, these were subsequently A-tailed using Platinum® DNA Polymerase prior to cloning.

2.5.4 DNA ligation

Following preparation of DNA fragments by restriction enzymes and purification (as above), complementary DNA overhangs were ligated together using T4 DNA ligase (NEB), overnight at 4°C.

2.5.5 Bacterial amplification of plasmids

2.5.5.1 Transformation of competent bacteria

Following ligation of an A-tailed amplicon into pGEM-Teasy, or of a complementary fragment into an expression vector, 1-2 µl of the ligation products were transformed into bacteria for subsequent amplification, according to the manufacturer's protocols. For One Shot® TOP10 Chemically Competent *E. coli* (Thermo Fisher): following addition of DNA, bacteria were incubated on ice for at least 5 minutes, and then heat shocked at 42°C for 45 seconds, before chilling on ice for 2 minutes. 250 µl of S.O.C. medium was added, and the bacteria were incubated at 37°C for 1 hour with horizontal shaking at 240 rpm, prior to plating. For Subcloning Efficiency™ DH5α™ Competent Cells: bacteria were incubated on ice for 15-30 minutes after DNA addition, and heat shock was carried out at 42°C for 20 seconds, before chilling on ice for 2 minutes. The volume was adjusted to 1 ml with 950 µl of S.O.C medium and bacteria were incubated at 37°C for 1 hour with horizontal shaking at 240 rpm.

2.5.5.2 Bacterial culture

Following bacterial transformation, bacteria were spread on solid LB-agar, containing appropriate antibiotics for the purpose of transformant selection (and, in the case of *lacZ*-complementing plasmids, X-gal: 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside), and grown for 14-20 hours at 37°C.

Single colonies and patches, and glycerol stocks, were amplified by growth in liquid LB. Cultures were maintained for 14-18 hours at 37°C with horizontal shaking at 240 rpm.

2.5.5.3 Glycerol stocks

Where transformed colonies were to be stored long term, glycerol stocks were produced by mixing bacterial liquid LB culture 1:1 with sterile 50% glycerol (final concentration 25% v/v) and storing at -80°C.

2.5.6 Sanger sequencing

Sequencing of DNA amplicons and constructs was carried out by Source BioScience LifeSciences, in the Department of Biochemistry, University of Oxford.

2.5.7 Construct preparation

2.5.7.1 *CEROX1* overexpression construct

Prior to this work, the full length *CEROX1* sequence was synthesised by Biomatik (Ontario, Canada), and then cloned into the pCAGGS overexpression vector by Dr Tamara Sirey using XhoI and BglII restriction sites. Features of the pCAGGS vector are summarised in Table 2.4.

Feature	Function
CAG 'promoter': CMV immediate-early enhancer; chicken β-actin promoter; rabbit β-globin splice acceptor	Promoter
Rabbit β-globin poly-A site	Polyadenylation signal
SV40 origin	Mammalian origin of replication
pMB1 (pBR322) origin	Bacterial origin of replication
Ampicillin resistance	Selectable marker

Table 2.4: Features of the pCAGGS overexpression vector. This vector was used to express *CEROX1* and variants thereof, as well as the EYFP control.

2.5.7.2 Human influenza haemagglutinin (HA)-tagged ORFs

In order to tag the ORFs within *CEROX1* and EYFP, primers were designed to target the sequence either side of the ORF/STOP codon boundary, in the case of the C-terminal tags, or following the start codon in the case of the N-terminal tags. The primers were flanked with the HA or MYC tag DNA sequence as appropriate (HA: 5'-TATCCGTATGATGTGCCGGATTATGCG-3'; MYC: 5'-GAACAAAAACTCATCTCAGAAGAGGATCTG-3'). Pairs of amplified DNA fragments were assembled by PCR-driven overlap extension (Figure 2.1). Sequences including the inserted epitope tag were then cut from the amplicons with pairs of restriction enzymes (Table 2.5), and ligated into the equivalently digested overexpression construct.

In order to create a start codon mutant via the introduction of a single nucleotide substitution (ATG>ATT), primers including the mutation were designed to cover the regions flanking it on both sides. These were used as above in PCR-driven overlap extension to produce the mutant fragment.

ORF modification site	5' restriction enzyme	3' restriction enzyme
EYFP C-terminal tagging	XhoI	Bsu36I / Eco81I
CEROX1-ORF1 start codon 1 mutation	XhoI	NotI
CEROX1-ORF1 N-terminal tagging	XhoI	NotI
CEROX1-ORF1 C-terminal tagging	BbsI / BpiI	SacI / Eco53I
CEROX1-ORF2 C-terminal tagging	AleI / OliI	SphI / PaeI
CEROX1-ORF3 C-terminal tagging	PacI	AleI / OliI

Table 2.5: Flanking restriction enzymes used to produce tagged CEROX1-ORF expression constructs. Restriction enzymes in this table were used in replacement of loci with PCR-driven overlap extension-derived mutant fragments. See also Figure 2.1.

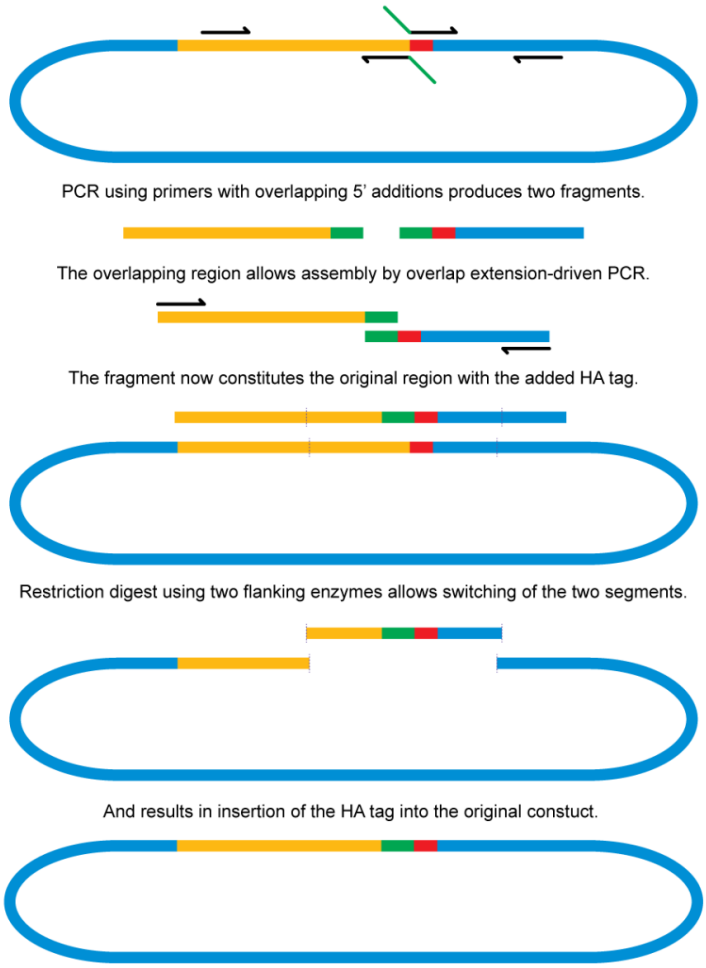


Figure 2.1: Generation of an HA-tagged ORF construct by PCR-driven overlap assembly. CEROX1-ORF, yellow; rest of overexpression vector and *CEROX1* sequence, blue; HA tag, green; STOP codon; red.

2.5.7.3 *CEROX1* / *SOX8* promoter-driven luciferase reporter constructs

The annotated sequence of the human genome between the 5' ends of *CEROX1* and *SOX8*, flanked with XhoI restriction sites, was synthesised by GenScript (Hong Kong), and provided in a pUC57 backbone. The promoter region was released by digestion with XhoI and ligated with into the pGL4.20 backbone (Promega), which was prepared by digestion with XhoI and dephosphorylation in order to prevent self-ligation. Bidirectional clones were screened by diagnostic digestion with NotI and SacI. To prepare pNL2.3 Nano-luc constructs, the promoter was digested from the pGL4.20 vectors with NheI and HindIII, and ligated into the parallel multiple cloning site of the pNL2.3 vector. In order to truncate the *CEROX1*-forward orientation of the promoter within the pNL2.3 vector, restriction enzymes that cut only within the promoter region of the vector were chosen and paired with SacI at the 5' end to release fragments. Where compatible sticky ends were not generated, the termini were blunted with the Klenow fragment of DNA polymerase I. The linearised vector fragments were then self-ligated with T4 DNA ligase.

2.5.7.4 *CEROX1* and *SOX8 in situ* hybridisation probe template constructs

Fragments of the human and mouse *CEROX1*/*Cerox1* and *SOX8*/*Sox8* genes were amplified by PCR using Platinum® DNA Polymerase, and then ligated into pCR™4-TOPO® (Thermo Fisher) according to the manufacturer's protocol.

2.6 Techniques for protein extraction and experimentation

2.6.1 Protein Extraction

2.6.1.1 For assay of mitochondrial electron transport chain complex enzymatic activities

Forty-eight hours post-transfection, cells were washed in PBS, re-suspended in KME buffer (100 mM KCl, 50 mM MOPS, 0.5 mM EGTA, pH 7.4), and then subjected to lysis by multiple rounds of flash freezing in liquid nitrogen followed by rapid thawing. Protein concentration was assayed using the Pierce™ BCA Assay Kit (Thermo Fisher Scientific) in microplate format, in triplicate.

2.6.1.2 For analysis of protein from cultured cells by SDS-PAGE and immunoblotting

Forty-eight hours post-transfection, cells were washed in ice-cold PBS, and then re-suspended and lysed in ice-cold RIPA buffer (150 mM sodium chloride, 1.0% IGEPAL® CA-630, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris, pH 8.0) with cOmplete™ Mini Protease Inhibitor Cocktail (both Sigma-Aldrich). Following incubation on ice, lysed samples were sonicated for 2 seconds, in order to shear chromatin and reduce viscosity. Insoluble cell debris was pelleted by centrifugation at 16,000 g for 20 minutes at 4°C, and the supernatant retained. Protein concentration was assayed using the Pierce™ BCA Assay Kit (Thermo Fisher Scientific) in microplate format, in triplicate.

2.6.1.3 For analysis of mitochondrial protein by native PAGE

Following isolation of mitochondria (see below), a small aliquot was lysed in RIPA according to the protocol for SDS-PAGE above, and this was used to determine protein concentration. Thereafter, a known amount of undenatured mitochondrial protein was solubilised with digitonin (8 g g^{-1}) in 1× NativePAGE™ sample buffer (Thermo). Insoluble debris was pelleted by centrifugation at 16,000 g at 4°C for 30 minutes, and the

supernatant retained. Coomassie blue G-250 was added to a final concentration one quarter that of the chosen detergent, prior to gel loading.

2.6.1.4 For analysis of protein from human tissue

Samples of human striatum tissue and a diffuse glioma tumour were obtained from the Oxford Brain Bank (Dr Olaf Ansorge). Around 1 g of frozen tissue was homogenised in 800 µl of ice-cold RIPA, with protease inhibitors as above, using a Precellys® Homogenizer (Bertin Instruments) with a CryoLys® cooling system (Bertin Instruments) to keep the equipment below 10°C. The resulting homogenate was processed, clarified, and quantified as above.

2.6.1.5 Ubiquitination

Ubiquitination is reversible, and so the action of de-ubiquitinating enzymes must be countered during the preparation of protein samples. There are six major de-ubiquitinating enzymes, one is a metallo-enzyme and the other five are cysteine enzymes[5]. To inhibit metallo-enzymes, EDTA was added to the lysis buffer at final concentration of 10 mM, in order to chelate metal ions. Additionally, iodoacetamide (IAA) was added at 100 mM to inhibit the cysteine-proteinases[5]. Samples were processed in the dark as quickly as possible, in order to prevent premature deactivation of IAA.

2.6.2 Mitochondrial isolation

Mitochondria for BN-PAGE were isolated using the Qproteome Mitochondria Isolation Kit (Qiagen), according to the manufacturer's instructions, but with two additional centrifuge clarification steps to minimise nuclear contamination of the mitochondria. Pellets of 1.5×10^8 HEK293T cells were used to produce 250 µg of mitochondrial protein.

2.6.3 Electron transport chain complex enzyme activity assays

Protein lysates were centrifuged at 700 g for 1 minute at 4°C to pellet cell debris, and the supernatant used in each assay. Enzyme assays were carried out in 1 ml disposable polystyrene cuvettes (Fisherbrand®) using

a Shimadzu UV-1800 spectrophotometer. See Tables 2.6 and 2.7 for measurement durations and wavelengths, respectively.

Specific activity of each enzyme assayed was calculated using Equation 2.1.

$$A = \frac{10^9 \times v \times \Delta_{abs}}{\epsilon \times p \times x}$$

Equation 2.1: Calculation of specific enzyme activity (A) using a spectrophotometric assay. Specific enzyme activity was calculated in nmol min^{-1} mg^{-1}. v is the total assay volume in dm^3, Δ_{abs} is the change of absorbance in min^{-1}, ϵ is the extinction coefficient for the chromophore (see Table 2.7), p is the assay pathlength in cm, and x is the amount of total protein added in mg.

Assay	Buffer composition (final concentrations in buffer)	Protocol	Time (s)
CI	25 mM potassium phosphate buffer, pH 7.2; 5 mM $MgCl_2$; 0.13 mM reduced nicotinamide adenine dinucleotide (NADH); 65 µM coenzyme Q; 2.5 mg ml^{-1} fatty acid-free albumin; 3.645 µM antimycin A.	Incubate 1 ml reaction buffer; add 100 µl of protein suspension, rapidly mix, and measure absorbance.	180
		Add 11 µl of 1 mM rotenone in DMSO, rapidly mix, and measure absorbance.	120
CII	Buffer A: 25 mM potassium phosphate buffer, pH 7.2; 5 mM $MgCl_2$; 2 mM sodium succinate hexahydrate. Buffer B: 365 µM antimycin A; 507 µM rotenone; 5 mM 2,6-dichloro-phenolindophenol (DCIP).	Pre-incubate 985 µl assay buffer A; add 100 µl of protein suspension, rapidly mix, and then incubate for 10 minutes. Add 10 µl of assay buffer B, rapidly mix, and measure absorbance.	60
		Add 5 µl of 13 mM coenzyme Q in absolute ethanol; rapidly mix, and measure absorbance.	120
CIII	30 mM potassium phosphate buffer, pH 7.2; 3 mM sodium azide; 1.5 µM rotenone; 50 µM oxidised cytochrome c; 25 µM decylubiquinol*.	Incubate 1 ml reaction buffer; add 100 µl of protein suspension and 5 µl of decylubiquinol solution, rapidly mix, and measure absorbance.	90
CIV	10 mM potassium phosphate buffer, pH 7.0; 50 µM reduced cytochrome c*.	Incubate 1 ml reaction buffer; add 100 µl of protein suspension, rapidly mix, and measure absorbance.	180
CS	1M Tris-HCl, pH 8.1; 1 mM 5,5'-dithio-bis(2-nitrobenzoic acid) (DTNB).	Aliquot 100 µl assay buffer and add 750 µl water. Add 50 µl of acetyl-CoA* solution and 100 µl of protein suspension; mix, and incubate for 5 minutes. Add 50 µl of 10 mM oxaloacetate solution, rapidly mix, and measure absorbance.	180
N.B.: All assays and incubations were carried out at 30°C.			
* See below for preparation of decylubiquinol, reduced cytochrome c, and acetyl-CoA.			

Table 2.6: Buffers and protocols used in mitochondrial enzyme assays. Protocols were sourced from Professor Lisa Heather (DPAG, University of Oxford).

2.6.3.2 Reduction of cytochrome c

A 10 mg ml^{-1} solution of cytochrome c in 100 mM potassium phosphate buffer, pH 7.0 was reduced by the addition of sodium ascorbate to a final concentration of 1 mg ml^{-1}. The resulting solution of reduced

cytochrome c was purified of excess ascorbate by dialysis using SnakeSkin Dialysis Tubing, 7K MWCO (Pierce™, Thermo Scientific) in 4 litres of potassium phosphate buffer overnight at 4°C.

2.6.3.3 Reduction of decylubiquinone to decylubiquinol

A 10 mM solution of decylubiquinone in ethanol was acidified to pH 2.0 using 6 M hydrochloric acid. The compound was then reduced by addition of sodium borohydride, in a fume hood. The solution was diluted two-fold with water, and then centrifuged to pellet excess sodium borohydride. The supernatant was retained and incorporated into the complex III assay buffer.

2.6.3.4 Acetylation of coenzyme A

Ten milligrams of coenzyme A (13 µmol) was dissolved in 1 ml of 100 mM potassium hydrogen carbonate and then acetylated by addition of 2 µl acetic anhydride and incubation upon ice for 10 minutes. The resulting mixture was diluted five-fold with water, and 50 µl was used in each citrate synthase assay reaction.

Assay	Wavelength (nm)	ϵ ($M^{-1} cm^{-1}$)
CI	340	6220
CII	600	21,000
CIII	550	21,000
CIV	550	21,840
CS	412	13,600

Table 2.7: Wavelengths of absorbance readings, and molar extinction coefficients (ϵ) used in calculations of enzyme activity for ETC complexes I to IV (CI-IV) and citrate synthase (CS).

2.6.4 SDS polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting

2.6.4.1 Small proteins (< 80 kDa)

For detection of tagged *CEROX1*-ORF proteins, and ETC complex subunits, SDS-PAGE and western blotting were carried out using the Novex® NuPAGE® system. For each sample, protein was combined with NuPAGE® 4× LDS Sample Buffer and NuPAGE® 10× Sample Reducing Agent, loaded onto a NuPAGE® 12% bis-tris gel (Invitrogen, Thermo Fisher), and separated by electrophoresis for around 150 minutes at 90 V in NuPAGE® MES SDS Running Buffer. Proteins were electroblotted onto a 0.2 µm pore-size polyvinylidene difluoride (PVDF) membrane in NuPAGE® Transfer Buffer with 10% methanol for 1 hour at 30-40 V, on ice. The

membrane was then blocked with blocking solution (5% skimmed milk in TBS-T) for 1 hour at room temperature before incubation with primary antibody in blocking solution overnight at 4°C. The membrane was rinsed and washed three times for 5 minutes in TBS-T, then incubated in secondary antibody in blocking solution for 1-2 hours at room temperature. The membrane was rinsed and washed three times for 5 minutes in TBS-T, and then imaged using ECL reagent on an ImageQuant LAS 4000 (GE Healthcare Life Sciences). Blot images were exported as GEL files and processed in ImageJ.

2.6.4.2 Larger proteins

For larger proteins, such as ubiquitinated species, modifications were made to the above protocol for western blotting. SDS-PAGE was carried out using NuPAGE® 10% bis-tris gels (Invitrogen, Thermo Fisher), and proteins were separated by electrophoresis in NuPAGE® MOPS SDS Running Buffer. Electroblotting transfer was carried out at 70 V for 2 hours at 4°C.

2.6.4.3 Re-probing blots

Where blots were to be re-probed with a second primary antibody, for the purposes of a loading control or examining multiple targets, existing HRP-conjugated secondary antibodies were inactivated using hydrogen peroxide as described by Sennepin et al.[6]. Briefly, blots were incubated with 30% hydrogen peroxide for 15-30 minutes at 37°C, and then washed thoroughly in TBS-T prior to new primary incubation, which proceeded as normal. Where possible, second primary antibodies raised in different host species were chosen, to avoid cross-amplification with the original primary antibody.

2.6.5 Blue native polyacrylamide gel electrophoresis (BN-PAGE)

In order to investigate complexes of the mitochondrial electron transport chain and their enzymatic activities, solubilised mitochondrial protein was loaded onto NativePAGE™ Novex® Bis-Tris 3-12% gradient gels. Initially, the cathode chamber was filled with 1× NativePAGE™ running buffer with the addition of 0.02% Coomassie blue G-250. After running at 150 V for one hour, the cathode buffer was changed to clear 1× NativePAGE™ running buffer, and the gel run at 300 V for up to two hours. All BN-PAGE was carried out in ice-water to preserve protein complex integrity.

2.6.5.1 In-gel enzyme activity assays

Immediately after electrophoresis, gels were rinsed in ice-cold water. Then gels were transferred into the appropriate assay buffer (Table 2.8) and incubated with light agitation for up to 8 hours. Following stain development, assays were halted by fixation (Table 2.8). Images were taken against an illuminated white or black (complex V) background with a Canon EOS D500 camera, and processed with ImageJ.

Assay	Buffer composition	Staining time	Post-stain fixing reagent
CI	50 mM potassium phosphate buffer pH 7.0; 0.2 mg ml^{-1} nitrotetrazolium blue (NTB); 0.1 mg ml^{-1} reduced nicotinamide adenine dinucleotide (NADH)	7 hours	50% methanol, 10% acetic acid
CII	5 mM Tris-HCl pH 7.4; 0.2 mg ml^{-1} nitrotetrazolium blue (NTB); 20 mM sodium succinate; 200 µM phenazine methosulphate	45 minutes	50% methanol, 10% acetic acid
CIV	50 mM sodium phosphate buffer pH 7.4; 1 mg ml^{-1} 3,3'-diaminobenzidine (DAB); 50 µM cytochrome c	7 hours	50% methanol, 10% acetic acid
CV	35 mM Tris pH 8.3; 270 mM glycine; 140 mM magnesium sulphate; 10 mM adenosine triphosphate (ATP); 0.2% lead nitrate	7 hours	50% methanol (no acid due to solubility of assay product)

Table 2.8: Assay buffers used in staining ETC complex activities following native PAGE.

2.6.5.2 In-gel complex III enzyme activity assays

Several assays were attempted to assess complex III activity in-gel, but all failed (Table 2.9).

Assay	Buffer composition	Comments
CIII	50 mM sodium phosphate buffer pH 7.4; 1 mg ml^{-1} 3,3'-diaminobenzidine (DAB)[7]	Weakly stained complex IV instead
	1-Step™ TMB-Blotting Substrate Solution (Thermo Fisher)[8]	Didn't stain any bands

Table 2.9: Assay buffers trialled for staining of complex III activity following native PAGE.

2.6.6 Densitometry

Where western blotting or BN-PAGE were used semi-quantitatively, band density was quantified using Image Studio Lite (LI-COR Biosciences).

2.6.7 Immunoprecipitation

2.6.7.1 Using bead-conjugated antibodies

HA-tagged proteins were purified by immunoprecipitation with EZview Red Anti-HA Affinity Gel beads (Sigma). Following equilibration in RIPA buffer, 50 µl of bead slurry was incubated with cell lysate, prepared as previously described, overnight at 4°C, in a total volume of 800-1200 µl. The immunoprecipitate was

washed three times in RIPA buffer and then eluted by heating to 70°C for ten minutes in 12 µl of NuPAGE® 4× LDS Sample Buffer. NuPAGE® 10× Sample Reducing Agent was added prior to gel loading.

2.6.7.2 Using non-conjugated antibodies

Cell lysate was incubated with the chosen primary antibody (5 to 20 µg) overnight at 4°C with gentle agitation. Antibody-target complexes were precipitated by addition of 50 µl of pre-equilibrated Protein G or Protein A Sepharose® beads or Dynabeads® and incubation with gentle agitation at 4°C for one hour. The immunoprecipitated proteins on the beads were pelleted, washed, and eluted as above.

2.6.7.3 Preparation for mass spectrometry

For confirmation of CEROX1-ORF1$^{C\text{-}HA}$ expression *in vitro*: following SDS-PAGE, the gel was fixed in 50% methanol, 10% acetic acid, and then stained with InstantBlue™ Coomassie® protein stain (Expedeon) prior to band excision.

For identification of anti-CEROX1-ORF1 immunoprecipitated proteins from brain tissues, Dynabead-protein A-antibody complexes were pelleted and washed, before storage at -80°C in a minimal volume of RIPA buffer prior to on-bead protein digestion.

2.6.8 Immunocytochemistry

HEK293T and SK-N-SH cells were grown on 13 mm glass coverslips that were coated with poly-L-lysine or poly-L-ornithine, respectively, for at least one hour at 37°C. Cells were stained according to a protocol obtained from Dr Lauren Watson (Becker laboratory, DPAG, University of Oxford). Cells were fixed in 4% PFA in PBS, and then permeabilised in 0.4% Triton X-100 in PBS, each for 20 minutes, before being incubated in blocking solution (10% skimmed milk and 1% normal goat serum in TBS with 0.02% Tween-20) for one hour. Cells were then incubated with primary antibody in antibody carrier solution (3% BSA and 0.02% sodium azide in TBS with 0.02% Tween-20) overnight at 4°C. Following three washes of five minutes in PBS, cells were incubated for one hour at room temperature in secondary antibody in antibody carrier solution. Cells were washed thrice more in PBS, and then mounted onto slides with VECTASHIELD™ Antifade

Mounting Medium with DAPI (Vector Laboratories) and sealed with clear nail varnish. Where fluorescently-labelled phalloidin was used to stain actin, it was added with the secondary antibody at 1:250.

2.6.9 Immunohistochemistry

Frozen sections of human control striatum and glioma tumour were dried for 15 minutes at room temperature. Sections were fixed in 4% PFA in PBS, and then washed three times for five minutes in PBS. Permeabilisation was carried out in 0.1% Triton X-100 in TBS for 10 minutes, followed by another three washes of five minutes, in TBS. Sections were blocked in 10% normal goat serum and 1% BSA in TBS with 0.1% Tween-20 for one hour at room temperature. Slides were then incubated with primary antibody in antibody carrier solution (1% BSA in TBS with 0.1% Tween-20) overnight at 4°C. Following three washes of five minutes in TBS, cells were incubated for one hour at room temperature in secondary antibody in antibody carrier solution. Sections were washed thrice more in TBS, and then mounted with VECTASHIELD™ Antifade Mounting Medium with DAPI (Vector Laboratories) and borosilicate glass cover slips (thickness #1; Thermo).

2.6.10 Antibodies

The following Tables 2.10-2.12 detail the commercial antibodies used in western blotting and immunocytochemistry, including their host species and dilutions in use.

Primary antibodies				
Target	Host species	WB dilution	ICC dilution	Supplier and code
HA tag	Rabbit	1:500	1:100	Sigma H6908
HA tag	Rat	1:500	-	Sigma Roche 11867423001
Histone H3	Rabbit	1:1,000	-	Abcam ab1791
Insulin	Guinea pig	1:500	-	Dako A0564
MYC tag	Mouse	1:500	-	Sigma M4439
Ubiquitin	Rabbit	1:200	-	Abcam ab7780
TUBA4A	Mouse	1:5,000	-	Sigma T5168
NDUFA13	Mouse	1:500	-	Santa Cruz sc-365978
NDUFA10	Mouse	1:500	-	Santa Cruz sc-376357
NDUFS1	Rabbit	1:10,000	-	Abcam ab169540
NDUFS3	Mouse	1:2,000	-	Abcam ab110246
NDUFAF2	Mouse	1:250	-	Santa Cruz sc-365592
NDUFAF4	Mouse	1:250	-	Santa Cruz sc-166928
SOX8	Mouse	1:200	1:100	Santa Cruz sc-374446
CALB1	Rabbit	-	1:200	Cell Signalling Technologies 13176
OPRM1	Rabbit	-	1:200	Sigma O0507

Primary antibodies				
Target	Host species	WB dilution	ICC dilution	Supplier and code
GFAP	Chicken	-	1:1000	Merck Millipore AB5541
CHAT	Goat	-	1:1000	Merck Millipore AB144P
IBA1	Goat	-	1:300	Abcam Ab5076
OLIG2	Goat	-	1:50	Biotechne R&D Systems AF2418
GAD65	Mouse	-	1:100	Sigma G1166
FOXP1	Mouse (JC12 hybridoma)	-	1:10	Banham laboratory, Radcliffe Department of Medicine, University of Oxford

Table 2.10: Primary antibodies used in western blotting (WB) and immunocytochemistry (ICC).

HRP-conjugated secondary antibodies for WB			
Target	Host species	WB dilution	Supplier
Rabbit IgG	Donkey	1:6,000	GE Healthcare NA934
Guinea pig IgG	Goat	1:5,000	Abcam ab6908
Mouse IgG	Sheep	1:5,000	GE Healthcare NA931

Table 2.11: Horseradish peroxidase-conjugated secondary antibodies used in western blotting.

Fluorescently-labelled secondary antibodies for FISH, ICC, IHC				
Target	Host species	Fluorophore	Dilution	Supplier
Sheep IgG	Donkey	AlexaFluor® 488	1:500	Thermo Fisher A11015
Rabbit IgG	Donkey	AlexaFluor® 488	1:1000	Thermo Fisher A21206
Rabbit IgG	Goat	AlexaFluor® 594	1:1000	Thermo Fisher A11037
Mouse IgG	Goat	AlexaFluor® 594	1:1000	Thermo Fisher A11005
Goat IgG	Donkey	AlexaFluor® 594	1:1000	Thermo Fisher A11058
Chicken IgG	Goat	AlexaFluor® 594	1:1000	Thermo Fisher A11042
Mouse IgG	Goat	AlexaFluor® 488	1:1000	Thermo Fisher A11001

Table 2.12: Fluorophore-conjugated secondary antibodies used in fluorescent RNA *in situ* hybridisation, immunohistochemistry, and immunocytochemistry.

2.6.10.2 Custom *CEROX1*-ORF1 antibody synthesis

Custom antibodies targeting the CEROX1-ORF1 protein were produced by AMSbio (Abingdon, Oxfordshire, UK). Briefly: following synthesis of a 17 amino acid peptide CPQESRHRPLGLTVPSG representing the C-terminal end of the predicted CEROX1-ORF1 reading frame, two New Zealand white rabbits were immunised over a course of ten weeks, by injection of the peptide suspended in Complete Freund's Adjuvant or Incomplete Freund's Adjuvant subcutaneously into the shoulders and intramuscularly into the hind leg. Intermittent and then terminal bleeds via the central ear artery were used to harvest blood, from which sera samples were produced. The resulting antibodies were subjected to affinity purification using the same immunogenic peptide, producing 2.4 mg and 2.49 mg of purified polyclonal antibody from the two rabbits, respectively.

Following testing by dot blot and western blotting, the affinity purified antibody from one rabbit was chosen as being the most sensitive and specific, and that antibody preparation was used thereon in detecting CEROX1-ORF1.

2.6.11 Mass spectrometry

For confirmation of *in vitro* CEROX1-ORF1 translation in HEK293T cells, MS was performed by Dr Sabrina Liberatori and Dr Svenja Hester (Advanced Proteomics Service, Department of Biochemistry, University of Oxford). Gel bands were subjected to in-gel trypsin digestion as previously described[9]. After digestion was completed, peptides were separated by nano-flow reversed-phase liquid chromatography coupled to a Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Scientific). Peptides were loaded on a C18 PepMap100 pre-column (inner diameter 300 µm × 5mm, 3 µm C18 beads; Thermo Scientific) and separated on a 50 cm reversed-phase C18 column (inner diameter 75 µm, 2 µm C18 beads) using a linear gradient from 10 to 35% of B for 30 min at a flow rate of 200 nl/min (A: 0.1% formic acid, B: 0.1% formic acid in acetonitrile). All data were acquired in a data-dependent mode, automatically switching from MS to collision-induced dissociation MS/MS on the top 10 most abundant ions with a precursor scan range of 350–2000 m/z. MS spectra were acquired at a resolution of 70,000 and MS/MS scans at 17,500. Dynamic exclusion was enabled with a duration of 5 seconds and charge exclusion was applied to unassigned and mono-charged ions. Raw data files were processed for protein identification using MaxQuant, version 1.5.0.35, integrated with the Andromeda search engine[10-12]. MS/MS spectra were searched against the human UniProt database; precursor mass tolerance was set to 20 ppm and MS/MS tolerance to 0.05 Da. Enzyme specificity was set to trypsin with a maximum of two missed cleavages. False discovery rate for protein and peptide spectral matches was set at 1%.

For identification of putative CEROX1-ORF1 peptides *in vivo* following IP from brain tissues, LC MS/MS was performed by Dr Sarah Bonham and Dr Roman Fisher (Target Discovery Institute, Nuffield Department of Medicine, University of Oxford). LC MS/MS analyses were performed using a Dionex Ultimate 3000 nano-ultra high pressure reverse phase chromatography coupled on-line to a Q Exactive High Field (HF) mass spectrometer (Thermo Scientific). Samples, injected in 2% acetonitrile, 0.1% formic acid, were separated on an EASY-Spray PepMap RSLC C18 column (500 mm × 75 µm, 2 µm particle size, Thermo Scientific) over a 60

minute gradient of 2-35% acetonitrile in 5% DMSO, 0.1% formic acid at 250 nl min^{-1}. MS1 scans were acquired at a resolution of 60,000 at 200 m/z and the top 12 most abundant precursor ions were selected for HCD fragmentation. Data were analysed using PEAKS, including *de novo* peptide sequencing and searches against human and gorilla UniProt databases; MS/MS tolerance was set to 0.05 Da. False discovery rate for protein and peptide spectral matches was set at 1%.

2.7 Assays

2.7.1 Nano-Glo® luciferase assay

For non-treated assays, HEK293T and SH-SY5Y cells were plated in 96-well plates at a density of 10,000 or 20,000 cells per well, respectively, in 125 µl of medium. Twenty-four hours after seeding, cells were transfected with 40 ng of firefly luciferase and 2 ng of *Renilla* luciferase constructs per well, using FuGENE® 6 at a FuGENE®:DNA ratio of 6 µl:1 µg. Twenty-four hours-post transfection, between 10 µl and 100 µl of medium was removed into an opaque white 96-well plate, and supplemented as necessary with PBS to a final volume of 100 µl. Luminescence activity was measured according to the manufacturer's protocol. Briefly: an equal volume of NanoGlo® Luciferase Assay Reagent (a 1:50 v/v solution of NanoGlo® Luciferase Assay Substrate in NanoGlo® Luciferase Assay Buffer) was added to the cell medium sample, and the mixture incubated for at least three minutes at room temperature, before luminescence was measured with a FLUOstar Omega Microplate Reader (BMG LABTECH).

2.7.1.1 Compound screen

Where cells were to be treated to assess promoter responsivity to stimuli, SH-SY5Y cells were plated in 96-well plates at a density of 10,000 cells per well in 100 µl of medium. Twenty-four hours post-transfection, the medium was removed and replaced with 125 µl per well of drug-, growth factor-, or hormone-laced media, prepared in a replica plate. Cells were growth under treatment for 20 hours before 100 µl of media were removed and assayed for secreted luciferase activity as above. Treatments were prepared in water, DMSO, or ethanol depending upon solubility, and subsequent dilutions ensured that the final concentration of non-aqueous solvents was 0.1% or less. Control readings were taken from wells treated only with fresh medium, 0.1% DMSO, or 0.1% ethanol.

2.7.1.2 Fold change and Z-score calculations

Fold change upon treatment was calculated using Equation 2.2. Fold changes were then converted to Z-scores using Equation 2.3.

$$\text{Fold change} = \frac{CEROX1\text{-}SOX8_{\text{Treated}} - \text{Empty}_{\text{Treated}}}{CEROX1\text{-}SOX8_{\text{Control}} - \text{Empty}_{\text{Control}}}$$

Equation 2.2: Calculation of fold change in promoter activity following drug treatment. For a given treatment, the fold change was calculated using the activity of the *CEROX1-SOX8* promoter and the empty promoter, under treatment and the appropriate control treatment (DMSO, ethanol, medium).

$$\text{Z-score}_{\text{k}} = \frac{\text{Fold change}_{\mu} - \text{Fold change}_{\text{k}}}{\text{Fold change}_{\sigma}}$$

Equation 2.3: Conversion of treatment-induced promoter activity fold change into a population Z-score. For a given treatment k, the fold change as calculated using Equation 2.2 was subtracted from the average fold change across all treatments (μ), and this was divided by the standard deviation (σ) of the same set of fold changes.

2.7.1.3 Cell number approximation

A parallel plate of treated cells was used to assess the proliferative effects of the screened treatments using the Cell Proliferation Assay Kit (Merck Millipore), according to the manufacturer's protocol.

2.7.2 Seahorse extracellular flux analysis

Respiratory activity was measured by Dr Tamara Sirey using an Agilent Seahorse Extracellular Flux Analyzer, with assistance from Roderick Carter, in the laboratory of Nicholas Morton, Centre for Cardiovascular Science, University of Edinburgh. HEK293T cells were plated on poly-L-lysine coated XFe24 microplates at 50,000 cells per well and grown for 24 hours prior to transfection with 300 ng of DNA. Respiratory flux was then assayed 24 hours following transfection: cells were washed three times with Seahorse Assay media (Agilent), supplemented with 10 mM glucose and 2 mM pyruvate and subjected to 30 minutes of pre-incubation at 37°C (ambient CO_2). After basal respiration measurements, 1 μM of oligomycin was injected to inhibit ATP synthase, then 0.2 μM of carbonyl cyanide 4-(trifluoromethoxy)phenylhydrozone (FCCP) was injected to uncouple respiration. Finally, 1 μM each of rotenone and antimycin A were injected to inhibit complex I and complex III respectively. Following respiratory assay, cells were fixed with trichloroacetic acid,

and total protein content was assayed by sulforhodamine B (SRB) colorimetric assay[13,14], in order to normalise oxygen consumption. Data were analysed using Agilent Wave software.

2.8 Imaging

2.8.1 Confocal microscopy

Cells grown and stained on cover slips were visualised by confocal microscopy using a Leica TCS SP5 II confocal microscope alongside Leica LAS AF acquisition software, with an HCX PL APO CS 63.0× 1.4 oil-immersion objective, used with Immersol 518N. Images were exported as TIF files and processed in ImageJ. Immuno-stained tissue sections were imaged by confocal microscopy using an Olympus FV3000 microscope with a 40× dry objective, with assistance from Dr Andrew Jefferson, Micron Imaging Facility, Department of Biochemistry, University of Oxford. Exported OIR files were manipulated with ImageJ using the OlympusViewer plugin.

2.8.2 Low-magnification imaging

Chromogenic ISH sections were imaged with a Stemi 2000-C stereo microscope and an AxioVision HRc CCD camera (both Zeiss). Panoramic images were assembled using the Photomerge function in Adobe Photoshop CS5.

2.9 Bioinformatics

2.9.1 Genome browsing

The UCSC Genome Browser (genome.ucsc.edu)[15] was used extensively in the examination of the *CEROX1* locus, synteny, and conservation (Multiz alignments of 100 vertebrates, and of 30 mammals, and associated PhyloCSF[16], PhyloP[17], and PhastCons[18] data). RNA-seq and ChIP-seq data from the ENCODE project were also accessed via the UCSC Genome Browser, as were the results of PLAR[19]. Additionally, the UCSC Table Browser[20] was utilised in downloading data ready for submission to iSeeRNA, and the FANTOM data browser Zenbu (fantom.gsc.riken.jp/zenbu)[21] was used to analyse CAGE data. The FANTOM5 Table Extraction Tool (fantom.gsc.riken.jp/5/tet) was used to access CAGE peaks for additional species including dog and rat.

2.9.2 Alignments

BLAT (BLAST-like alignment tool)[22], accessed via the UCSC Genome Browser, was used in interpreting sequencing of RACE-PCR products and during the production of vector constructs. NCBI BLAST (blast.ncbi.nlm.nih.gov) was used to search for homologues of *CEROX1* and CEROX1-ORFs, as well as to map Laurasiatheria transcriptomes in manual *CEROX1* orthologue annotation, and map *CEROX1*/*Cerox1* reads in all other transcriptomic/gene expression analyses. The NCBI Conserved Domain Database (ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)[23-26] was used to identify remnant domains with the *CEROX1* locus. MUSCLE (EMBL-EBI; ebi.ac.uk/tools/msa/muscle)[27-30] was used to align the DNA sequences of multiple species' *CEROX1* orthologues, as well as the putative protein sequences of human and primate CEROX1-ORF1 peptides.

2.9.3 Open reading frame annotations

To assess the coding potential of *CEROX1*, ATGpr (atgpr.dbcls.jp)[31] and the ExPASy Translate Tool (web.expasy.org/translate) were used to annotate putative ORFs. Six programs were used to score coding likelihood: Coding Potential Calculator (CPC, cpc.cbi.pku.edu.cn)[32]; CPC2 (cpc2.cbi.pku.edu.cn)[33]; Coding Potential Assessment Tool (CPAT, lilab.research.bcm.edu/cpat)[34]; PORTRAIT (bioinformatics.org/portrait)[35]; iSeeRNA (sunlab.cpy.cuhk.edu.hk/iSeeRNA)[36], and RNAcon (crdd.osdd.net/raghava/rnacon)[37].

2.9.4 MicroRNA response element predictions

MREs were predicted by Professor Ana Marques, Dr Jennifer Tan, Dr Oscar Bedoya-Reina, and Dr Wilfried Haerty, using TargetScanHuman v7.0 and TargetScanMouse v7.1[38].

Predicted MREs shared between *CEROX1* and mRNA transcripts were represented using Circos plots (circos.ca)[39].

2.9.5 Transcription factor predictions

Transcription factor binding sites in the *CEROX1-SOX8* promoter were predicted using MATCH and PATCH (gene-regulation.com/pub/programs.html)[40,41] based upon the TRANSFAC database[41], and Signal Scan

(www-bimas.cit.nih.gov/molbio/signal)[42,43]. Candidate transcription factor sites were then analysed for evolutionary conservation using ConTra v3 (bioit2.irc.ugent.be/contra)[44-46].

2.9.6 Expression data

RNA expression data were sourced from the Human Developmental Biology Resource (HDBR; hdbr.org/expression)[47], the genotype-tissue expression (GTEx) project (gtexportal.org)[48], Stemformatics (stemformatics.org)[49], and the SH-SY5Y Sequencing Project (systemsbiology.uni.lu/shsy5y).

Other next generation sequencing data, including ribosome profiling and Olig2 ChIP-seq data were obtained via the NCBI Sequence Read Archive (SRA; ncbi.nlm.nih.gov/sra)[50]. Individual references are provided in the text and figures.

Protein expression data were obtained from the Human Protein Atlas (proteinatlas.org)[51-53].

2.9.7 Statistical analysis

Statistical analyses were carried out using Microsoft Excel and Prism (GraphPad). Unless otherwise specified, all *p* values in this thesis result from application of Student's t-test (2-tailed, assuming heteroscedasticity for stringency). For all figures: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Where used, # $p < 0.1$. All error bars represent the standard error of the mean unless otherwise specified.

Following calculation of transcript GC contents in subsets using Microsoft Excel, group standard deviations were used to calculate the composite population standard deviation using Equation 2.4 (courtesy of Professor David Dickey, Department of Statistics, North Carolina State University, NC, US).

$$\sigma = \sqrt{\frac{\sum_1^N s_i^2(n_i - 1) + \sum_1^N n_i(\bar{X} - \bar{x_i})^2}{N - 1}}$$

Equation 2.4: Calculation of the population standard deviation, σ. Where: s is the standard deviation of each subgroup; $\bar{x}$ is the mean of each subgroup; $\bar{X}$ is the population mean; N is the number of subgroups, here being 10.

Calculation of the standard error of the mean for averaged values already associated with errors was calculated using Equation 2.5 (courtesy of Ken Tatebe, University of California, Berkeley).

$$\bar{c} = \frac{n_a}{N}\bar{a} + \frac{n_b}{N}\bar{b}$$

$$\epsilon_c = \sqrt{\frac{n_a^2 - n_a}{N}\epsilon_a^2 + \frac{n_b^2 - n_b}{N}\epsilon_b^2 + \frac{n_a n_b\left(\bar{a} - \bar{b}\right)^2}{N^3 - N^2}}$$

Equation 2.5: Calculation of the weighted average $\bar{c}$, and combined standard error of the mean, ϵ. Where: n is the number of samples in each subgroup a and b; N is the total number of samples.

2.9.8 Other

Following alignment of *CEROX1* orthologues using MUSCLE, the resulting alignment was used to build a phylogenetic tree using MEGA7[54].

Galaxy (usegalaxy.org)[55-57] was used to format and truncate GENCODE transcript sequences for GC content calculation.

2.10 Solutions

Recipes for all solutions may be found in Table 2.13.

Ubiquitous	
PBS (phosphate-buffered saline)	10 mM phosphate, 2.7 mM potassium chloride, 137 mM sodium chloride, pH 7.4
Solutions for DNA and RNA work	
EB	10 mM Tris-HCl, pH 8.5
TE (Tris-EDTA)	10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0
TAE (Tris-acetate-EDTA)	40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 8.3
SSC buffer (saline-sodium citrate buffer)	150 mM sodium chloride, 15 mM sodium citrate, pH 7.2
Protein extraction	
RIPA (radioimmunoprecipitation assay) buffer	150 mM sodium chloride, 50 mM Tris pH 8.0, 1% IGEPAL® CA-630, 0.5% sodium deoxycholate, 0.1% SDS
KME buffer	100 mM potassium chloride, 50 mM MOPS, 0.5 mM EGTA, pH 7.4
Western blotting	
TBS-T for western blotting	150 mM sodium chloride, 20 mM Tris-HCl pH 7.6, 0.1% (v/v) Tween® 20
Blocking solution for western blotting	5% skimmed milk (w/v) in TBS-T for western blotting
Mild antibody stripping buffer	200 mM glycine, 1% Tween® 20, 0.1% SDS, pH 2.2
Immunocytochemistry	
Paraformaldehyde (PFA) fixing solution	4% paraformaldehyde in PBS
TBS-T for immunocytochemistry	150 mM sodium chloride, 10 mM Tris-HCl pH 8, 0.02% (v/v) Tween-20
Blocking solution for immunocytochemistry	10% skimmed milk and 1% normal goat serum in TBS-T for immunocytochemistry
Permeabilising solution	PBS with 0.4% Triton X-100
Antibody carrier solution for immunocytochemistry	3% bovine serum albumin and 0.02% sodium azide in TBS-T for immunocytochemistry
Immunohistochemistry	
4% paraformaldehyde in PBS	4% paraformaldehyde in PBS
TBS for immunohistochemistry	150 mM sodium chloride, 20 mM Tris-HCl pH 7.6
TBS-T for immunohistochemistry	10% normal goat serum, 1% BSA in TBS-T for immunohistochemistry
Permeabilising solution for immunohistochemistry	PBS with 0.1% Triton X-100
Antibody carrier solution for immunocytochemistry	1% bovine serum albumin in TBS-T for immunohistochemistry

Table 2.13: Recipes for solutions used in the protocols above. Recipes for underlined components may also be found in this table.

REFERENCES

1. Merglen, A., Theander, S., Rubi, B., Chaffard, G. *et al.* Glucose sensitivity and metabolism-secretion coupling studied during two-year continuous culture in INS-1E insulinoma cells. *Endocrinology* **145**, 667-78 (2004).
2. Shi, X. & Jarvis, D.L. A new RACE method for extremely GC-rich genes. *Analytical biochemistry* **356**, 222-228 (2006).
3. Schnoor, M., Voß, P., Cullen, P., Böking, T. *et al.* Characterization of the synthetic compatible solute homoectoine as a potent PCR enhancer. *Biochemical and Biophysical Research Communications* **322**, 867-872 (2004).
4. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).
5. Emmerich, C.H. & Cohen, P. Optimising methods for the preservation, capture and identification of ubiquitin chains and ubiquitylated proteins by immunoblotting. *Biochemical and Biophysical Research Communications* **466**, 1-14 (2015).
6. Sennepin, A.D., Charpentier, S., Normand, T., Sarre, C. *et al.* Multiple reprobing of Western blots after inactivation of peroxidase activity by its substrate, hydrogen peroxide. *Anal Biochem* **393**, 129-31 (2009).
7. Wittig, I., Karas, M. & Schagger, H. High resolution clear native electrophoresis for in-gel functional assays and fluorescence studies of membrane protein complexes. *Mol Cell Proteomics* **6**, 1215-25 (2007).
8. Smet, J., De Paepe, B., Seneca, S., Lissens, W. *et al.* Complex III staining in blue native polyacrylamide gels. *Journal of Inherited Metabolic Disease* **34**, 741 (2011).
9. Shevchenko, A., Tomas, H., Havli, J., Olsen, J.V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols* **1**, 2856 (2007).
10. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367 (2008).
11. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **10**, 1794-1805 (2011).
12. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**, 731 (2016).
13. Skehan, P., Storeng, R., Scudiero, D., Monks, A. *et al.* New Colorimetric Cytotoxicity Assay for Anticancer-Drug Screening. *JNCI: Journal of the National Cancer Institute* **82**, 1107-1112 (1990).
14. Vichai, V. & Kirtikara, K. Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nature Protocols* **1**, 1112 (2006).
15. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006 (2002).
16. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-82 (2011).
17. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**, 110-121 (2010).
18. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034-1050 (2005).
19. Hezroni, H., Koppstein, D., Schwartz, Matthew G., Avrutin, A. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11**, 1110-1122 (2015).
20. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32**, D493-D496 (2004).
21. Severin, J., Lizio, M., Harshbarger, J., Kawaji, H. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nature Biotechnology* **32**, 217 (2014).
22. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Research* **12**, 656-664 (2002).
23. Marchler-Bauer, A. & Bryant, S.H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* **32**, W327-W331 (2004).
24. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, D225-D229 (2011).

25. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Research* **43**, D222-D226 (2015).
26. Marchler-Bauer, A., Bo, Y., Han, L., He, J. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* **45**, D200-D203 (2017).
27. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).
28. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics* **5**, 1-19 (2004).
29. McWilliam, H., Li, W., Uludag, M., Squizzato, S. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013).
30. Li, W., Cowley, A., Uludag, M., Gur, T. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* **43**, W580-4 (2015).
31. Nishikawa, T., Ota, T. & Isogai, T. Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics* **16**, 960-7 (2000).
32. Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345-9 (2007).
33. Kang, Y.J., Yang, D.C., Kong, L., Hou, M. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* **45**, W12-w16 (2017).
34. Wang, L., Park, H.J., Dasari, S., Wang, S. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41**, e74-e74 (2013).
35. Arrial, R.T., Togawa, R.C. & Brigido Mde, M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics* **10**, 239 (2009).
36. Sun, K., Chen, X., Jiang, P., Song, X. *et al.* iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* **14**, S7-S7 (2013).
37. Panwar, B., Arora, A. & Raghava, G.P. Prediction and classification of ncRNAs using structural information. *BMC Genomics* **15**, 127 (2014).
38. Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**(2015).
39. Krzywinski, M., Schein, J., Birol, I., Connors, J. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).
40. Kel, A.E., Gößling, E., Reuter, I., Cheremushkin, E. *et al.* MATCH(TM): a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* **31**, 3576-3579 (2003).
41. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I. *et al.* TRANSFAC(®) and its module TRANSCompel(®): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108-D110 (2006).
42. Prestridge, D.S. Signal Scan - a Computer-Program That Scans DNA-Sequences for Eukaryotic Transcriptional Elements. *Computer Applications in the Biosciences* **7**, 203-206 (1991).
43. Prestridge, D.S. SIGNAL SCAN 4.0: additional databases and sequence formats. *Bioinformatics* **12**, 157-160 (1996).
44. Hooghe, B., Hulpiau, P., van Roy, F. & De Bleser, P. ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species. *Nucleic Acids Research* **36**, W128-W132 (2008).
45. Broos, S., Hulpiau, P., Galle, J., Hooghe, B. *et al.* ConTra v2: a tool to identify transcription factor binding sites across species, update 2011. *Nucleic Acids Research* **39**, W74-W78 (2011).
46. Kreft, Ł., Soete, A., Hulpiau, P., Botzki, A. *et al.* ConTra v3: a tool to identify transcription factor binding sites across species, update 2017. *Nucleic Acids Research* **45**, W490-W494 (2017).
47. Lindsay, S.J., Xu, Y., Lisgo, S.N., Harkin, L.F. *et al.* HDBR Expression: A Unique Resource for Global and Individual Gene Expression Studies during Early Human Brain Development. *Front Neuroanat* **10**, 86 (2016).
48. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580 (2013).
49. Wells, C.A., Mosbergen, R., Korn, O., Choi, J. *et al.* Stemformatics: Visualisation and sharing of stem cell gene expression. *Stem Cell Research* **10**, 387-395 (2013).
50. Leinonen, R., Sugawara, H., Shumway, M. & on behalf of the International Nucleotide Sequence Database, C. The Sequence Read Archive. *Nucleic Acids Research* **39**, D19-D21 (2011).

51. Pontén, F., Jirström, K. & Uhlen, M. The Human Protein Atlas—a tool for pathology. *The Journal of Pathology* **216**, 387-393 (2008).
52. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C. *et al.* Tissue-based map of the human proteome. *Science* **347**(2015).
53. Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D. *et al.* A subcellular map of the human proteome. *Science* **356**(2017).
54. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-4 (2016).
55. Giardine, B., Riemer, C., Hardison, R.C., Burhans, R. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-5 (2005).
56. Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783-5 (2010).
57. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86 (2010).

CHAPTER 3: CHARACTERISATION OF *CEROX1* AS A CONSERVED LONG NON-CODING RNA

INTRODUCTION

3.1 Many lncRNAs are poorly conserved

One of the impediments raised in arguing for the functionality of lncRNAs as a class of genes is their relatively low level of conservation. Around 80% of lncRNAs are primate-specific[1], having evolved after the Euarchonta-Glires divergence. The sequences of mammalian lncRNAs have evolved rapidly in comparison to protein-coding genes, resulting in poor sequence conservation, and implying that they are subject to only modest purifying selection[2]. This has led some to question to what extent this class of genes represents "transcriptional noise"[3,4]. Of the relatively few well-characterised lncRNAs, many are poorly conserved at the sequence level, with some instead demonstrating conservation of gene architecture[4]. Indeed, lncRNA GC content and splicing sites are more highly conserved than whole-gene sequence[5], as are lncRNA promoters[4], implying that the act of transcription, or the production of a transcript with potentially minimal conserved sequence is in some cases the instigator of function. However, analysis of 3122 mouse long ncRNAs by Ponjavic et al. revealed that the 4.1-5.5% of evolutionarily constrained sequence within these transcripts is of a similar proportion to the coverage of exons within protein-coding transcripts[4]. Interestingly, some syntenically-conserved loci demonstrate expression of non-sequence-homologous non-coding transcripts throughout evolution[6,7]. The extremely highly expressed lncRNA *MALAT1* was originally presumed mammalian-specific[8], but an orthologue in zebrafish, with sequence conservation in only a motif at the 3' end, has now been identified[9,10]. Function may be conferred by a relative small region of sequence conservation[5,11] amid an otherwise dramatic absence of conserved exonic architecture[7]. For examples, the vertebrate lncRNA *Cyrano*[10], in which a short region of high conservation (96% across 52 species from human to fishes) comprising a binding site for the equally conserved miR-7, appears to confer a conserved miRNA-regulating function[12-14].

In 2009, Ponjavic et al.[15] identified 659 murine long ncRNAs, from the set of 3122 introduced above, that demonstrated evolutionary constraint. Although mouse *Cerox1* (AK079380) was among those analysed, it was there classified as non-constrained, owing presumably to a murine lineage-specific deletion that

reduces mouse-human homology; this will be illustrated later. In fact, as this chapter will show, *CEROX1* is an example of a lncRNA that exhibits not only alignable sequence conservation throughout eutherian mammals, but also conservation of expression patterns on a sub-tissue and cell type basis. Conservation of tissue specificity is a trend amongst lncRNAs[1], and here is highly consistent with *CEROX1* performing a conserved function in the CNS.

3.1.1 *CEROX1* in context: *CEROX1* neighbours two protein-coding genes, *LMF1* and *SOX8*

3.1.1.1 Some lncRNAs regulate their neighbouring genes

Within their sample of evolutionarily constrained lncRNAs, Ponjavic et al.[15] identified several trends, including genomic proximity of lncRNAs to transcription factors with developmental roles in the CNS, and correlation of expression with these neighbouring protein-coding genes[15]. As will be described during this chapter, *CEROX1* exemplifies these traits: it is well-conserved evolutionarily, situated adjacent to a developmentally important transcription factor *SOX8*, and co-expressed with this gene in the CNS. However, it is unknown whether the activity of *CEROX1* is consistent with the hypothesis posited by Ponjavic et al.[15], that such lncRNAs may be involved with regulation of their neighbouring genes *in cis*. Other examples from within that set of conserved lncRNAs that have since been characterised include *Paupar*[16] and *Dali*[17], both of which act in *cis* to regulate their neighbouring transcription factors, Pax6 and Pou3f3, respectively. On the other hand, *lnc-SCA7* instead acts as a ceRNA to regulate expression of ATXN7 (ataxin 7) by competitive binding of miR-124[18]. Furthermore, both *Paupar*[16,19] and *Dali*[17] have further roles in regulating gene expression in *trans*; though the roles of lncRNAs may include regulation of neighbouring genes in *cis*, they are by no means restricted to this mechanism. Based upon the conserved similarity to its mouse orthologue *Cerox1*, it seems likely that *CEROX1* would conduct a role in post-transcriptional regulation in *trans*, not co-transcriptional regulation in *cis* (Sirey et al., 2019[20]).

3.1.1.2 Erroneous annotation linked *CEROX1* to its neighbour *LMF1*, lipase maturation factor 1

Human *CEROX1* is a multi-exonic RNA transcribed from within the terminal band p13.3 on the small arm of chromosome 16. It is transcribed divergently from SRY-related HMG-box protein 8 (*SOX8*), while downstream of *CEROX1* is lipase maturation factor 1 (*LMF1*) (Figure 3.1A-B). Human *CEROX1* was originally

annotated as a non-coding variant transcript of *LMF1*, owing to the annotation of a 5' alternative exon as part of several *LMF1* isoforms. However, closer inspection of these putative transcripts indicates that they are poorly supported predictions or transcripts targeted for nonsense-mediated decay (Ensembl, ensembl.org; Figure 3.1C), and that *CEROX1* is an intergenic RNA, unlike a number of other non-coding genes that exist within the introns of *LMF1* (Figure 3.1A). *LMF1* encodes a protein that is localised to the endoplasmic reticulum, where it is a required chaperone for the post-translational maturation of vascular lipase enzymes including lipoprotein lipase, hepatic lipase, and endothelial lipase[21]. Loss-of-function mutations in human *LMF1* are associated with combined lipase deficiency and hypertriglyceridemia[21,22], and these phenotypes are recapitulated in the *cld* (combined lipase deficiency) mouse. Homozygous *cld* mice develop lethal hyperchylomicronemia within several days after birth following the consumption of lipid-rich nursing milk, which leads to excessively elevated plasma triglyceride levels[23,24]. There is currently no evidence to suggest a regulatory or otherwise functional relationship between *CEROX1* and *LMF1*, though it cannot be ruled out and is worthy of greater investigation.

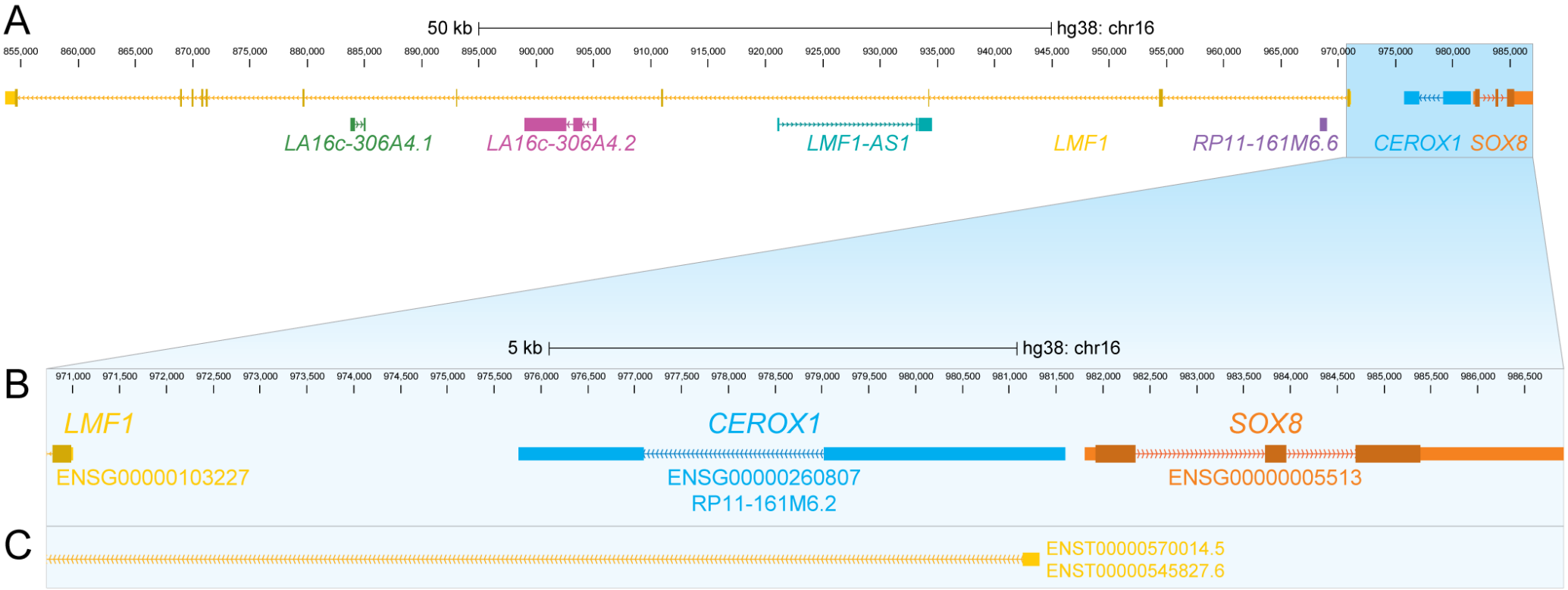


Figure 3.1: The genomic locus and annotation of human *CEROX1*. (A) *CEROX1*, also known as RP11-161M6.2 and ENSG00000260807, is situated on chromosome 16, between two protein coding genes, *SOX8* (SRY-related HMG-box protein 8) and *LMF1* (lipase maturation factor 1). There are a number of other non-coding transcripts within the locus of *LMF1*, including three intergenic long non-coding RNAs, and one antisense RNA that complements an *LMF1* exon. **(B)** A closer inspection of the *CEROX1* locus illustrates its closeness to *SOX8*, from which it is divergently transcribed from a bidirectional promoter region. **(C)** One alternative *CEROX1* exon was originally annotated as an exon of *LMF1*, and is annotated as such in at least two *LMF1* transcripts ENST00000570014.5 and ENST00000545827.6, but these are classed, respectively, as transcript support level 5 (having no evidence from mRNA or EST transcripts), and a transcript targeted for nonsense-mediated decay, suggesting that erroneous transcription events can bridge the *CEROX1* and *LMF1* loci but that the products are recognised by the cell as nonsense and subsequently destroyed. Gene annotations were accessed via the UCSC Genome Browser (genome.ucsc.edu) and illustrate those from GENCODE version 27, Ensembl 91, December 2017.

3.1.1.3 *SOX8*, SRY-related HMG-box protein 8 is implicated in diverse developmental functions

SOX8/*Sox8* shares a putative bidirectional promoter with *CEROX1*/*Cerox1* in both human and mouse (this will be the subject of greater investigation in Chapter 6). Characterised examples of bidirectionally transcribed pairs of genes include functionally-related genes whose expression levels must be co-ordinately regulated[25], such as the highly evolutionarily conserved H2A-H2B histone pair[26] and three pairs of collagen genes[27-29], and so the possibility that functionality of *CEROX1* is related directly or indirectly to the roles of *SOX8* must be considered. It is crucial to investigate any possible functions of *SOX8* that might be augmented with synergistic or antagonistic regulation by *CEROX1*.

SOX8 was originally identified due to its similarity to two other proteins, SOX9 and SOX10[30,31]. Together these form the E sub-family of SOX proteins, and it is highly likely that they have originated from a common genetic ancestor by gene duplication: single SoxE protein is present in invertebrates such as *Drosophila* (Sox100B), while all vertebrates have three, suggesting that these paralogues arose by gene duplication early in vertebrate divergence[32,33]. Despite identifying a broad and developmentally dynamic expression pattern, initial studies of Sox8-deficient mice indicated that the gene has no essential function, in stark contrast to the lethality of Sox9- and Sox10-deficient models. Close analysis has since identified various phenotypes: Sox8 knockout mice display reduced adipose tissue, leading to weight reduction[34], as well as abnormal bone development[35] and loss of fertility[36]. However, no overt effect is observed in many of the tissues in which Sox9 or Sox10 is co-expressed and in which deletion of these genes leads to lethal or otherwise profound developmental defects, such as the deletion of either Sox9 or Sox10 in the nervous system, which results in impaired myelination[37-39], or the absence of Sox9 in the developing gonad, which leads to sex reversal[40]. Given these observations, it is widely hypothesised that there is functional redundancy between Sox8 and Sox9, and Sox8 and Sox10[34,41-43]. Two or more of these proteins are expressed in many of the same tissues under similar developmental restrictions[34]. In the developing spinal cord, Sox9 is expressed in neuroepithelial cells that will form both motor neurons and oligodendrocytes. The appearance of Sox8 expression during or immediately after the onset of oligodendrocyte lineage specification is soon followed by Sox10 expression[39]. While deletion of Sox8 alone does not impede early spinal cord development[34], co-deletion of Sox8 worsens the effects of Sox9- or Sox10-deficient mouse models; for example, the few cells that are specified to the oligodendrocyte lineage in the spinal cord of Sox9-deficient mice[39] are no longer present upon co-deletion of Sox8[43]. A mild delay of postnatal terminal

oligodendrocyte differentiation is exhibited when Sox8 is absent, but heterozygous co-deletion of Sox10 leads to a dramatic delay[42]. This suggested a non-equal redundancy model, by which Sox9 or Sox10 may compensate for Sox8 absence, but not the other way around[43]. It was suggested that this may be in part due to the relatively low expression levels of Sox8[42], or a lower level of transcriptional activation capacity[44], however direct replacement of Sox8 into the Sox10 locus also fails to rescue Sox10-deletion phenotypes[33], confirming that they are not functionally equivalent. This has led some to question whether the continued expression of SOX8 in some tissues is simply an evolutionary remnant, and that perhaps of the three E group SOX proteins it most faithfully retains the expression pattern of an ancestral sequence[43]. The identification of a divergently transcribed genetic partner in the form of *CEROX1* raises the possibility that expression of the *SOX8* locus has been conserved by association through a shared promoter. This hypothesis will be re-visited following investigation of the expression pattern of *CEROX1*. However, not all functions of Sox8 may be compensated for by other E group SOX proteins, and indeed there are tissues in which only Sox8 of the three is expressed, such as striatal neurons[34]. Some functions appear to be complementary without redundancy. For example, both Sox8 and Sox9 are required for development of connective tissue: Sox9 is crucial for chondrocyte[45] production and differentiation in forming cartilage, and Schmidt et al.[35] showed that Sox8 is an essential regulator of osteoblast differentiation and so bone development, via the master regulator transcription factor Runx2. Sox8 negatively regulates Runx2 and so Sox8-deficient mice are osteopenic, due to premature Runx2 activation and osteoblast differentiation prior to sufficient proliferation. Transgenic constitutive expression of Sox8 during bone development, when it is usually downregulated, leads to osteoporosis, due to absence of Runx2 activity and subsequent failure of osteoblast differentiation and mineralisation. Additionally in the testes, which is the adult tissue outside the CNS in which SOX8/Sox8 expression is highest, the expression of Sox8 in Sertoli cells has been found to be essential in maintaining male fertility[36], and expression of Sox8 in spermatogonial cells is upregulated in response to Wnt signalling and modulates meiotic commitment during spermatogenesis[46].

3.1.2 Examination of the expression and genomic history of *CEROX1* will inform hypotheses of its functional relevance

While a putative function of *CEROX1* in regulating the electron transport chain has recently been described (Sirey et al., 2019[20]; this will be the subject of Chapter 5), its potential as an endogenous regulator, and the extent to which its putative function is consistent with its expression has not yet been fully established. In this chapter, a full characterisation of the *CEROX1* locus and of *CEROX1* expression will be undertaken, with a focus upon evolutionary conservation, the most widely lauded hallmark used in inferring functionality. Following an appraisal of the architecture of the *CEROX1* locus using bioinformatic and experimental molecular biology approaches, these will be extended to: a) a comparison of human *CEROX1* with the already partly characterised mouse *Cerox1*; and b) a more distal evolutionary investigation of the gene throughout mammals. By ascertaining the spatiotemporal expression patterns of *CEROX1* and its conserved orthologues, further hypotheses will be formed about the role that the lncRNA may play in both normal and disease physiology.

RESULTS

3.2 Transcription at the *CEROX1* locus produces a number of isoforms

3.2.1 Transcriptional activity at the *CEROX1* locus is supported by epigenetic evidence

In order to garner initial evidence to support the annotation of the *CEROX1* gene, characteristics of the locus including epigenetic features were examined using the UCSC Genome Browser. All evidence supports the classification of *CEROX1* as a classical promoter-associated lncRNA[47].

3.2.1.1 *CEROX1* is one of many genes with a CpG-rich promoter

Examination of the UCSC Genome Browser (genome.ucsc.edu) indicates that the putative bidirectional promoter shared by *SOX8* and *LMF1* is situated within a large CpG island. CpG islands are stretches of sequence that have resisted so-called *CG suppression*. Over 80% of CG dinucleotides are methylated on the cytosine residue to form 5-methylcytosine. This is advantageous during DNA replication, since methylation or absence thereof identifies the old and newly synthesised strands. However, 5-methylcytosine can undergo spontaneous mutation by deamination to form thymine and over the course of evolution this process has depleted the incidence of CG dinucleotides to around 1% of the genome, much lower than the ~4.4% that ought to occur by chance[48]. Stretches of DNA that remain largely non-methylated are not subject to this mutating process, and so retain their CG dinucleotides, being known as *CpG islands*. These regions are therefore indicative of low DNA methylation status, and higher transcriptional activity. Around 70% of human promoters are situated within these CpG-rich sequences[49]. The CpG island surrounding the *CEROX1-SOX8* promoter is over 5.4 kb in length, and contains 448 CG dinucleotides (excluding repeat regions), giving it an observation ratio of 0.77 (0.6-0.65 being a commonly-recognised threshold[50]) (Figure 3.2A). Only 1.11% of human CpG islands are longer than 3 kb[51], making this CpG island amongst the longest in the genome. Promoters situated within long CpG islands (those longer than 2 kb) tend to: a) have a higher number of RNA polymerase II binding sites; b) be associated with genes of intermediate levels of tissue specificity; and c) exhibit roles in development and gene regulation[52]. Evidence in support of the latter two hypotheses with regards to *CEROX1* will be highlighted later.



Figure 3.2: Epigenetic features of the *CEROX1* locus support its transcriptional activity. (A) The *CEROX1-SOX8* promoter sits within a long CpG island of over 5.4 kb. **(B)** *CEROX1* expression, assessed by RNA-seq (Transcription), is consistent with post-translational histone modifications in HSMM (human skeletal muscle myoblast) and K562 (immortalised erythroleukemia) cells. In HSMM cells, the *CEROX1-SOX8* promoter is flanked by a peak of $H3K4Me_3$, and the *CEROX1* gene body is marked by $H3K36Me_3$, together indicating that the locus is transcriptionally active, while the inactivity of the locus in K652 cells is accompanied by no $H3K4Me_3$ peak and little $H3K36Me_3$. RNA-seq and ChIP-seq data were generated by the Wold and Bernstein labs of the ENCODE Project Consortium, respectively. Those data, as well as CpG island annotation, based on software by G. Miklem and L. Hillier, were accessed via the UCSC Genome Browser.

3.2.1.2 The *CEROX1* locus exhibits histone modifications that are consistent with its transcriptional activity on a cell type-by-cell type basis

Chromatin immunoprecipitation (ChIP)-seq data from the ENCODE project indicate that the histones of the chromatin of the *CEROX1* gene are, like all others, subject to post-translational modifications, and that these correlate with transcriptional activity of the locus (Figure 3.2B). In HSMM (human skeletal muscle myoblast) cells, RNA-seq indicates that *CEROX1* is expressed highly. The *CEROX1-SOX8* promoter region is flanked by peaks of histone 3 lysine 4 trimethylation ($H3K4Me_3$), which is associated with active promoter regions, as well as $H3K4Me_2$, which is found at active and poised promoters. $H3K4Me_2$ has been found to correlate highly with the position of transcription factor binding regions[53] and this may be the reason that its prevalence within gene bodies (thereby indicating a larger area of $H3K4Me_2$ coverage and an accordingly larger TBFR) is an indicator of tissue-specific expression[54], the control of which relies upon a greater number of transcriptional regulatory signals. The gene body of *CEROX1* (and *SOX8*) is covered by $H3K36Me_3$, which is a mark of active transcription. Around the 3' end of *CEROX1*, $H3K4Me_1$ is prevalent, alongside H3K27 acetylation (H3K27Ac): these are indicators of an active enhancer element (Figure 3.2). On the other hand, in K562 (immortalised erythroleukaemia) cells, where transcription of *CEROX1* is very low, there is no peak of $H3K4Me_3$ at the *CEROX1-SOX8* promoter, and lower prevalence of $H3K36Me_3$, $H3K4Me_2$, $H3K4Me_1$, and H3K27Ac throughout the locus. Transcriptional activity at the *CEROX1* locus thereby correlates with the presence or absence of these histone modifications.

3.2.1.3 The *CEROX1* locus is in proximity to a putative enhancer element but *CEROX1* is not an eRNA

The presence of a tissue-specific region of $H3K4Me_1$ density around the 3' end of *CEROX1*, especially in conjunction with H3K27Ac, which is a marker of active, as opposed to poised, enhancers[55], suggests that the locus contains an enhancer element whose activity correlates with *CEROX1* expression. This might lead to the suggestion that *CEROX1* is an enhancer RNA (eRNA). However, the putative enhancer in question is at the 3' end of *CEROX1*, and therefore it is clear that it is not responsible for direct transcriptional initiation of this promoter-associated lncRNA, which is instead expressed from the bidirectional promoter shared with *SOX8*. However, a relationship between the activity of this putative enhancer element and the transcription and expression of *CEROX1, SOX8*, or *LMF1* (or even a more distal gene) cannot be ruled out.

3.2.2 Retrotransposon insertion might have contributed to *CEROX1* evolution

Several studies have identified that retro-transposable elements are common within the transcriptional initiation regions of both enhancer- and promoter-driven lncRNAs[7,56]. While BLAST (blast.ncbi.nlm.nih.gov) searches of the *CEROX1* sequence indicate no significant hits, a search against the conserved domain database (ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)[57-60] reveals significant similarity within the 5' proximal ~1.8 kb of the transcript to the large tegument protein UL36[61] (*E* value = 5.64×10^{-3}) and the transcriptional regulator ICP4[62] (*E* value = 8.89×10^{-4}), both of which are indicative that this region contains a retrotransposed herpes simplex virus (HSV) remnant (Figure 3.3).

Additionally, the 3' end of first exon is situated within a type 2A LINE (long interspersed nuclear element) (Repeat Masker, GIRI Repbase, girinst.org/repbase)[63] (Figure 3.3). This element therefore includes at least one splicing site for each *CEROX1* isoform, and it is possible that it has shaped the exonic architecture of the transcript. The significance of these two elements with respect to the evolution of *CEROX1* will be further discussed in Section 3.3.1.2.

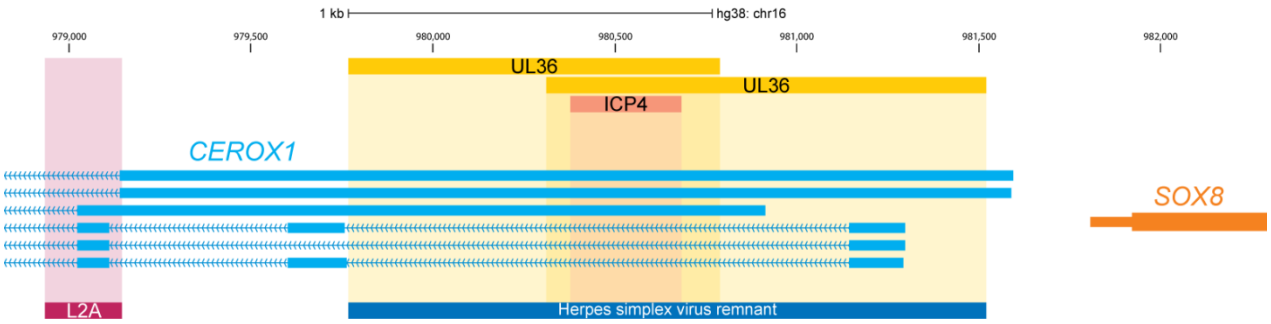


Figure 3.3: The 5' end of *CEROX1* contains two retrotransposed elements. Much of the 5' exon of *CEROX1* bears resemblance to conserved domains found in the Herpes simplex virus proteins UL36 and ICP4, suggesting that the locus includes a retrotransposed fragment of viral genome. Additionally, a L2A family LINE element sits at the end of the 5' exon of *CEROX1*, hosting the splice site to produce the largest intron present in most isoforms, as well as an additional splice acceptor site in several isoforms.

3.2.3 At least twelve isoforms of *CEROX1* are annotated at present

The current GENCODE annotation (version 29, Ensembl 94, October 2018) of the *CEROX1* locus includes eight different transcripts, resulting from four putative transcriptional start sites (TSSs) and several alternative splicing sites (Figure 3.4A). Preliminarily, two major bioinformatic datasets were consulted in order to assess which of the *CEROX1* isoforms are genuinely expressed, and which, if any, dominates in particular tissues or cell types.

3.2.3.1 *In silico* approaches offer differing conclusions regarding *CEROX1* isoform dominance

One of the most comprehensive human gene expression databases is the NIH GTEx (Genotype-Tissue Expression) project, which provides high-quality RNA-seq data for a diverse selection of post-mortem human tissues. Isoform specific data are available from the GTEx project, and based upon the summation of expression for transcripts sharing a TSS these indicate that the second TSS (here denoted β) is the most active, with roughly equal expression promoted from the first and third TSSs (α and γ, respectively), and little from the most downstream TSS (δ) (Figure 3.4B).

However, RNA-seq assesses the steady-state equilibrium of RNA levels, and while informative this does not necessarily reflect the actual level of expression at the level of transcriptional activity at alternative TSSs. In order to focus specifically on isoform production, CAGE (cap analysis of gene expression)[64] data was also consulted. The FANTOM5 (Functional Analysis of the Mammalian Genome 5) CAGE-seq project combined CAP trapper technology[65,66], which utilises a 5'-methylguanosine-specific diol in biotinylation and subsequent isolation of full-length cDNAs, with HeliScope single molecule sequencing technology[67]. The resulting CAGE dataset includes a diverse range of human and mouse cell types and tissue samples, comprising 975 human and 399 mouse samples in total[68]. Data from the human FANTOM5 dataset (accessed via Zenbu[69], fantom.gsc.riken.jp/zenbu) indicate that, of the four putative transcriptional start sites (TSSs), the bulk of transcriptional initiation takes place at the α TSS, followed by β, with little signal observed at γ, and none at the most downstream δ (Figure 3.4A,C). Indeed there is no CAGE-tag read within 380 bp of the δ TSS in any samples in the dataset. The two datasets thereby prompt different conclusions as to the relative expression of the *CEROX1* isoforms (Figure 3.4D).

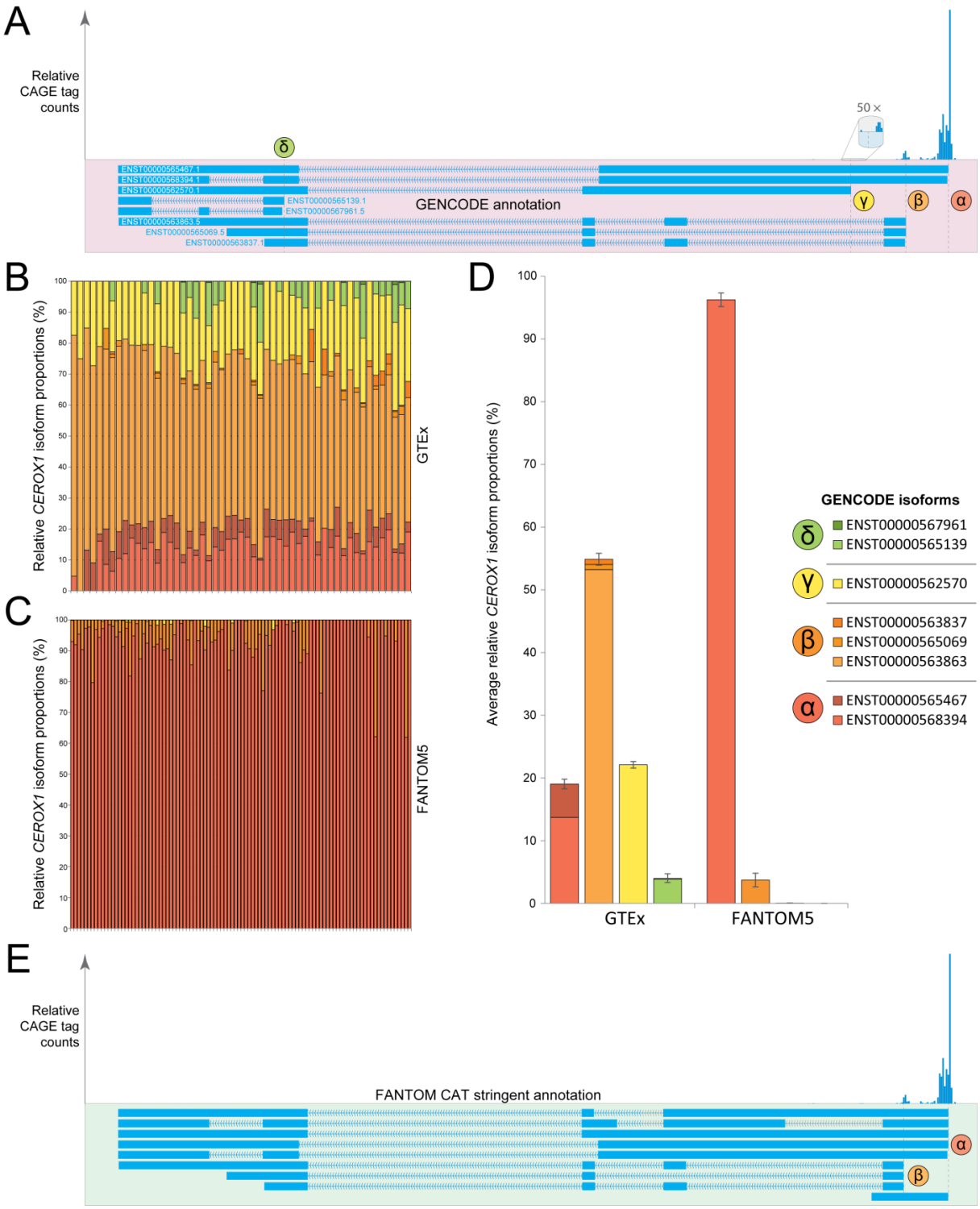


Figure 3.4: Data from the CAGE-seq FANTOM5 project and the GTEx RNA-seq project offer differing results as to the relative activities of the transcriptional start sites of the *CEROX1* locus. (A) Mapping of CAGE-tag peaks against the annotated isoforms of *CEROX1* indicates a very strong peak at the most upstream TSS (α), with progressively smaller peaks visible at the two subsequent TSSs downstream (β and γ). No CAGE-tag reads were detected within 200 bp of the most downstream putative TSS (δ). **(B)** The relative bias of the α TSS-aligning CAGE-tag peak is clearly visible when expressed as a proportion of the total *CEROX1* CAGE-tag reads observed for the top 100 FANTOM5 samples (sorted by total *CEROX1* expression). The isoform specific expression data derived from the GTEx dataset **(C)** illustrate a very different trend, with the two isoforms originating from the α TSS comprising less than 30% of total *CEROX1* expression across all 53 tissues represented. **(D)** On average, only 19% of *CEROX1* transcripts originate from the α TSS according to the GTEx data, with the β TSS producing the greatest proportion of transcripts (54.9%), while the FANTOM5 data indicate an average of wildly different value of 96.2% for TSS α. **(E)** The FANTOM CAT annotation of transcripts from the *CEROX1* locus has been amended to reflect the CAGE-derived TSS bias.

3.2.3.2 The discrepancy in isoform dominance may be addressed by amended annotation

Transcription activity at a locus, which the FANTOM5 project aims to measure, is not necessarily directly in proportion with steady-state transcript concentration, but the CAGE-seq approach actually measures the latter, albeit focusing only on the very 5' end, as a proxy for transcriptional start site usage. This does not, therefore, explain the discrepancy. If a post-transcriptional process were to alter the isoform bias of the gene, this would be reflected in the RNA-seq data of the GTEx project, but also to some degree in the CAGE-seq data of the FANTOM5 dataset. The discrepancy might reflect in part the inclusion of many cell lines in the FANTOM5 sample set, unlike the GTEx samples that are exclusively post-mortem tissues.

Although the GTEx data are a very robust assessment of gene expression across the population, analyses of RNA-seq data on an isoform-by-isoform basis are only as accurate as the annotation of those transcripts. In addition to a thorough examination of the human transcriptome, the FANTOM5 project has focused on lncRNAs, integrating CAGE data with existing transcript model collections such as GENCODE, as well as an RNA-seq-based model transcriptome from FANTOM5 samples. In order to maximise accuracy of transcriptional start site annotation, transcripts and associated CAGE data were assessed with the novel TIEScore (transcription initiation evidence score) metric, which accounts for expression and architecture of a transcript as well as CAGE cluster proximity to the 5' end[56]. This metric was used to define 'stringent', 'robust' and 'permissive' sets of genes in the meta-assembly known as FANTOM CAT (CAGE-associated transcriptome)[56]. Almost 28,000 lncRNAs were defined within the robust FANTOM CAT dataset, and the 5' ends of these are more accurate and more completely annotated than those in other catalogues[56]. In light of the FANTOM5 data that suggests that the α TSS is by far the dominant TSS and that γ and δ are poorly active or inactive, the annotation of *CEROX1* isoforms has been amended in FANTOM CAT (fantom.gsc.riken.jp/cat/v1). The 'stringent' model for *CEROX1* comprises only isoforms initiated from the two upstream-most TSSs, and this is far more consistent with CAGE data (Figure 3.4E).

3.2.4 Isoforms were investigated *in vitro* using rapid amplification of cDNA ends (RACE)

Ultimately, it was decided that the most effective approach for assessing the presence of putative isoforms within the cellular models and tissues of interest was to apply to each a non-bioinformatic experimental approach, which centred on the molecular biology technique of RACE. Appending a known sequence onto

the 3’ or 5’ end of a RNA of incomplete sequence allows for PCR amplification across the unknown termini, using a gene-specific primer complementary to the region of known sequences, and a universal one complementary to the appended sequence. A variety of RACE approaches were required to identify the transcripts produced from the *CEROX1* locus.

3.2.4.1 *CEROX1* is highly refractory to traditional 5’ RACE protocols and optimisation strategies

Initially, in order to assess the presence of multiple TSSs *in vitro*, RLM (RNA ligase-mediated) -RACE was applied to the 5’ end of *CEROX1* in total RNA from HEK293T cells and human brain using the GeneRacer™ Kit (Invitrogen); this was chosen due to its previous success in amplifying the 5’ end of mouse *Cerox1*. In this protocol, de-capping of the total RNA sample is used to form an accessible terminus at the 5’ end, allowing ligation of an RNA oligonucleotide of known sequence. This is followed by reverse transcription using a gene-specific primer in order to heighten specificity, and then PCR using a gene-specific primer, and one complementary to the previously-ligated oligonucleotide. Initial RACE-PCR using the recommended Platinum® Taq DNA Polymerase produced several products, but subsequent sequencing revealed that none represented *CEROX1*. Several other products were amplified that included the GeneRacer™ Oligo adapter sequence, and included the 5’ end of an annotated transcript, suggesting that the earlier RNA-manipulating enzymatic steps were successful. It was hypothesised that the failure to produce any *CEROX1* products was due to the highly refractory nature of the GC-rich 5’ end of the transcript to PCR. Many other experiments during the course of this body of work have indicated that it is a very difficult transcript to target by PCR. Comparison to the GC content of the remainder of the human transcriptome reveals that *CEROX1* is exceptionally GC-rich, both averaged across its full length and especially at the 5’ end (Figure 3.5A-C).

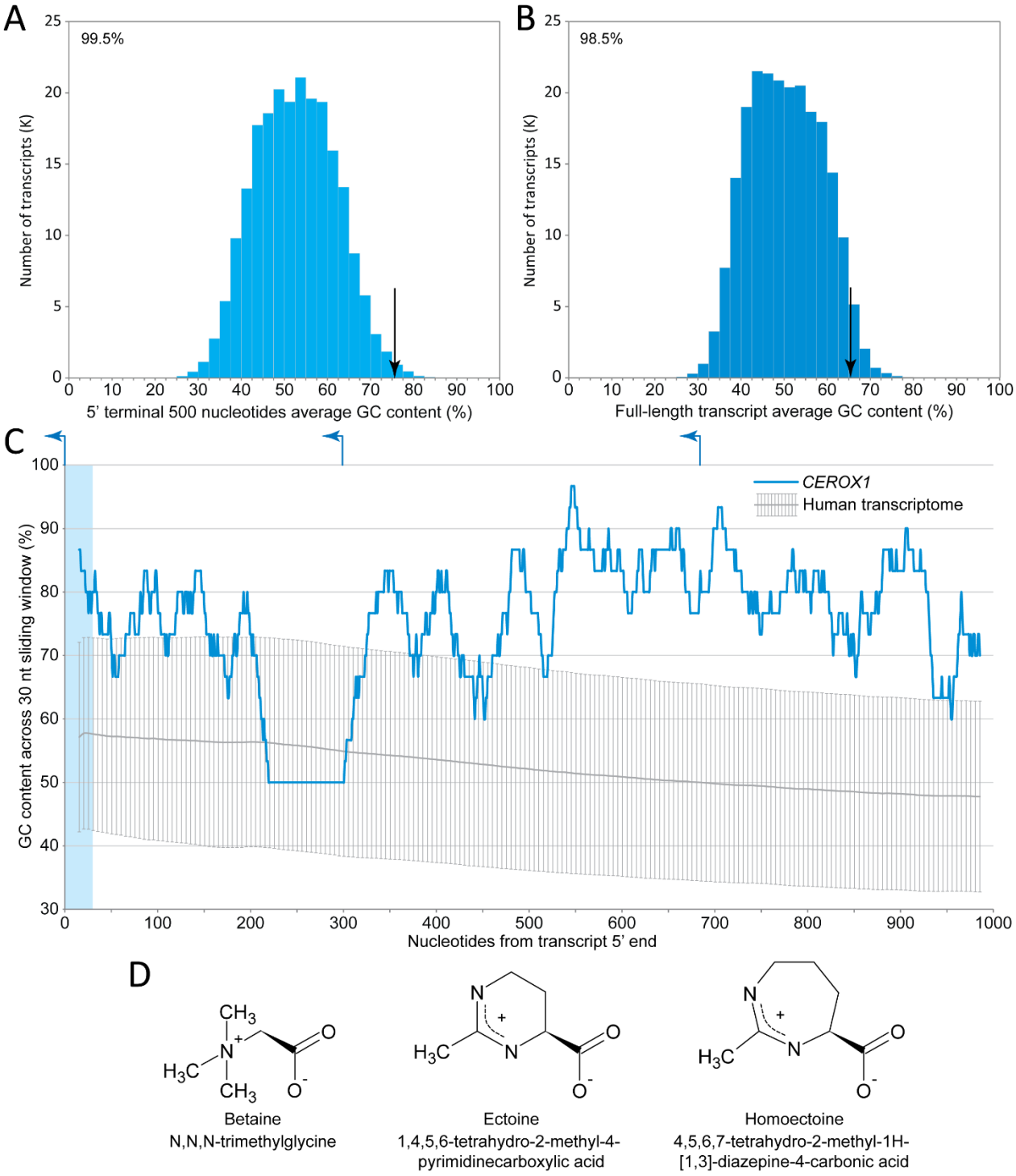


Figure 3.5: The high GC content of *CEROX1* prompted the use of DNA denaturation-promoting additives. Distributions of GC content across the human transcriptome (219,288 transcripts, GENCODE v27, Ensembl 90), both localised to the 5'-proximal 500 nucleotides **(A)** and across entire transcripts **(B)**, reveal that *CEROX1* has an exceptionally high GC content that places it within the top 0.5-1.5% of all transcripts. **(C)** A distribution of GC content for the 5'-proximal 1 kb of sequence for all 200,846 transcripts in GENCODE v28, averaged over 30 nucleotide sliding windows (light blue box for scale) and shown ± the standard deviation. The TSSs for CEROX1 are shown with arrows. The putative 5' ends of *CEROX1* are clearly highly GC-rich relative to the transcriptome. A microsatellite of ~110 nt is responsible for the plateau of low GC content around 200 nt from the 5'-most end. **(D)** As a result, adjuvants that promote DNA melting were added to RACE-PCR reactions. Betaine, ectoine, and the synthetic homoectoine share some common chemical features: they all exist in aqueous solution as zwitterions, with carboxyl anions and quaternary ammonium cations.

3.2.4.2 The final 5' RACE protocol for *CEROX1* required high-temperature polymerase enzymes and a combination of natural and synthetic osmolyte additives

The failure to amplify *CEROX1* 5' end amplicons led increasingly to the conclusion that the GC-richness of *CEROX1*, or a putative inherent difficulty in its secondary structure, was likely interfering. After many attempts (detailed in Appendix A3), a highly optimised RACE protocol adapted from that of Shi et al.[70] for heavily GC-rich genes, utilising the high temperature enzymes ThermoScript™ Reverse Transcriptase and Phusion® High-Fidelity Polymerase along with betaine and a rare synthetic osmolyte homoectoine[71] (kindly provided by Professor Erwin Galinski, University of Bonn, Germany) (Figure 3.5D) eventually produced RACE products that sequencing revealed to be the furthest upstream 5' end (Figure 3.6A). This is in accordance with the updated annotation of the FANTOM CAT database above (Figure 3.4E).

3.2.5 3' RACE supports the existence of at least two transcriptional end sites for *CEROX1*

In addition to the multiplicity of 5' TSSs and splicing sites, differential termination at the 3' terminus creates further variation of *CEROX1* isoforms. Accordingly, this was investigated using 3' RACE (Figure X.A). Briefly, a primer containing a known sequence and a tract of deoxythymidine (dT) nucleotides, therefore complementary to the polyadenylated (polyA) tail of mature mRNA was used to prime the sequence for reverse transcription. The resulting cDNA was then amplified using a gene specific primer and a primer complementary to the reverse transcription primer.

These experiments confirmed the presence of multiple 3' transcriptional termination sites (TTSs), and differential splicing patterns following the longest intron (Figure 3.6B). One of the cloned products is not consistent with any of the eight GENCODE-annotated *CEROX1* isoforms, again supporting the need for amended annotations such as those made by the FANTOM CAT project (Figure 3.6C).

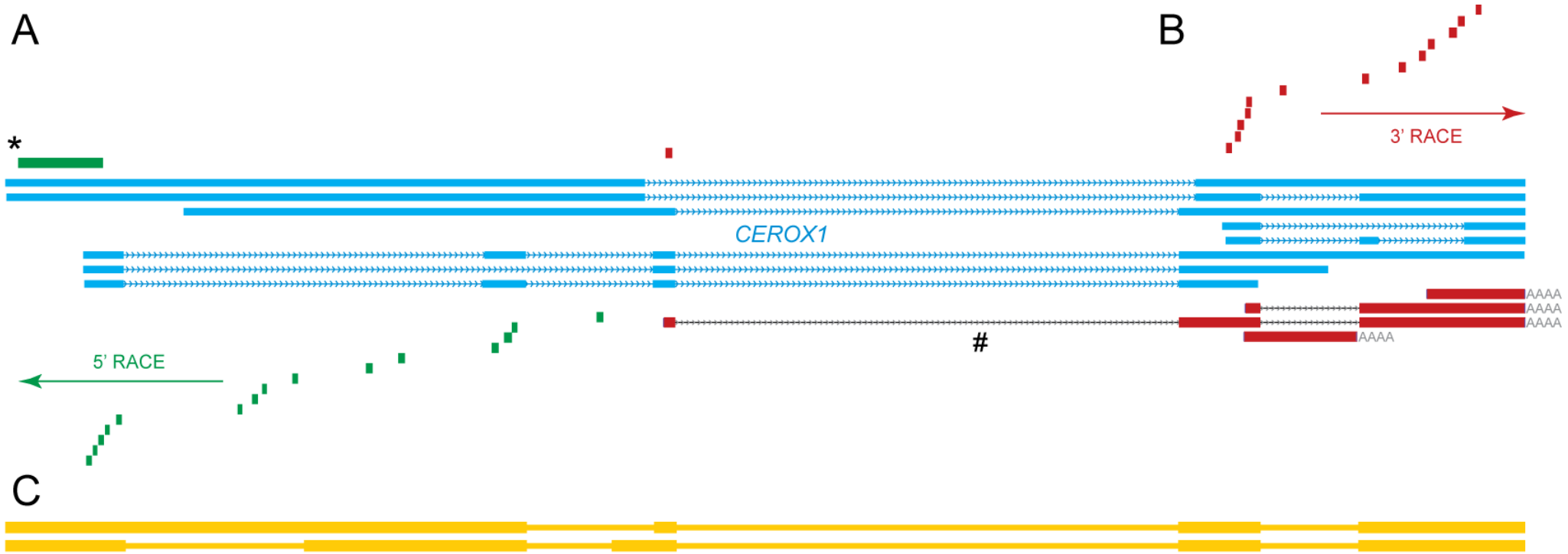


Figure 3.6: Rapid amplification of cDNA ends (RACE) confirms 3' and 5' ends of multiple *CEROX1* isoforms. (A) A single 5' end (green) was confirmed by 5' RACE, supporting the upstream-most transcriptional start site. The cloned product (*) is slightly short, likely due to light RNA degradation. Despite the use of 15 primers (small green rectangles) across the 5' half of *CEROX1*, neither of the other 5' ends was represented by RACE-PCR clones. **(B)** 3' RACE was more successful in identifying multiple transcriptional termination sites and splice variants, detecting at least two different TTSs. The longest 3' RACE product (#) does not fully correspond to any of the eight GENCODE-annotated isoforms that are illustrated here. It does, however, correspond to two isoforms annotated in FANTOM CAT **(C)** (yellow).

3.3 Human *CEROX1* and mouse *Cerox1* demonstrate a high level of conservation

3.3.1 The human and mouse *CEROX1* loci are syntenic and partially conserved

The human *CEROX1* gene is flanked by two protein-coding genes, lipase maturation factor 1 (*LMF1*) and SRY-related HMG-box protein 8 (*SOX8*). This arrangement is syntenic to that of mouse *Cerox1*, which is flanked by the murine orthologues of the same genes, *Lmf1* and *Sox8* (Figure 3.7A-B). These genes are amid a syntenic block shared by mouse chromosome 17 and human chromosome 16[31].

3.3.1.1 Epigenetic marks indicate conservation of a putative enhancer element

Interestingly, another non-coding transcript is annotated in the intergenic space between *Cerox1* and *Lmf1* in mouse (Figure 3.7B). This transcript, AK033307, was annotated in GenBank as a result of FANTOM CAGE data, but is very lowly expressed, and absent from GENCODE and Ensembl. Three or more CAGE tags were identified in only five of 1195 samples (0.42%) in the FANTOM5 CAGE Phase 1 and 2 dataset, with the highest expression being 1.05 relative log expression units (rle), while *Cerox1*, for comparison, has a maximal expression of 121.08 rle (115-fold higher) (CAGE-seq data obtained from Zenbu[69], fantom.gsc.riken.jp/zenbu). Examination of the TSS for this transcript reveals that it falls within a tissue-specific peak of $H3K4Me_1$ and H3K27Ac syntenic to that identified in human (Figure 3.2). ChIP-seq data from the ENCODE project illustrate that $H3K4Me_1$ and H3K27Ac peaks are present in testis, the tissue from which two of the five samples in which AK033307-associated CAGE tags were found was derived, while no peak is visible in liver (Figure 3.7C). This tissue specificity, along with its low expression level, strongly suggests that AK033307 is a transient eRNA associated with the putative enhancer identified previously.

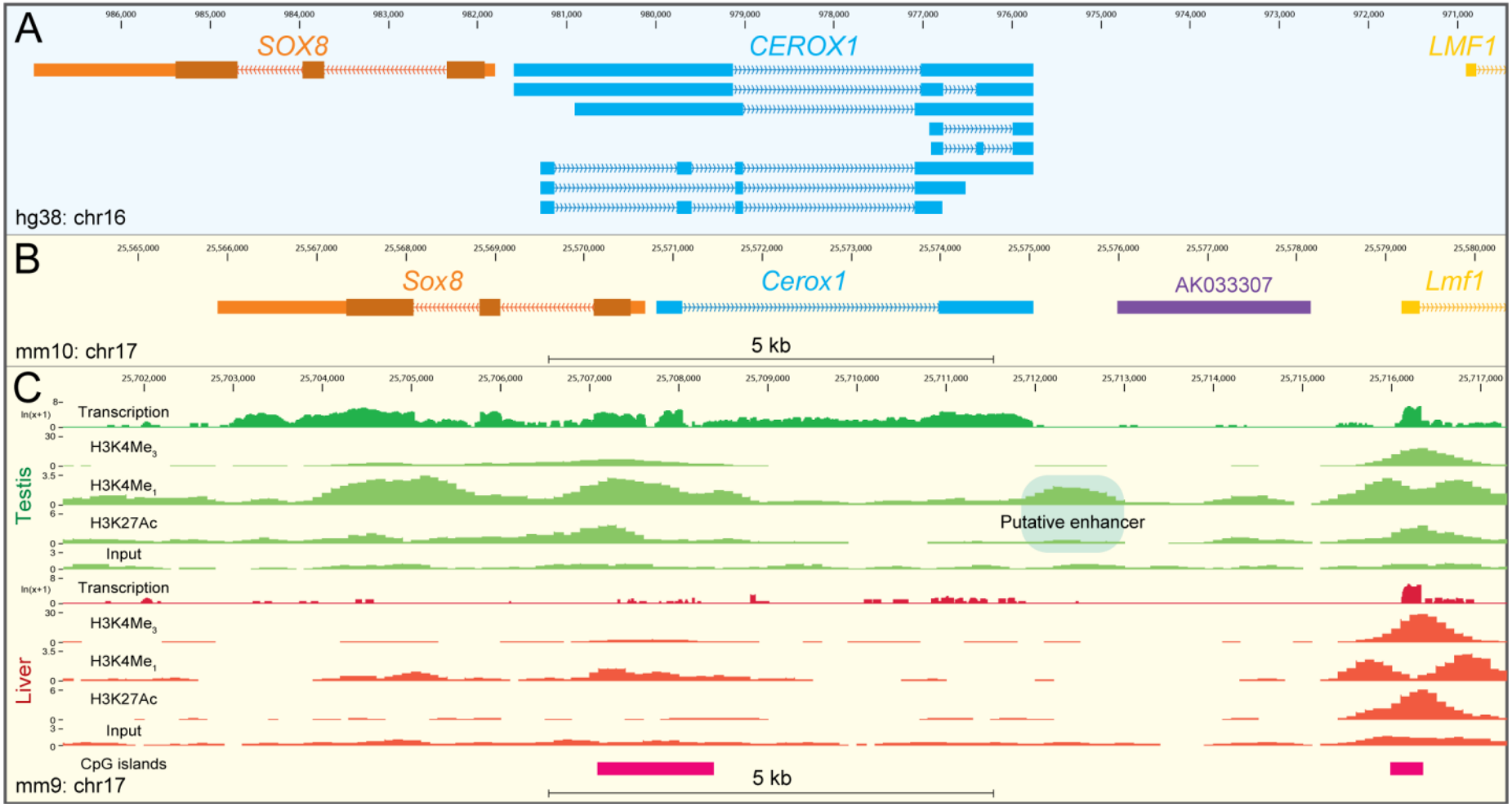


Figure 3.7: The loci of human *CEROX1* and mouse *Cerox1* are syntenic. Both human *CEROX1* **(A)** and mouse *Cerox1* **(B)** are flanked by two protein-coding genes, *SOX8*/*Sox8* and *LMF1*/*Lmf1*. An additional non-coding transcript AK033307 is annotated in the intergenic region between *Cerox1* and *Lmf1* in mouse[66]. **(C)** As in human, expression at the mouse *Cerox1* locus is in accordance with traditional epigenetic markers: its promoter lies within a CpG island, and transcriptional activity correlates with the presence of $H3K4Me_3$ and $H3K4Me_1$ peaks. A tissue-specific peak of $H3K4Me_1$ and H3K27Ac between *Cerox1* and AK033307 is indicative of an active enhancer element, possibly associated with the transcriptional initiation of AK033307, which may be an eRNA. ChIP-seq data were generated by the Ludwig Institute for Cancer Research, University of California San Diego, as part of the ENCODE Project Consortium, and accessed via the UCSC Genome Browser.

3.3.1.2 The architecture of mouse *Cerox1* is different to human *CEROX1*, owing to a genomic deletion

Notably, the 5' exon of mouse *Cerox1* is substantially shorter than that of its human orthologue. It appears that the mouse *Cerox1* locus has undergone a deletion of ~2 kb that covers ~1.5 kb of the sequence comprising the 5' exon in human *CEROX1*. Possibly due to the inclusion of the L2A LINE and associated splicing sites in this deletion, the splicing site has moved ~0.5 kb upstream, resulting in the loss of ~2 kb from the exon (Figure 3.8A). The singular splicing donor site in the absence of the LINE supports the hypothesis that this element has contributed to the isoform diversity of human *CEROX1*, as is the case for SINE (short interspersed nuclear repeat) and LTR (long terminal repeat) elements in the lncRNA *Cyrano*[7].

Based upon examination of the sequence alignment of the locus in other Rodentia species (Figure 3.8A), the deletion is not observed in the Hystricomorpha clade, with homology to the human locus seen throughout the 5' exon (e.g. guinea pig). The deletion is present in rat, and appears to be present in golden hamster, prairie vole, and lesser Egyptian jerboa. However, the absence of the deletion in Chinese hamster suggests that the apparent presence of the deletion in these other Myomorpha species is due to poor quality of assembly (which likely used the mouse genome as a reference), and that the deletion is specific to the Muroidea clade (Figure 3.8B).

With the mechanism of action of human *CEROX1* currently uncertain, this deletion may be useful later in identifying the functional elements of the lncRNA. For example, the absence of the LINE element in mouse precludes its acting in a conserved manner upon translation machinery akin to the previously described retrotransposon-containing lncRNAs[72].

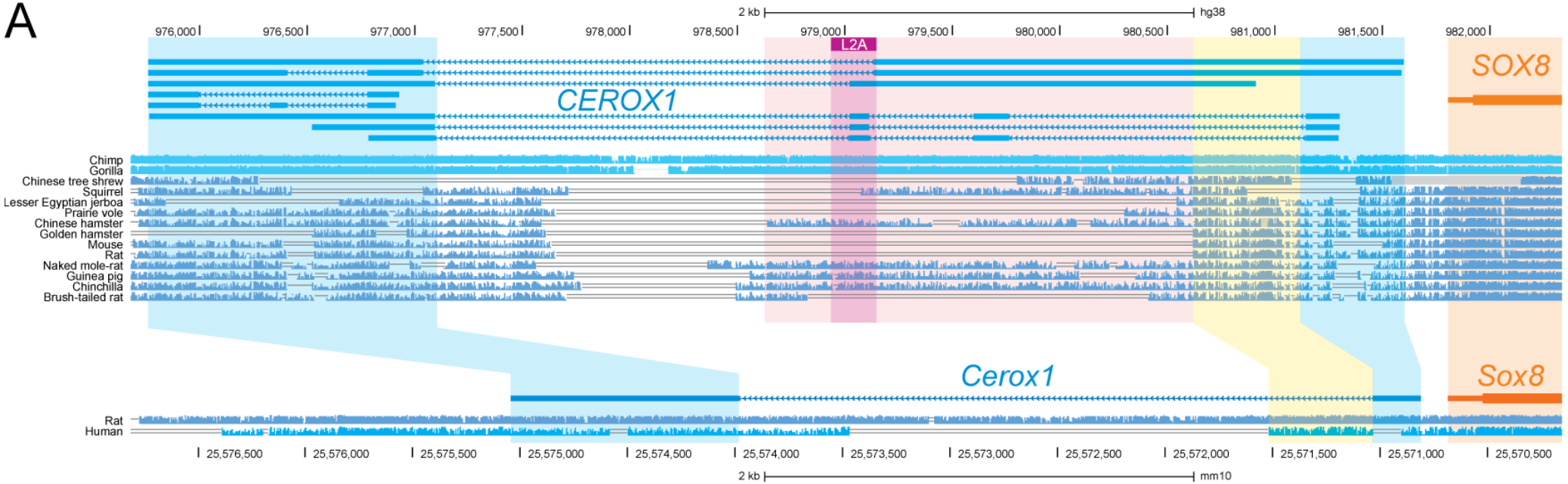




Figure 3.8: Genomic deletion in the *Cerox1* locus of some rodent species in has altered the architecture of the lncRNA. (A) While the majority of the 3' exon (left) sequence is homologous between the human and mouse lncRNAs (blue shading), a deletion of ~2 kb (pink; including the LINE, dark pink) has altered the mouse *Cerox1* locus. The splice site has moved upstream an additional ~0.5 kb, through a region of conservation (yellow). **(B)** Shown in grey are the species in which the above deletion is found. Although most Myomorpha species contain the deletion, suggesting that it occurred at the top of that clade (#), the absence of the deletion in Chinese hamster suggests that the deletion occurred in the Muroidea divergence (*), and that poor genome assembly is the reason for the absence in other species. Multiz alignments were obtained from the UCSC Genome Browser (genome.ucsc.edu).

3.3.2 The overall expression profile of human *CEROX1* mirrors that of mouse *Cerox1*

Previous investigation of mouse *Cerox1* with RT-qPCR using a mouse tissue panel and CAGE-tag data from the FANTOM5 project[68,69] indicated that the lncRNA is highly expressed amongst the murine lncRNA population, and across a range of adult mouse tissues (Sirey et al., 2019[20]). The human lncRNA population demonstrates a very high level of variation in expression level, wider than that of protein-coding genes[73], with some being expressed at less than one molecule per cell on average, such as *HOTTIP*[74], and some being very highly expressed, such as *MALAT1*[75], among the most highly expressed human genes[76,77] (top 1%)[78]. In order to investigate the expression level of human *CEROX1*, RNA-seq data from the GTEx project[78] were examined. It was found that the expression level of *CEROX1* is within the top 0.3% of that for the 5161 lncRNAs assayed, and higher than that of 87.5% of protein-coding genes (Figure 3.9A). This identifies human *CEROX1* as an extremely highly expressed transcript, a decile above the expression of mouse *Cerox1*, which is still more highly expressed than 87% of other mouse lncRNAs (Dr Wilfried Haerty) (Figure 3.9B).

With regards to its tissue localisation, RT-qPCR using a human tissue RNA panel indicated a near-ubiquitous expression profile for *CEROX1* (Figure 3.9C) (*these RT-qPCR data were generated prior to the work of this doctoral thesis, and have been incorporated into another degree submission, as stated previously in the Declaration of Authorship*). Expression was found to be highest in the brain and central nervous system, followed by testes. Similarity in gene expression patterns between the brain and testes has been described before[79,80], as has the preferential expression of lncRNAs within these tissues[81-83]. The transcript was not detected in the liver, however. Again, these observations parallel the findings of Sirey et al.[20] that show that the expression of mouse *Cerox1* (Figure 3.9E) is also ubiquitous across adult tissues, but elevated in the brain, and minimal in the liver. Human GTEx data[78], which include a greater number of tissues as well as specific brain regions, support these results (Figure 3.9D), again indicating enrichment in the central nervous system, including all regions of the brain assayed, and an absence or very low level of expression in the liver. The visually apparent enrichment of *CEROX1* within the central nervous system is supported by sample ontology analysis within FANTOM CAT[56], with significant enrichment of expression exhibited for “brain” (21.1-fold, $p = 2.46 \times 10^{-42}$), “central nervous system” (17.9-fold, $p = 1.86 \times 10^{-37}$), and specific brain regions, e.g. “striatum” (13.1-fold, $p = 2.65 \times 10^{-6}$).

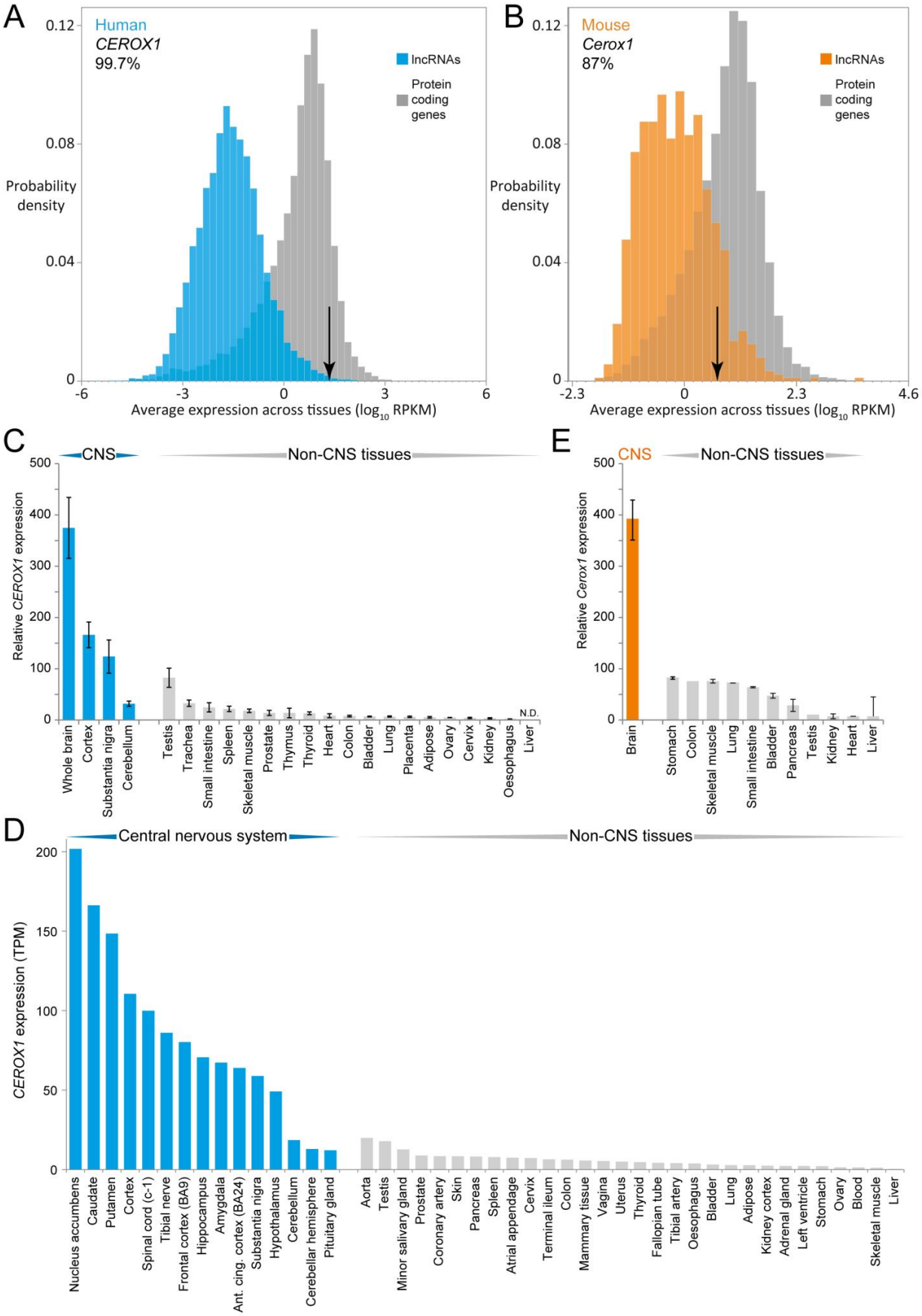


Figure 3.9: Expression patterns of human *CEROX1* and mouse *Cerox1* are highly similar. Expression, averaged across all tissues, of human *CEROX1* **(A)** and mouse *Cerox1* **(B)** within their respective species' transcriptomes indicate that both are highly expressed lncRNAs (data analysis by Dr Wilfried Haerty). Human *CEROX1* is in the top 0.3% of lncRNAs, and mouse *Cerox1* the top 13%. RT-qPCR data **(C)** and RNA-seq data **(D)** from the GTEx project[78] indicate that human *CEROX1* is ubiquitous but highly enriched in the CNS. **(E)** This parallels expression of mouse *Cerox1* (Sirey et al., 2019[20]).

As would be expected, the relative rank of *CEROX1* within the transcriptome increases with expression levels, peaking in tissues of the CNS. Here, *CEROX1* expression outranks that of almost all nuclear-encoded complex I genes, which are essential and highly and constitutively expressed. In the basal ganglia, *CEROX1* is within the top 1% of expressed genes, top 0.2% of expressed lncRNAs, and more highly expressed than ~40 of 44 nuclear-encoded complex I subunits and assembly factors (Figure 3.10); analyses of other tissues are shown in Appendix A6.

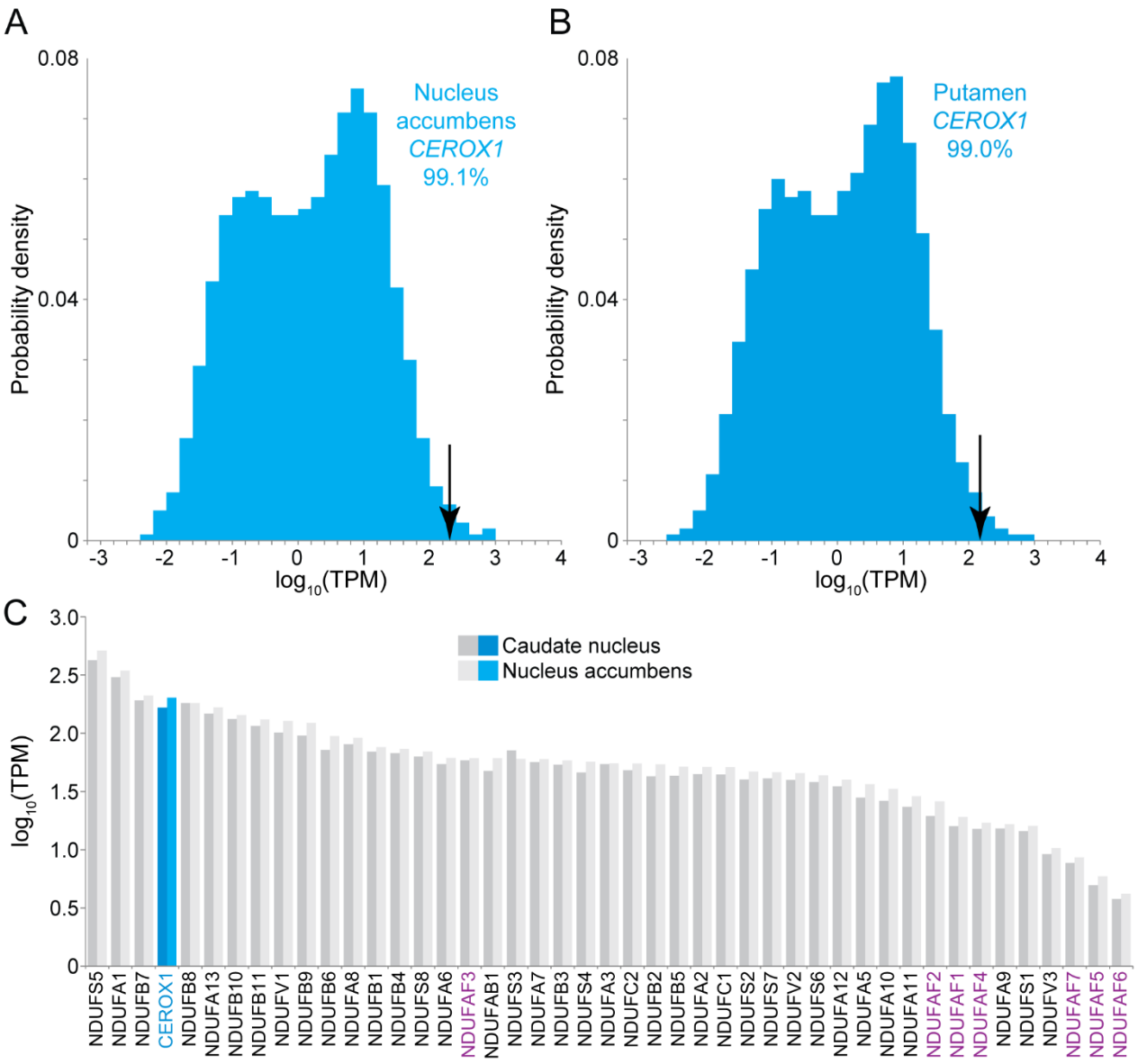


Figure 3.10: *CEROX1* expression is most elevated in the basal ganglia, where it outranks that of mitochondrial genes. In the nucleus accumbens **(A)** and putamen **(B)**, *CEROX1* expression is higher than 99% of the expressed transcriptome (both mRNAs and non-coding RNAs). **(C)** This extraordinary expression is highlighted by the fact that *CEROX1* is more highly expressed than 41 (nucleus accumbens) and 40 (caudate) of 44 nuclear-encoded complex I-associated genes (37 subunits and 7 assembly factors, labels shown in purple). Data were obtained from the GTEx Project[78] (gtexportal.org/home/datasets).

Consistent with the aforementioned hypothesis based upon the long length of the *CEROX1-SOX8* CpG island and the tissue-specific presence of $H3K4Me_2$ throughout the gene body of *CEROX1* (Figure 3.2), the expression of *CEROX1* does indeed show an intermediate level of tissue specificity: it is expressed in almost all tissues, but shows significant enrichment in the central nervous system.

3.4 The expression of *CEROX1* is varied, dynamic, and conserved, to the cellular level

The above expression data are useful in envisaging the distribution of *CEROX1* expression throughout the body, and are highly consistent in identifying the CNS as the main site of expression, supporting physiological conservation of *CEROX1*. Due especially to the complexity and heterogeneity of the cells comprising this tissue, it is essential to determine the expression of *CEROX1* on a cell type-specific basis. Different cell types are responsive to different stimuli, and have their own roles at a sub-organismal level, and so knowledge of cell type-level *CEROX1* expression patterns may inform other hypotheses and experiments, including those investigating the function and regulation of the lncRNA. From a translational perspective, the disease relevance of expression of *CEROX1* may well be dependent on its placement within specific cell types. This prompted two routes of investigation: first, to look at specificity of expression within the major gross cell types of the brain; second, to examine whether *CEROX1* expression is varied between spatiotemporally distinct populations of cells within the same tissue, including throughout development.

3.4.1 *CEROX1* is enriched in cells of the oligodendrocyte lineage, in both mouse and human

Several studies have sought to assess the differential transcriptomes of cell types across the brain. Data from the Qian Wu lab[84], who isolated cell types from adult mouse cortex by immunopanning, indicate that mouse *Cerox1* is highly enriched in developing and mature oligodendrocytes (Figure 3.11A), expressed at 6- to 15-fold higher levels in these cells than the whole cortex on average. This is in agreement with data obtained from two single cell RNA-seq projects that assessed gene expression in two specific brain sub-regions: the SVZ Cell Atlas, which examined the sub-ventricular zone (SVZ) of the adult cortex[85]; and a study of cell populations within the striatum[86] (Figure 3.11B-C).

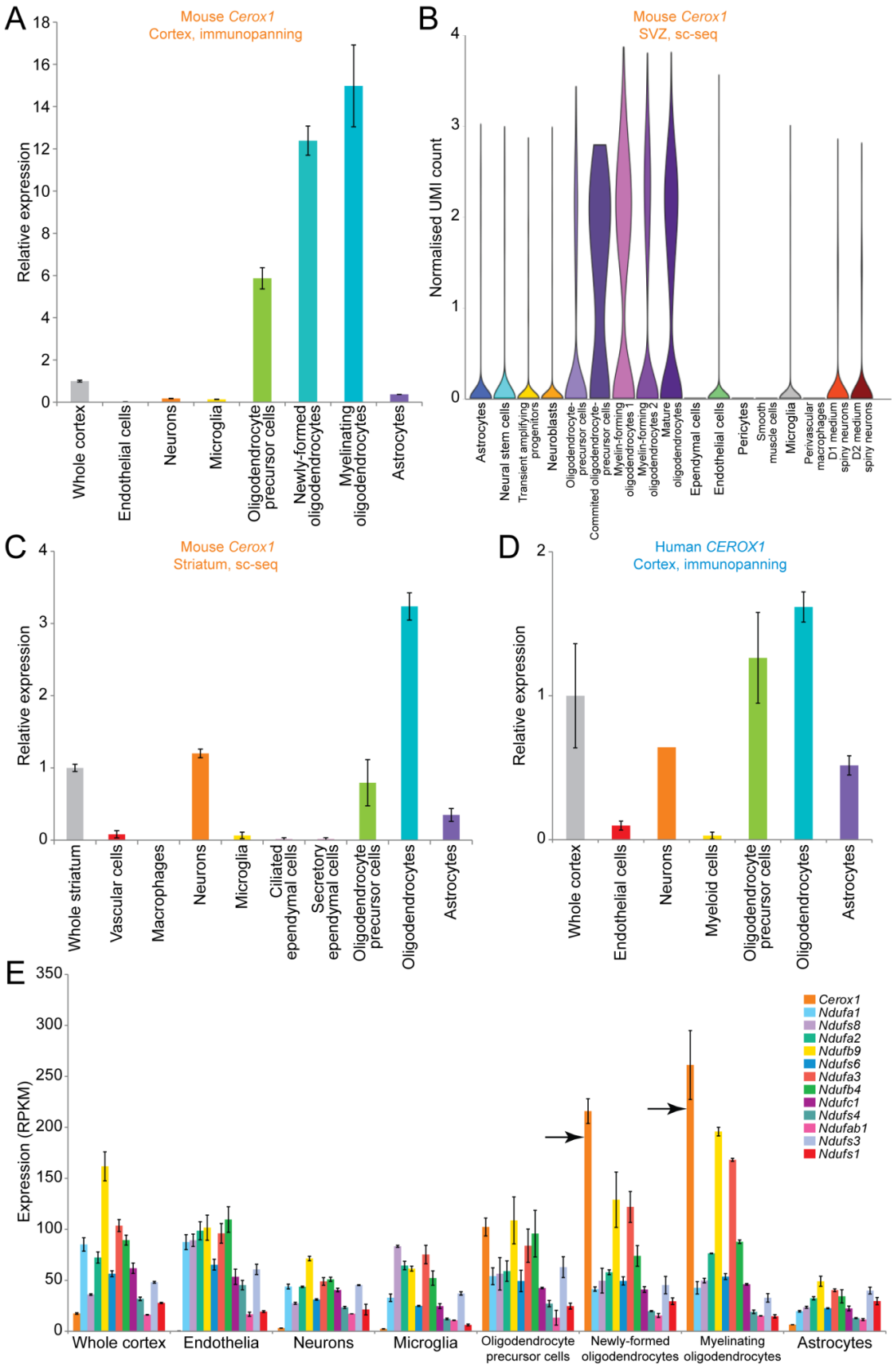


Figure 3.11: Both human *CEROX1* and mouse *Cerox1* are enriched in oligodendrocytes. (A) RNA-seq data[84] from immune-purified cortex cell populations indicate that the expression of mouse *Cerox1* is highly elevated in oligodendrocytes. Single cell sequencing (sc-seq) of **(B)** cortical SVZ populations (shiny.mdc-berlin.de/SVZapp)[85]; and **(C)** striatal cells both corroborate the previous trend of enrichment in oligodendrocytes. **(D)** *CEROX1* also peaks in oligodendrocytes and cells of their precursory lineage. **(E)** Not only is *Cerox1* enriched in oligodendrocytes, but it is also expressed at unusually high levels in these cells (arrows), above the expression of the complex I subunit genes whose levels are altered the greatest extent by *Cerox1* modulation in N2A cells. sc-seq – single cell sequencing.

Human *CEROX1* expression, also derived from temporal cortex RNA-seq data from the Qian Wu group[87], also peaks in cells of the oligodendrocyte lineage, through the enrichment of expression in these cells is to a lower degree (1.3- to 1.6-fold) (Figure 3.11D). In both human and mouse, expression of *CEROX1* is lower in endothelial cells, neurons, myeloid cells or microglia, and astrocytes – all other major cell types found in the brain – than across the whole cortex as an average. In newly-formed and myelinating oligodendrocytes, levels of mouse *Cerox1* RNA are higher than those of all twelve subunits of complex I identified by Sirey et al.[20] whose expression levels are significantly altered by *Cerox1* modulation in N2A cells (Figure 3.11E).

3.4.1.1 The *SOX8-CEROX1* locus is epigenetically regulated in oligodendrocytes

Like *Cerox1*, *Sox8* is also highly enriched in oligodendrocytes, reflecting its role in their development, and suggesting that: a) the two genes are co-regulated, either transcriptionally though their common promoter region or post-transcriptionally; or b) that one of the two genes is regulated in a co- or post-transcriptional manner by the other. The *CEROX1-SOX8* promoter will be the subject of deeper investigation in Chapter 6. Preliminarily, however, ChIP-seq data were examined in order to identify possible epigenetic regulatory co-ordination of *Cerox1* and *Sox8*.

Olig2 (oligodendrocyte transcription factor 2) is a basic helix-loop-helix transcription factor involved in specification of oligodendrocytes and motor neurons in the developing CNS. Data from the literature indicate peaks of Olig2 occupancy at the locus surrounding *Cerox1* and its neighbours in both mouse[88] and rat[89] oligodendrocytes (Figure 3.12). Peaks include especially the putative enhancer element downstream of the 3' end of *Cerox1*, as identified above, in both mouse and rat, and the *Cerox1-Sox8* promoter in mouse. These conserved sites of Olig2 enrichment suggest that the similarity in gene expression between *Cerox1* and *Sox8* is potentially due to co-transcriptional regulation at their shared locus.

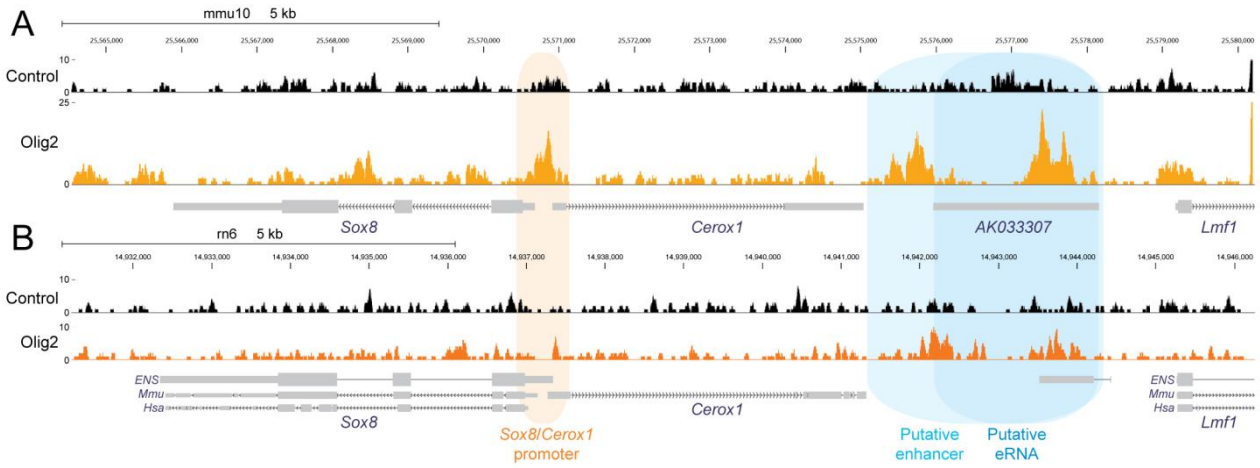


Figure 3.12: Occupancy of the oligodendrocyte transcription factor Olig2 around the rodent *Cerox1* loci. In both mouse **(A)** and rat **(B)**, ChIP-seq data[88,89] reveal peaks of Olig2 occupancy at the promoter of *Cerox1* (orange shading), and especially at the putative enhancer element downstream of the 3' end (light blue shading), as well as across the potential eRNA associated with it (dark blue shading). Ensembl (ENS) transcripts for rat *Sox8* and *Lmf1* are shown alongside alignments of human (Hsa) and mouse (Mmu) orthologues.

3.4.2 Even in the mostly highly expressing tissues, *CEROX1* expression is varied

Human *CEROX1* shows a lower degree of cell type-based enrichment compared to its murine homologue according to the immunopanning RNA-seq[84,87] (Figure 3.11A,D). However, the complex differences between cell types, or even within cells of the same broad 'type' may not be accurately depicted by a single immunopanning experiment. In order to complement these sequencing data, RNA *in situ* hybridisation (ISH) was used to examine the distribution of *CEROX1* across the striatum, a sub-region of the brain in which average human *CEROX1* expression is consistently high (Figure 3.9D – nucleus accumbens, caudate, and putamen), and in the oligodendrocytes of which mouse *Cerox1* shows a lower level of enrichment than those in cortex (Figure 3.11C).

3.4.2.1 *CEROX1* and *SOX8* might be associated reciprocally with the striosome-matrix network

Staining by ISH of human striatal sections (obtained from the Oxford Brain Bank, Dr Olaf Ansorge) indicated expression of both *CEROX1* and *SOX8* across the tissue (Figure 3.13). Neither gene had an entirely homogeneous expression profile, however. *CEROX1* staining was concentrated in irregular patches (Figure 3.13B-B'') while *SOX8* staining showed a partly opposing pattern, with ubiquitous expression except for small patches of little to no staining (Figure 3.13C-C'). The staining for *SOX8* is corroborated by publically available IHC data from the GENSAT project (Gene Expression Nervous System Atlas; gensat.org)[90], which demonstrate similar patches of reduced staining of Sox8 protein throughout the mouse striatum (Figure 3.13D).

These patterns are consistent with an infrastructure of the striatum known as striosomes, or striatal patches. These regions are distinguished from the remainder of the striatum, known as the matrix, by their differential expression of neurotransmitters and receptors. Indeed, the initial identification of striosomes by Ann Graybiel was upon the basis of their reduced acetylcholinesterase (ACHE) expression[91]. Other histochemical markers for striosomes include enrichment of μ-opioid receptors (OPRM1), dopamine receptor D1 (D1R), and substance P, while these markers are reduced in the matrix, where others, including calbindin (CALB1), choline acetyltransferase (CHAT) and the dopamine receptor D2 (D2R), are elevated[92]. Striosomes form a complex and dynamic network of compartments throughout the striatum, occupying ~10-15% of the tissue by volume on average, but this proportion, as well as the typical cross-section of a striosome, varies along the anteroposterior and mediolateral axes[92]. Differential striosome-matrix histochemistry is also altered according to these axes[93-95]. These factors make absolute identification based upon a small number of sections challenging.

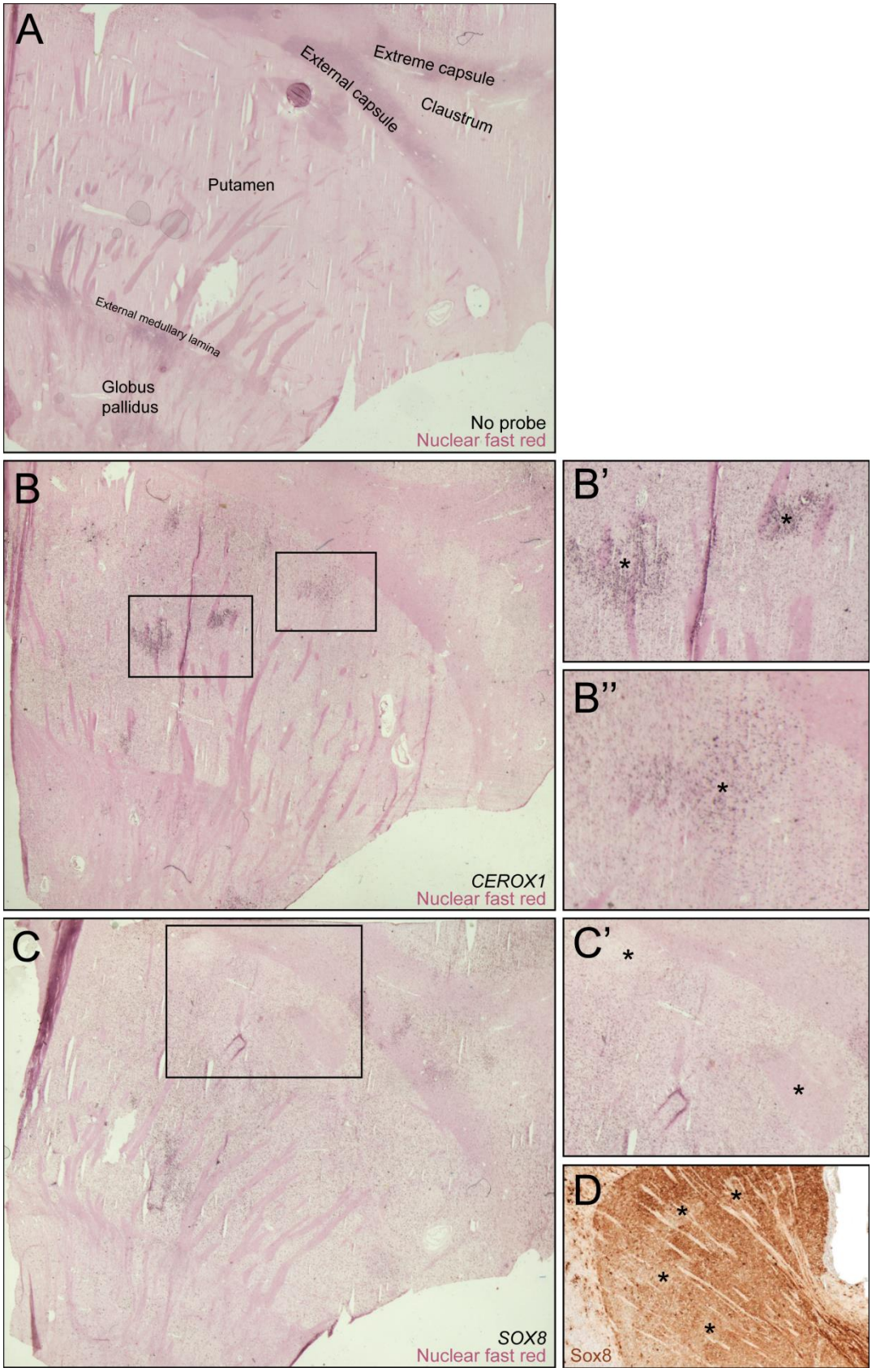


Figure 3.13: *In situ* hybridisation (ISH) indicates heterogeneity in *CEROX1* and *SOX8* expression throughout the striatum. Sections of striatum were subjected to ISH using DIG-labelled RNA probes with chromogenic alkaline phosphatase development (black/dark), and then counter-stained with nuclear fast red (pink/purple). **(A)** Staining in the absence of a DIG-labelled probe produced no staining. **(B)** Following incubation with a *CEROX1* probe staining was observed across the sample, but strongly in irregular patches **(B',B'')**. Likewise, probing for *SOX8* largely revealed staining throughout the sample, but with irregular patches of little to no staining **(C')**. **(D)** Mouse Sox8 IHC from GENSAT (gensat.org)[90] demonstrated similar patches of low staining in the striatum of a sagittal mouse brain section.

3.4.3 *CEROX1* expression is dynamic throughout development, but tissue specificity is retained

As well as providing information about the expression profile of murine *Cerox1* on a tissue-by-tissue basis, RNA-seq data from Schmitt et al.[96] also indicate that mouse *Cerox1* is expressed dynamically throughout late embryonic and early post-natal development, increasing in the brain by 1.3-fold within the final week of gestation, and then by 11.4-fold in the first month of life (Figure 3.14). Expression of *Sox8* in the brain follows a similar trend, increasing by 1.4-fold and 2.2-fold in the same respective periods. In the liver, on the other hand, where adult expression of both *Cerox1* and *Sox8* is minimal or negligible, expression across these developmental stages is consistently very low. *Cerox1* expression in the mouse embryo therefore mimics that in the adult in terms of tissue specificity, and, in the brain, increases throughout development to reach the eventual extremely high level observed in the adult.

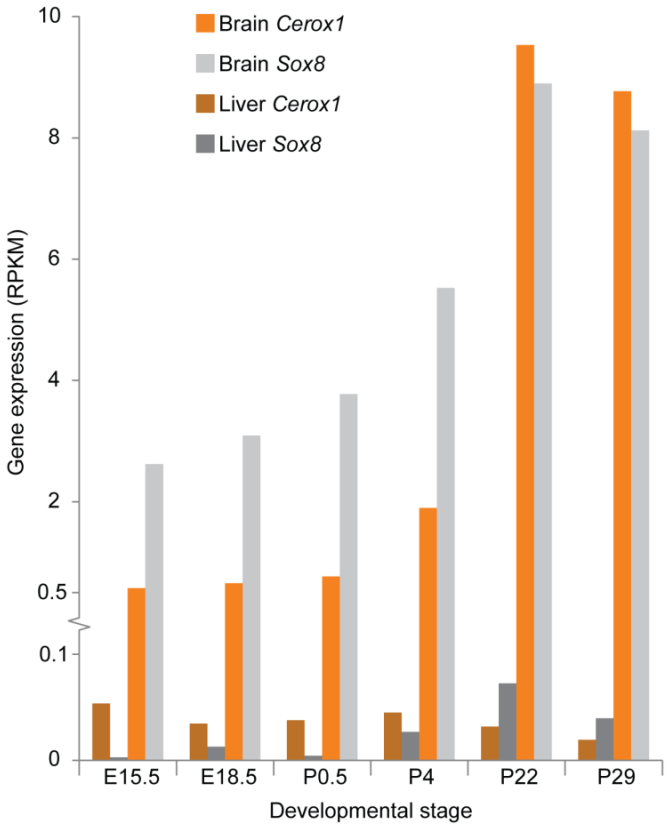


Figure 3.14: Expression of *Cerox1* and *Sox8* in the developing mouse brain and liver are dynamic, and consistent with adult expression data. RNA-seq[96] data indicate the patterns of expression of *Cerox1* and *Sox8* in developing brain and liver between E15.5 and P29. In the brain, expression of both *Cerox1* and *Sox8* increases steadily between each time point, until it peaks ~3 weeks after birth. Meanwhile, in the liver, expression of both genes is extremely low or undetectable, and does not appear to change substantially across the course of these observation points. E – embryonic day; P – post-natal day.

3.4.3.2 Expression of human *CEROX1* during brain development parallels those of *Cerox1* in mouse

In order to extend the investigation of developmental *CEROX1* expression into the human brain, gene expression data from the Human Developmental Biology Resource Expression database[97] (hdbr.org/expression) were used to assess expression in brain regions of human embryos and foetuses from 4 to 17 weeks post-conception. Akin to *Cerox1* in the developing mouse brain (Figure 3.14), *CEROX1* is expressed throughout human brain development, in all regions represented in the data (Figure 3.15). Data for some brain regions are more variable, such as the basal ganglia, and this likely reflects the higher variance of expression in the adult across the sub-regions of the basal ganglia. For some tissues, such as the spinal cord, there is a clear trend of increasing expression throughout the period of gestation represented (Spearman's $\rho = 0.74$, $p = 0.0027$), and this is consistent with the steady increase seen in the levels of *Cerox1* in the developing mouse brain, and with a potential role for the gene in brain development.

3.4.3.3 *In vitro* differentiation recapitulates the increase of *CEROX1* observed in development *in vivo*

Following the observation of dynamic *CEROX1* expression in the developing brain, further gene expression data derived from stem cell models of differentiation may provide an insight into how *CEROX1* might: a) be dynamic in the development of other tissues; and b) be dysregulated in disease.

Data from Nayler et al., 2017[98], accessed via Stemformatics (stemformatics.org)[99], show that *CEROX1* expression is increased 3.7-fold during the differentiation of iPSCs (induced pluripotent stem cells) into cerebellar neurons. In dorsal forebrain-like neurons differentiated from Down's syndrome and control iPSCs, *CEROX1* was upregulated 10- to 11-fold[100]. These data indicate that *in vitro* differentiation of *CEROX1* does recapitulate changes observed by RNA-seq of *in vivo* development in the brain. Looking at other non-CNS cell types, FANTOM CAT data[56] indicate that *CEROX1* is upregulated during the differentiation-associated fusion of myoblasts into myotubes, and during differentiation of hESCs into cardiomyocytes, suggesting that a trend of increasing *CEROX1* expression might be observed in developing skeletal and cardiac muscle.

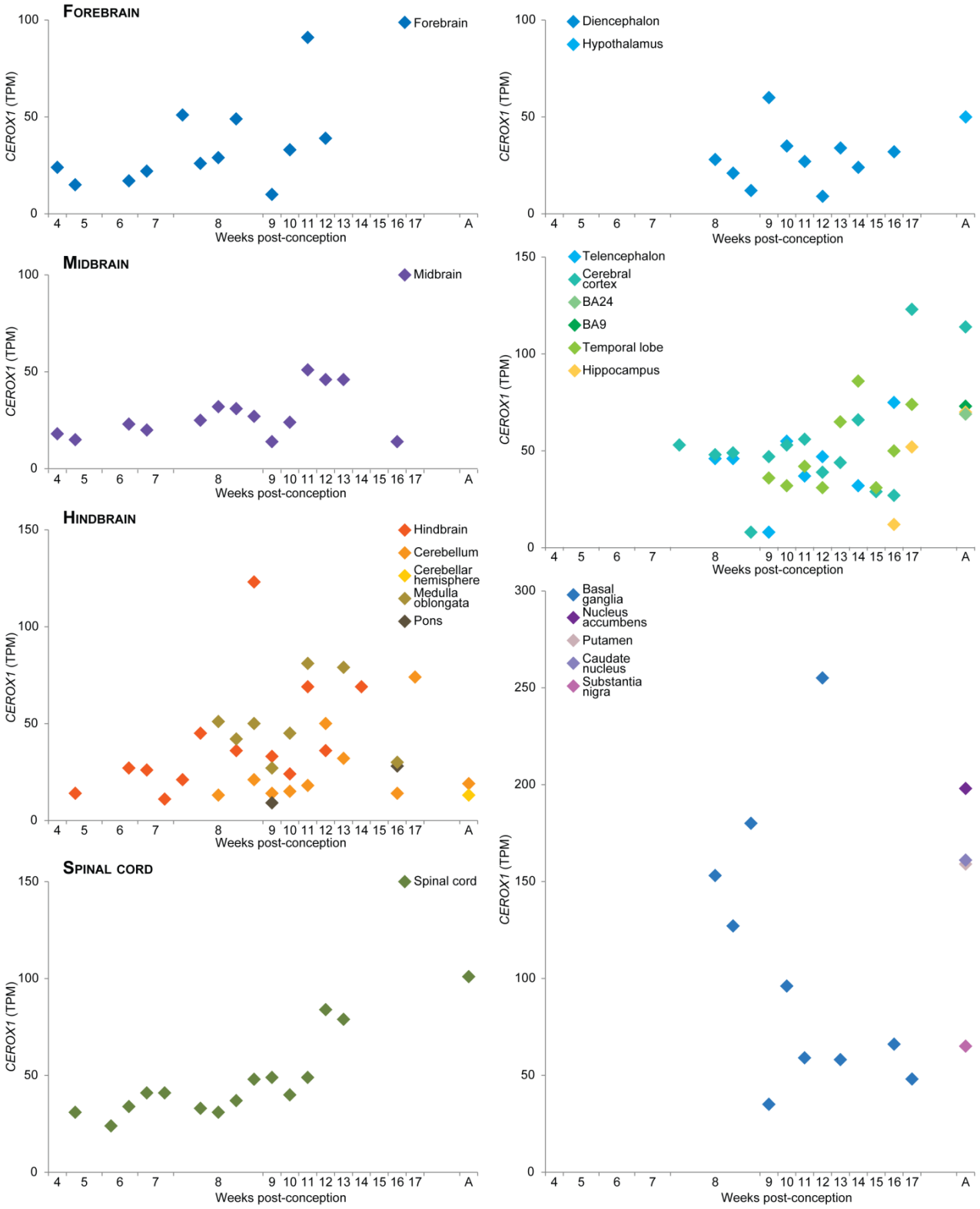


Figure 3.15: Human *CEROX1* is expressed throughout the developing human brain. RNA-seq data from the Human Developmental Biology Resource Expression database[97] (hdbr.org/expression) indicate that *CEROX1* is expressed across all regions of the developing brain throughout the period represented, from four to 17 weeks post-conception. The data are stratified into regions of the developing central nervous system: the forebrain, including the diencephalon and telencephalon, which forms the cortex and basal ganglia (subgroups in the right-hand side graphs); the midbrain; the hindbrain, which forms the cerebellum, medulla oblongata, and pons; and the spinal cord. Where available, adult expression from the GTEx project dataset (A) is included for comparison.

Ataxia-telangiectasia (AT) is a neurodegenerative disease that causes cerebellar ataxia as well as increased predisposition to cancer and infection. Data from the first example above show that the overall expression of *CEROX1* is unaltered in neurons differentiated from AT patient-derived iPSCs (induced pluripotent stem cells) relative to controls. However, there may be a change in the balance of isoforms (Dr Samuel Nayler, personal communication). In dorsal telencephalon-like neural progenitor cells, on the other hand, a 1.5-fold increase in *CEROX1* expression was observed in neural progenitors differentiated from autism spectrum disorder patient-derived iPSCs over those derived from controls[101].

3.5 Conservation of *CEROX1* is older than the divergence between human and mouse

3.5.1 The DNA sequence of the *CEROX1* locus is conserved throughout Eutheria

Examining the conservation tracks of the UCSC Genome Browser (genome.ucsc.edu) indicates that the sequence of the *CEROX1* locus is conserved not only back to the common ancestor with mouse but throughout the mammalian clade of Eutheria (Figure 3.16). It does not appear to be present in Metatheria, suggesting that it arose during the radiation of placental mammals around 100-90 Mya[102-104]. This is in contrast to the highly conserved transcription factor *SOX8*, from which *CEROX1* is divergently expressed, whose sequence (Figure 3.16) and expression are conserved throughout vertebrates[31], all the way back to reptiles[105] and fish[30]. The *LMF1-SOX8* synteny appears to be conserved to zebrafish (with which mammals last shared a common ancestor over 450 Mya[106]) (Figure 3.17).

The most conspicuous region of conservation that does not fall within *SOX8* or *LMF1* is in fact across the putative conserved enhancer element identified previously in this chapter (Figure 3.2; Figure 3.7), which is downstream of the 3' end of *CEROX1*. PhastCons scores[107] (Figure 3.16A,E) are strongly in support of a conserved element, as is basewise conservation by PhyloP across 100 vertebrates (Figure 3.16B). The more evolutionarily distal conservation at this element than across the *CEROX1* gene suggests that the enhancer element is older than the lncRNA, and is, or at least was originally, involved in the regulation of expression of a neighbouring gene such as *SOX8*, or even a distal gene.

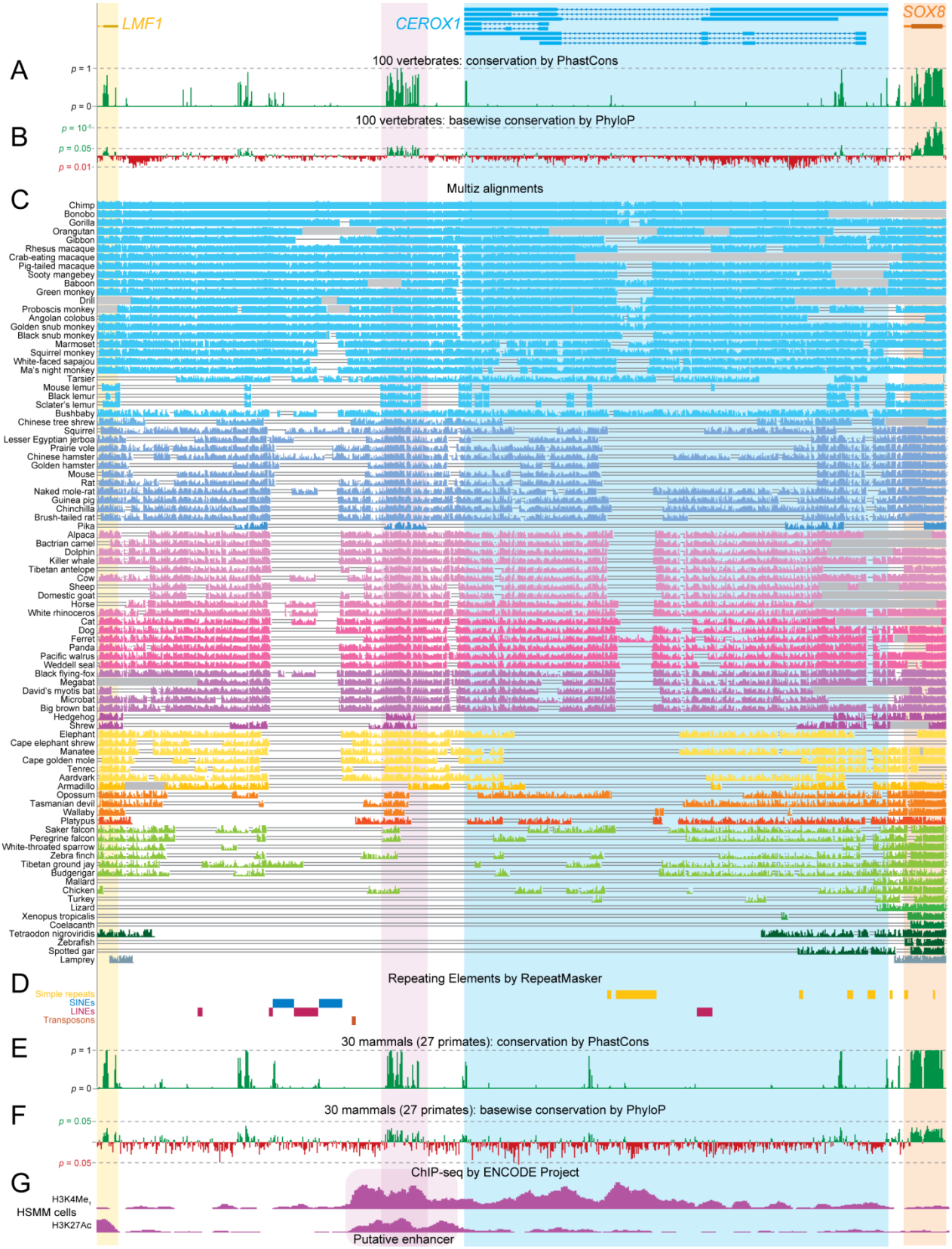
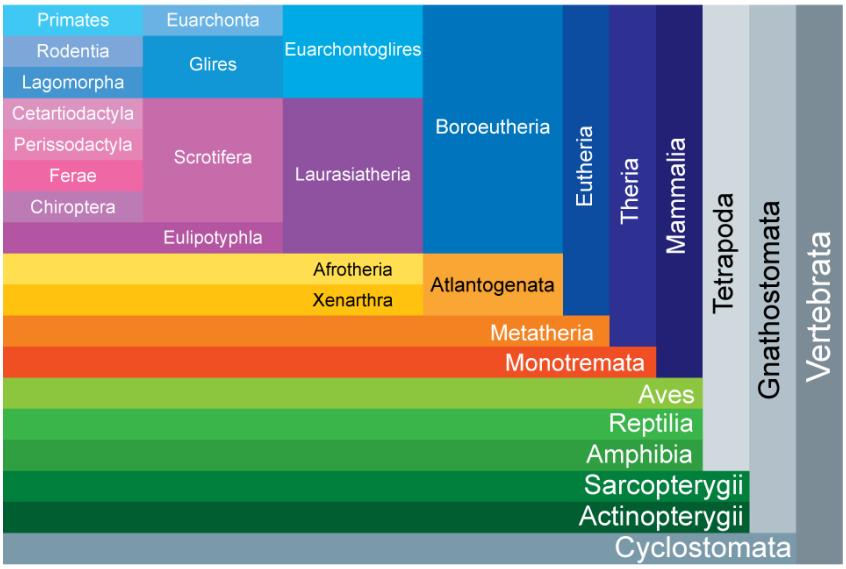


Figure 3.16: Evolutionary conservation of the DNA sequence of, and surrounding, the *CEROX1* genomic locus. Conservation metrics calculated for the *CEROX1* across the aligned sequences of 100 vertebrates indicate that the mixed levels of conservation throughout the gene. **(A)** PhastCons scores[107], which represent the probability that each nucleotide is involved in a conserved element, indicates peaks of conservation at the TSSs and TTSs of *CEROX1*. **(B)** PhyloP, which assesses the probability of rejecting a hypothesis of neutral evolution in favour of one of conservation or accelerated evolution, also indicates a significant likelihood (indicated by dashed lines) of conservation at the transcriptional start sites. Individual species' Multiz alignments **(C)** illustrate that the gene sequence of *CEROX1* is highly conserved throughout eutherian (placental) mammals, and appears to be conserved into the few examples of metatherians (e.g. opossum), across small stretches of sequence. Closer examination of apparent peaks of conservation in platypus and

birds reveals that these are incidental artefacts of alignment. Regions most lacking in conservation correspond with repetitive elements, as annotated by RepeatMasker **(D)**. Conservation metrics for the sequences of only 30 mammals (27 primates) agree with a hypothesis of recent neutral evolution at the *CEROX1* locus **(E, F)**, with sequences showing non-significant scores in both directions. **(G)** The putative enhancer identified above aligns with a region of high PhastCons and PhyloP scores, supporting its classification as a conserved regulatory element. For comparison, the first exon of *SOX8* (orange) shows a high density of high PhastCons conservation scores **(A)**, and a very high probability of featuring in a conserved element, i.e. the coding sequence of this protein-coding gene, with some nucleotides having a conservation probability of $p < 10^{-5}$ **(B)**. **(C)** Individual species alignments illustrate conservation throughout all vertebrate representatives, back to the jawless fish, the lamprey. The data support that *LMF1* (yellow) has been conserved throughout mammals and birds. ChIP-seq data were generated by the Bernstein lab of the ENCODE Project Consortium. All conservation metrics and alignments were obtained from the UCSC Genome Browser (genome.ucsc.edu). Below is shown a key to the colours used for alignment bars, which reflect vertebrate phylogenetic groupings.



Rodentia - rodents; Cetartiodactyla - even-toed ungulates; Perissodactyla - odd-toed ungulates; Ferae - carnivores; Chiroptera - bats; Afrotheria - African mammals; Eutheria - placental mammals; Metatheria - marsupials; Monotremata - egg-laying mammals; Tetrapoda - four-limbed vertebrates; Sarcopterygii - lobe-finned fishes; Actinopterygii - ray-finned fishes; Gnathostomata - jawed vertebrates; Cyclostomata - jawless vertebrates; Vertebrata - backboned animals.

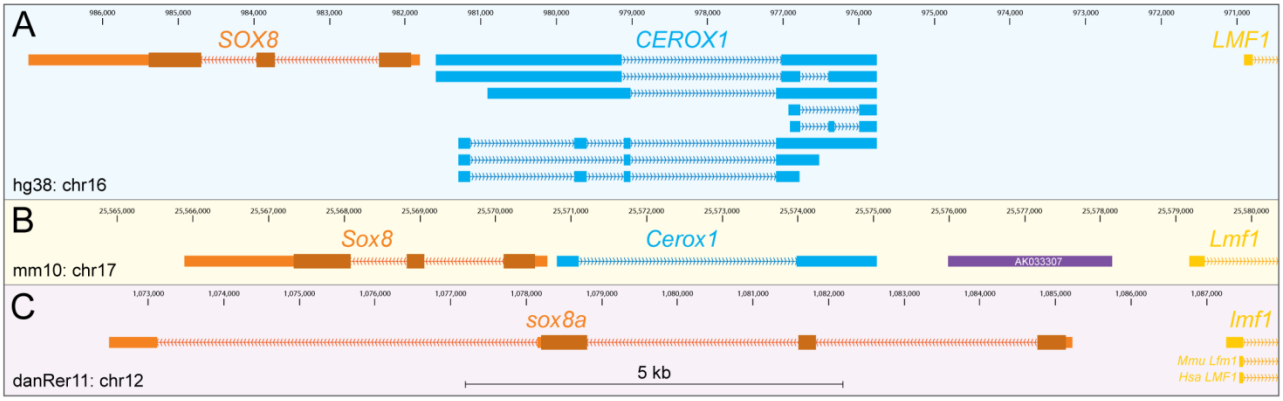


Figure 3.17: The synteny of *CEROX1*'s neighbouring protein-coding genes is highly conserved. The close proximity of *SOX8* and *LMF1* orthologues is conserved not only throughout mammals including human **(A)** and mouse **(B)**, but also back to zebrafish **(C)**, among the most evolutionarily distant vertebrates.

3.5.2 The conserved *CEROX1* locus is transcribed throughout Eutheria

3.5.2.1 *CEROX1* is expressed throughout Euarchontoglires

While the sequence of the *CEROX1* locus is conserved throughout Eutheria, this does not necessarily indicate that the lncRNA is expressed in all of the species constituting this super-order. For example, three lncRNAs are annotated in the proximity of the *SOX21* locus in human but, even though the DNA sequence of the region is conserved throughout mammals, and some stretches throughout vertebrates, only one of the three lncRNAs has an expressed orthologue in mouse[7]. In order to assess the conservation of expression amongst other species, and rule out the, albeit unlikely, scenario that human and mouse (*Mus musculus*) are exceptional in expressing a *CEROX1* orthologue transcript, RACE was used to assess *Cerox1* expression in other mammals. Specifically, 3' RACE was utilised due to its ability to distinguish transcribed RNA based upon the polyadenylated tail. Initially, three additional species that are evolutionarily close to human or mouse were investigated, namely the rhesus macaque (*Macaca mulatta*) and western lowland gorilla (*Gorilla gorilla gorilla*), and the common rat (*Rattus norvegicus*), respectively (Figure 3.18A). Total brain RNA samples from the five mammals were probed by 3' RACE, and *Cerox1* transcript termini were successfully derived from each (Figure 3.18B). The terminal 90 nucleotides from the 3' ends of the identified transcripts share 52% sequence identity, supporting the previous identification of a putative conserved TTS in PhastCons data (Figure 3.16).

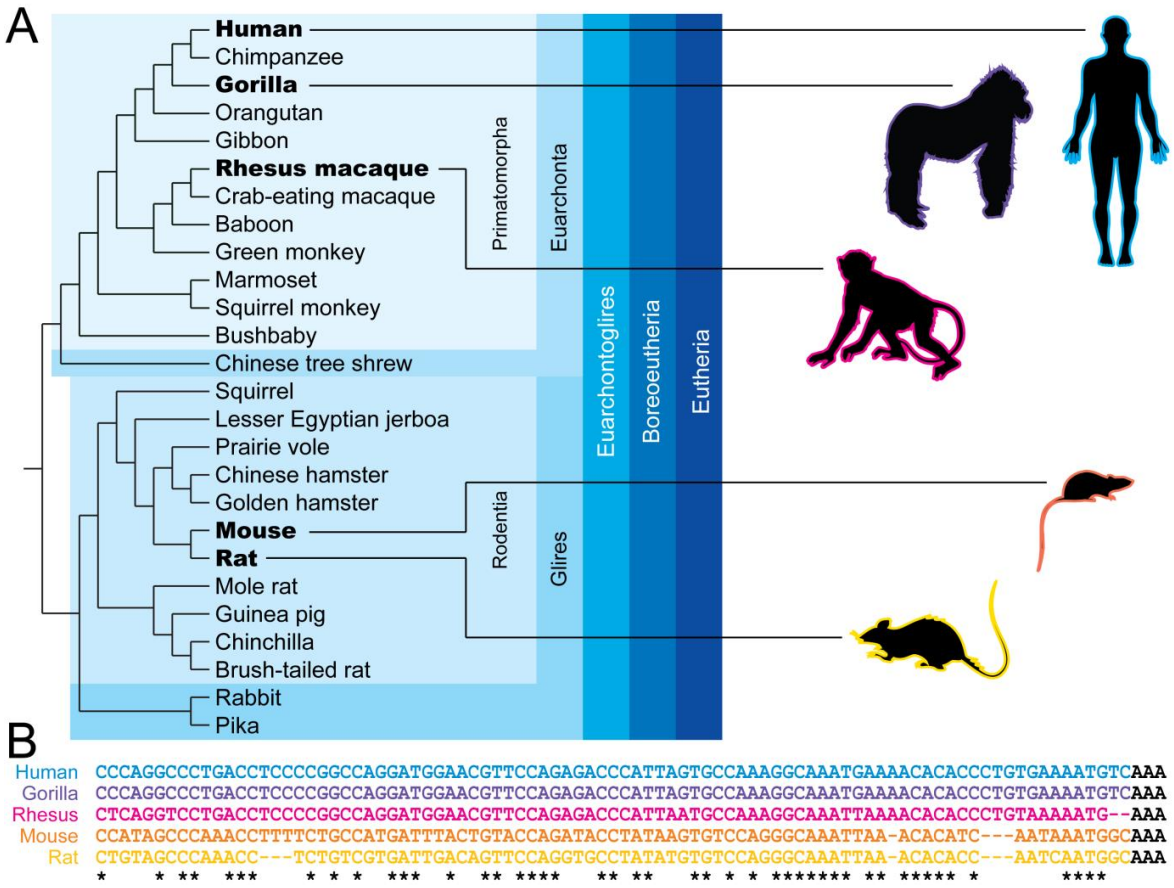


Figure 3.18: Conservation of the *CEROX1* 3' end in Euarchontoglires. 3' RACE was used to characterise the presence of *CEROX1* orthologues across representative Euarchontoglires species (**A**): human, gorilla, rhesus macaque, mouse, and rat. (**B**) The 3' terminal 90 nucleotides of the orthologous *Cerox1* transcripts in these five Euarchontoglires species share 52% sequence identity.

3.5.2.2 The search for *CEROX1* orthologues in Laurasiatheria primarily utilised bioinformatic data

The previous data support conservation of expression of *CEROX1* orthologues throughout Euarchontoglires. To investigate more evolutionarily divergent eutherian mammals, species from the Laurasiatheria clade were also examined. While sequencing 3' and 5' RACE-PCR amplicons is the surest way to infer the termini of transcribed sequences *in vitro*, bioinformatic datasets can both inform the design of such experiments, and complement them with samples that would not be readily available for study. The rise of RNA-seq and the resulting explosion of transcriptomics data have provided expansive datasets for a number of diverse mammalian species, including Laurasiatheria representatives. Even more divergently with respect to human evolution, some are derived from species of the remaining eutherian super-orders Afrotheria and Xenarthra, which are relatively few and not routinely subject to research. Given the highly conserved synteny of the genomic region throughout Mammalia, the search for *CEROX1* orthologous transcripts in

Laurasiatheria species was focused in the intergenic space between the 5' ends of *SOX8* and *LMF1* orthologues.

Two approaches were taken in analysing RNA-seq data: examination of existing annotations where available; and manual mapping where prior annotation is not present. In response to the difficulty of annotating lncRNAs in poorly-studied and in some cases poorly-annotated genomes, the Ulitsky group developed a 'pipeline for lncRNA annotation from RNA-seq data' (PLAR), and subsequently applied this to transcriptomics data from 18 species, from human to sea urchin[7,108]. The resulting read mappings and annotations, available through the UCSC Genome Browser (genome.ucsc.edu) were consulted in order to identify *CEROX1* orthologues, and carry the weighty advantage of strand-specific mapping. In dog, ferret, and sheep, PLAR analyses identified transcription at the *CEROX1*-syntenic loci. Expression in sheep was identified in almost all tissues, and substantially enriched in the brain (Figure 3.19A). This is consistent with the expected distribution for a *CEROX1* orthologue. Likewise, the expression mapped in dog was almost entirely observed in brain and testis samples (Figure 3.19B), and the resulting annotation is consistent with a putative lncRNA annotation resulting from an improved canine transcriptome annotation by the Broad Institute and Uppsala University[109] (UCSC Track Hub). Interestingly, the latter project also annotated another putative non-coding transcript in the *Sox8-Lmf1* syntenic intergenic locus, divergent from *Lmf1*, and this transcript overlaps the conserved enhancer putatively identified above based upon epigenetic data in human and mouse, suggesting that it might be associated with enhancer-promoted transcription.

The recent addition of rhesus macaque, rat, and dog, among others, to FANTOM5[110] provides CAGE-seq for these species, but no peaks are annotated for *Sox8* or *Cerox1* orthologues, likely due to low coverage.

3' RACE of the putative *Cerox1* transcript in ferret was attempted using RNA extracted from an adult brain, but the poor quality of the resulting RNA precluded success.

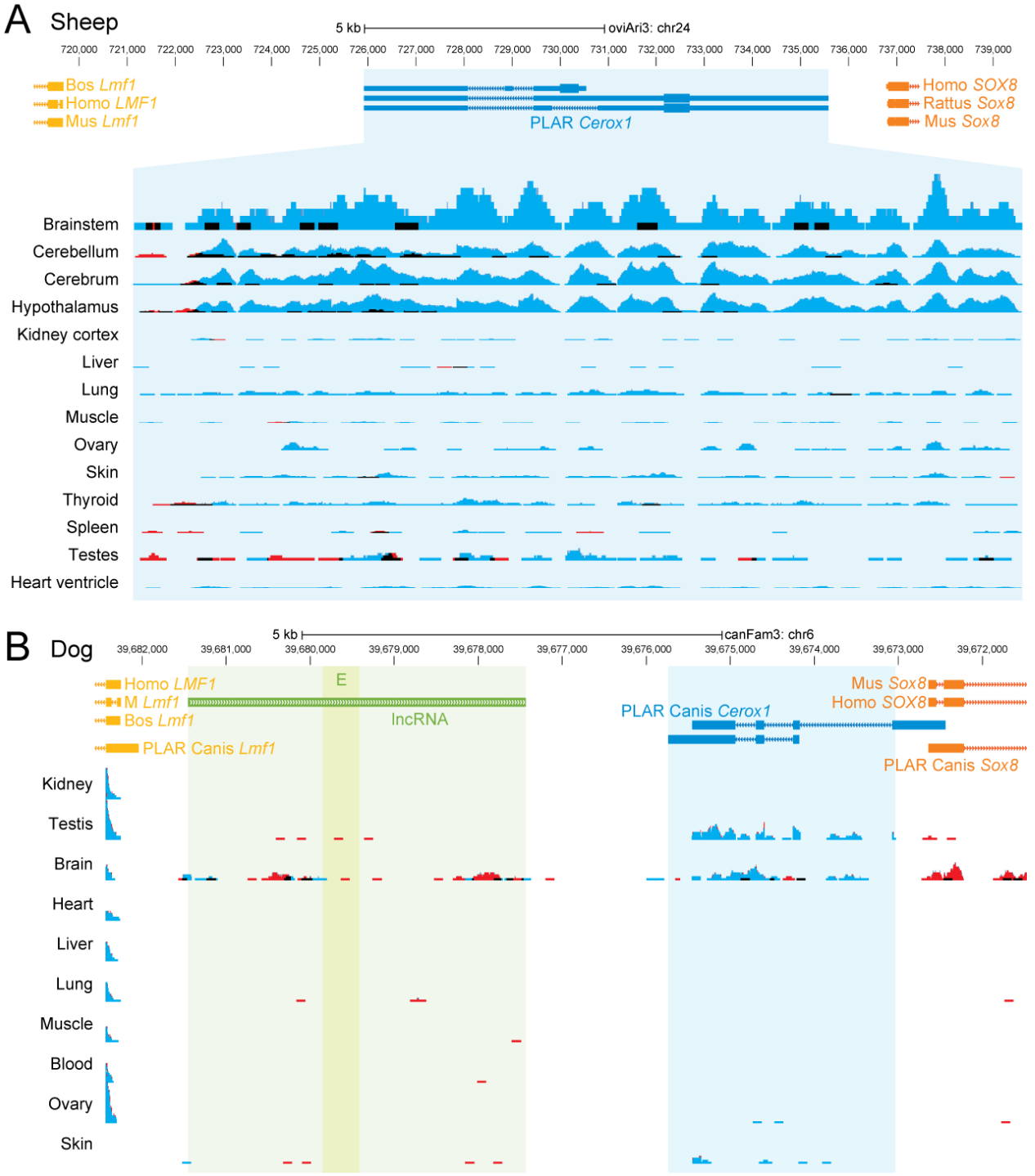


Figure 3.19: PLAR analyses identified *CEROX1*-orthologous transcripts in sheep (*Ovis aries*) and dog (*Canis lupus familiaris*). RNA-seq data analysed by PLAR[7,108], and accessed via the UCSC Genome Browser (genome.ucsc.edu) led to the annotation of putative transcripts at the *Cerox1* syntenic locus in **(A)** sheep and **(B)** dog. **(A)** In sheep, RNA-seq traces indicate a similar tissue localisation profile to that expected, with enrichment in the central nervous system. **(B)** Lower expression in dog has led to likely less precise annotation of *Cerox1*, with some disparity between peaks and the annotation, but there is clear evidence of transcription. A second transcript is annotated between *Lmf1* and *Cerox1* (green). Interestingly, this overlaps the putative conserved enhancer identified previously (E). Blue peaks indicate transcription on the strand consistent with *Cerox1* synteny, orange on the opposite strand, and black non-stranded sequencing reads. Mapped homology of *Lmf1* and *Sox8* loci to human, mouse, and cow genomes is shown to evidence synteny.

3.5.2.3 Further annotation and phylogenetic analysis supports the mammalian origins of *CEROX1*

In order to extend the investigation to some of the many species upon which PLAR has not yet been applied, the intergenic sequences between *SOX8* and *LMF1* orthologues were derived from the UCSC (genome.ucsc.edu) and Ensembl (ensembl.org) genome browsers and then used to search, using BLAST (blast.ncbi.nlm.nih.gov), against the read sequences identified in the various transcriptomic datasets, identified using the NCBI Short Read Archive (SRA; ncbi.nlm.nih.gov/sra). Informed by the conserved expression pattern of *CEROX1* in mouse and human, which both exhibit substantial central nervous system enrichment (Figure 3.9), the search for Laurasiatheria representative transcripts was initially focused on data derived from brain tissues where available. However, it is possible that these more evolutionarily diverse species would show a different distribution of the transcript (this will be investigated later) and so other tissues in which *CEROX1* is known to be expressed in human were considered where central nervous system data were not available. A number of Euarchontoglires species including those represented within PLAR analyses were included for comparison to previously obtained data. Examined species loci and transcriptomics datasets used are listed in Appendix A4. Successfully mapped reads were used as guides for defining putative *CEROX1* orthologous transcripts in the respective species. In order to assess any clade-specific evolutionary changes that have occurred locus during its evolution, putative *CEROX1*-orthologous loci defined by transcriptomics data were then aligned using MUSCLE (EMBL-EBI; ebi.ac.uk/tools/msa/muscle)[111-114], and the resulting alignment was used to produce a phylogenetic tree in MEGA7[115] (Figure 3.20). The tree recapitulates the consensus phylogeny of mammals: primate and rodent groups each cluster, and then join together to form the Euarchontoglires clade; the carnivores (Ferae) cluster and then join with the odd-toed ungulates (Perissodactyla) and even-toed ungulates (Cetartiodactyla) to form Laurasiatheria; meanwhile, armadillo, the single Xenarthra representative, is the furthest out-group. This supports the relatively neutral evolution of the *CEROX1* locus in accordance with the consensus evolution relationships of eutherian mammals. No Afrotherian *CEROX1* orthologue was identified, likely due simply to the lack of data: the clade includes few species and the only transcriptome available was that of an elephant placenta[116]. The rodent clade notably has long branches, indicating a greater rate of sequence divergence in mouse and rat especially. This is likely due to the fact that rodent genomes are subject to a higher evolutionary rate than those of primates[117].

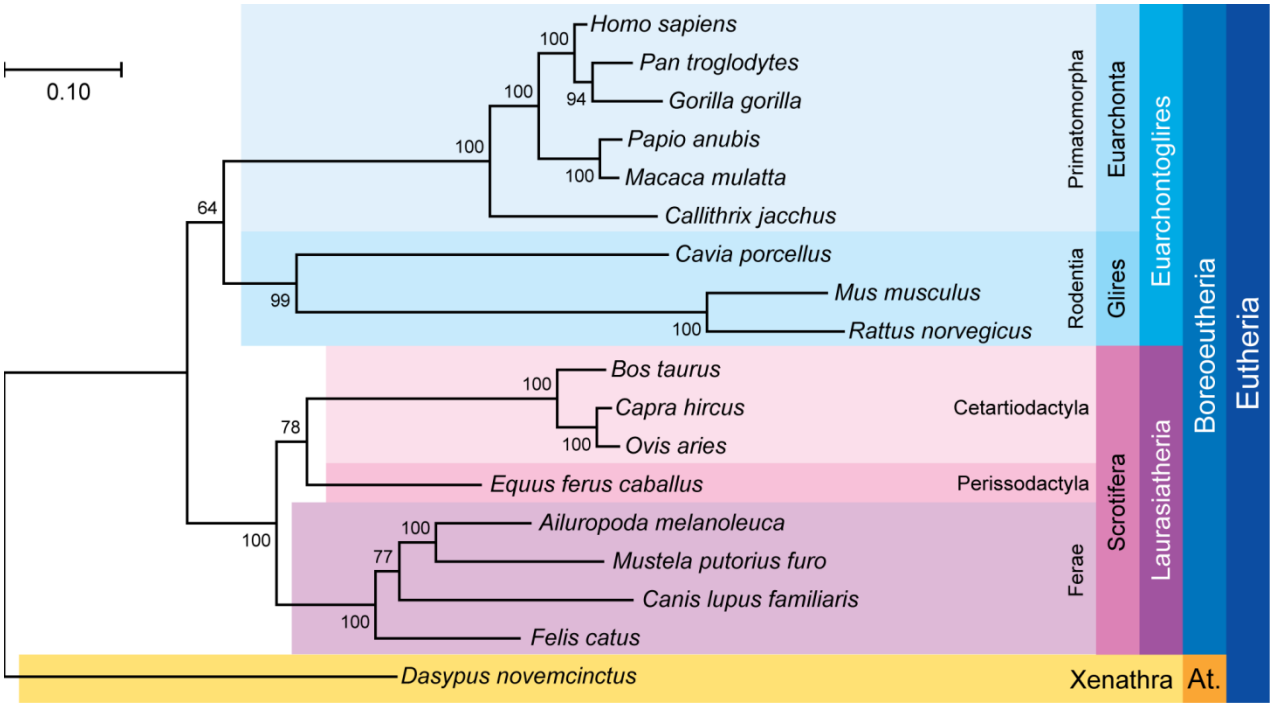


Figure 3.20: Phylogenetic relationships of *CEROX1*-orthologous sequences in 18 eutherian mammals. The phylogenetic tree indicates the evolutionary divergence of sequence in representative mammalian species. The nine-banded armadillo (*Dasypus novemcinctus*) which is the sole Atlantogenata (At.) species represented, acts as an outgroup, while the remaining mammals cluster in accordance with the consensus phylogenetic groups, listed at the right-hand side. The tree was generated in MEGA7[115] using a maximum likelihood nearest-neighbour-interchange method, assuming a uniform general time reversible model of substitution, analysing all sites with at least 75% nucleotide coverage. The key at top-left shows the branch-length, a metric of sequence dissimilarity equivalent to substitutions per site. Meanwhile, the values at internal nodes represent bootstrap statistic[118] support values for each clade, derived from 1000 iterations.

Interestingly, while no overt sequence conservation is present at the *CEROX1* locus past the eutherian mammal divergence, PLAR analysis of some more distal vertebrate genomes does suggest that the locus has previously hosted lncRNA expression. In elephant shark and coelacanth, areas of transcription are present between the orthologues of *SOX8* and *LMF1*. These might represent eRNAs associated with conserved enhancer elements of the surrounding protein-coding genes, or evolutionarily short-lived bona fide lncRNAs.

3.5.3 Expression at the *CEROX1* locus has always been enriched in the central nervous system

RNA-seq data have consistently indicated that human *CEROX1* is ubiquitously expressed but highly enriched in the central nervous system (Figure 3.9). Building upon the above data, which indicate that *CEROX1* is evolutionarily conserved both at the DNA sequence level and in its transcription into an RNA molecule, the subsequent step in characterising the evolutionary expression of the lncRNA was to examine the profile of *CEROX1* orthologues across the tissues of diverse species, thereby probing whether its conservation is likely due to a tissue-specific functionality.

For some of the more commonly studied species, RNA-seq data are available for multiple tissues; these data were used to examine the tissue-level expression distribution of putative *CEROX1* orthologues, as well as validate the assumption used above: that if a sequence transcribed from the syntenic locus is to be identified from transcriptomics data, this will be most readily d 0000one using those from brain tissues. Analysis of eight non-human mammals (rhesus macaque, mouse, rat, cow, dog, cat, giant panda, and nine-banded armadillo) supported this conclusion, with all showing highest expression in the brain and central nervous system (Figure 3.21). The smallest enrichment for central nervous system tissue was observed in the armadillo, where the expression level in the spleen was almost equal (99.9%) to that in cerebellum (Figure 3.21I). This might represent a specific evolutionary difference, since armadillo is the most evolutionarily diverse species represented in this analysis, or simply an anomaly, though it is also notable that the cerebellum is amongst the most lowly expressing brain regions in human, rhesus macaque, mouse, cow, and cat (Figure 3.21A,B,C,E,G) and so it seems likely that other armadillo brain regions might present a greater *Cerox1* enrichment. No multiple-tissue data was available for any other Xenarthra representatives, or any Afrotheria species, in order to probe a possible evolutionary trend.

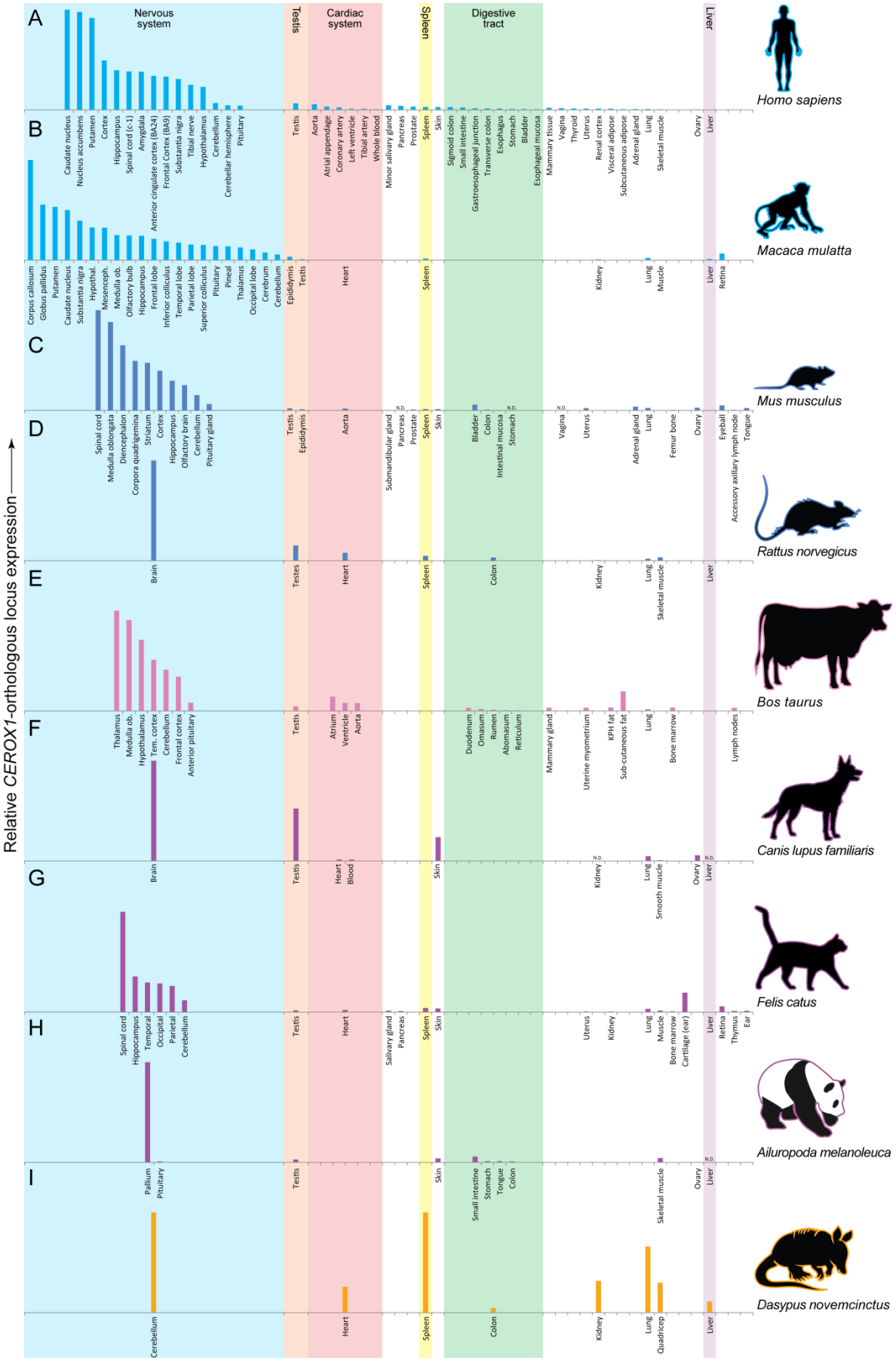


Figure 3.21: Tissue expression profile of *CEROX1* orthologues is conserved throughout Eutheria. Examination of relative transcription levels across the loci of putative *Cerox1* orthologues in tissues of nine mammals indicates that the trend of expression on a tissue basis is largely conserved. In particular, enrichment in the CNS (blue) and absence or minimal expression in the liver (purple). Data points are partially stratified by physiological system and otherwise ordered by descending expression in human tissues.

DISCUSSION

Evolutionarily, *CEROX1* is more highly conserved than over 80% of human lncRNAs[1], which are primate lineage-specific. Considering also that even the ~15 My of evolutionary separation between mouse and rat has yielded conserved transcription at only 60% of rodent lncRNA loci[119], the comparatively extensive conservation of *CEROX1* is highly indicative of functionality. On a broad tissue level, conserved enrichment of *CEROX1* and its orthologues in the central nervous system, where cellular energy burden and ATP production requirements are highest (the brain comprises 20% of the body's energy requirement[120]), is in accordance with the notion that *CEROX1* carries out a conserved role in regulating energy homeostasis. Since mitochondria are responsible for 90% of cellular energy production[121], this is additionally consistent with the hypothesis that human *CEROX1*, like mouse *Cerox1*, acts upon the mitochondrial electron transport chain. Why then is *CEROX1* consistently absent in the liver, given that it is a metabolically active tissue? One important facet of the competitive endogenous RNA (ceRNA) mechanism is the potential for cross-regulation between numerous transcripts[122,123], and so *CEROX1* may be only one of a number of RNA molecules involved in a homeostatic regulatory network that modulates electron transport chain activity. While liver is undoubtedly highly metabolically active, its metabolism is also substantially different to that of other tissues, in that many of its activities function to maintain whole-body homeostasis and the energy substrate needs of other organs such as the brain, rather than simply to produce energy in the form of ATP. It may therefore be the case that the liver uses a different set of transcripts in order to integrate control of the ETC with alternative signalling and metabolic pathways.

The extremely high expression level of *CEROX1* and *Cerox1* within their respective species' transcriptomes addresses one of the potential questions regarding the ability of *CEROX1* to regulate genes of the ETC via a ceRNA mechanism: is *CEROX1* expressed highly enough to compete to a meaningful degree with these transcripts, which are both highly expressed and stable? A common criticism of the ceRNA mechanism is the requirement for appropriate stoichiometry of the interacting RNAs[124,125]. Multiple models have found that the most potent cross-regulation of competing transcripts was achieved when their expression was closer in magnitude and that more effective regulation was of the less abundant transcript following perturbation of the more abundant one[126-128]. The genes of the ETC are highly and constitutively expressed and stable, at both the mRNA and protein level[129]. It would thus be expected that a competitive endogenous

regulator of this system would require a very high abundance in order to effect regulation. Being amid the most highly expressed lncRNAs in the transcriptome, *CEROX1* has the requisitely high expression for this role, especially in the CNS, where expression is enriched above complex I subunit levels in both human and mouse. Mouse *Cerox1* is amongst the top three lncRNAs whose expression is enriched in precursory, newly-formed, and mature myelinating oligodendrocytes[84].

The expression of *CEROX1* is correlated with that of *SOX8*. Given the apparent difference in the evolution of the two genes – *SOX8* is considerably older, based upon its extensive vertebrate conservation – it is possible that the existing promoter architecture of *SOX8* facilitated the evolution of *CEROX1* expression[130]. This will be explored further in Chapter 6, but it is noteworthy that regulatory sites within the genomic locus of the two genes are bound by Olig2 in a conserved manner, suggesting that they are potentially regulated in a concerted manner in oligodendrocytes, where expression of both genes is enriched. Oligodendrocytes are specialised CNS glial cells whose key function is insulation and stabilisation of neuronal axons via production and maintenance of the myelin sheath, a lipid-dense layer. The propagation of action potentials along an axon is accelerated by saltatory conduction as it jumps between the unmyelinated internodes. Oligodendrocytes are the most common type of glial cell (57%) found in the neocortex (29% of the total cell number), and the most common cell type in the basal ganglia, representing almost 50% of the tissue (calculated using data from Karlsen and Pakkenberg[131]). This is consistent with the extraordinary level of expression of *CEROX1* in the basal ganglia.

Enrichment of *SOX8* expression in oligodendrocytes is to be expected, based upon its role in oligodendrocyte specification and differentiation as discussed previously though, as also highlighted then, the contribution that *SOX8* makes to oligodendrocyte development seems relatively small, and its roles may be assumed without lasting phenotype by its paralogues *SOX9* and *SOX10*. Perhaps an additional role conveyed from the same locus by *CEROX1* in these cells has provided the selective pressure to retain *SOX8* expression evolutionarily in spite of *SOX9* and *SOX10*. Expression of mouse *Cerox1* has been linked to oligodendrocyte lineage development previously: in a study investigating lncRNAs differentially expressed during differentiation of mouse embryonic forebrain-derived neural stem cells, Mercer et al.[132] identified subsets of lncRNAs whose upregulation correlated with specification of GABAergic neurons or oligodendrocyte progenitor lineages during differentiation. *Cerox1* was the only lncRNA identified as upregulated in oligodendrocyte progenitors while being downregulated in GABAergic neurons. In rodents,

most oligodendrocytes are generated in the first few weeks of post-natal development, peaking in the second week[133]. This coincides with the period in which *Cerox1* expression exhibits the greatest increase in the brain, in the first three weeks post-birth. These data support a role for *Cerox1* in oligodendrocyte development, which potentially coincides with its putative energy-regulating function: the differentiation of oligodendrocytes from precursor cells involves upregulation of mitochondrial genes and activity[134], and is highly sensitive to perturbation of respiratory activity, being blocked by even low doses of the complex I inhibitor rotenone[134]. Oligodendrocyte precursor cells and differentiating oligodendrocytes are highly susceptible to oxidative damage *in vitro*[135-138] and *in vivo*[139]. While the differentiation of oligodendrocytes in culture, as used by Schoenfeld et al.[134], will not completely recapitulate the energy requirements of *in vivo* development, these results likely reflect an underlying switch in metabolism intended to meet the substantial costs[140,141] of producing myelin and enveloping axons. Interestingly, while myelination evolutionarily originated in earlier vertebrates, its prevalence seems to have increased over time, with ~50% myelination of the brain in primates[142] compared to ~15% in rodents[143], and this correlates with the increased relative expression levels of *CEROX1*/*Cerox1* within the transcriptomes of these species. From this it might be posited that ongoing *CEROX1* enrichment in developing oligodendrocytes has contributed to the expansion of myelination and the resulting connectivity of cortical areas that contributes to increased evolutionary cognition.

However, the role of the consistent enrichment of *CEROX1* expression in mature oligodendrocytes is less clear. While the human adult CNS does continue to generate new myelinating oligodendrocytes, the rate of turnover is very low (~0.3% per year, equating to only ~15% total turnover by the age of 60)[144], and so does not explain the elevation of *CEROX1* expression. Schoenfeld et al.[134] also showed that fully differentiated oligodendrocytes are much less sensitive to respiratory inhibition by rotenone[134]. Initially, this invokes the observation that *CEROX1* overexpression *in vitro* protects cells against the detrimental effects of rotenone (Sirey et al., 2019[20]), and prompts the suggestion that the high level of *CEROX1* in these cells contributes to this resistance, especially since mature oligodendrocytes are still highly susceptible to oxidative stress[139]. However, the data of Schoenfeld et al.[134] are also in agreement with Fünfschilling et al.[145], who showed that genetic ablation of mitochondrial respiration via complex IV in mature oligodendrocytes *in vivo* does not produce an overt effect upon CNS health, with no death of oligodendrocytes or neurons, and no demyelination, axonal degeneration or inflammation. This is in stark contrast to the effect of analogous

inhibition of respiration in the periphery, which results in severe demyelination and neuropathy, despite the survival of Schwann cells themselves under glycolytic conditions[145]. *CEROX1* is expressed in Schwann cells of the peripheral nervous system, though at a lower level than even oligodendrocyte precursor cells (FANTOM5)[68]. As well as insulation and physical support, the envelopment of neuronal axons by mature oligodendrocytes also provides metabolic sustenance[146]. Respiration deficient oligodendrocytes contribute to an increased pool of intracellular lactate, which is rapidly metabolised by other cells including neurons, being detectable only under anaesthetic inhibition of neuronal mitochondrial activity[145]. This suggests that they are able to survive by aerobic glycolysis[147], and channel lactate or pyruvate into the metabolism of other cells. This is reminiscent of the previously described astrocyte-neuron lactate shuttle[148]. Oligodendrocytes express one of the same monocarboxylate transporter channels MCT1[149] (despite initial studies of MCT1 in astrocytes, it seems that MCT1 is highly enriched in oligodendrocytes, and MCT4[150] is the likely agent in the former[151]) that would allow them to channel lactate and pyruvate into the periaxonal space at the heart of myelin bundles, whence neurons may uptake them via MCT2[151]. Heterozygous MCT1 deletion leads to degenerative axonopathy, despite unaltered myelination by the unaffected oligodendrocyte population, as does oligodendrocyte-specific MCT1 deletion[151]. This supports an MCT1-dependent insulation-independent role for oligodendrocytes that is essential for neuronal health. Accordingly, it has also been shown that the activation of NMDA receptors in oligodendrocytes in response to glutamate released by axonal action potentials stimulates uptake of glucose via GLUT1[152]. This promotes myelination during development, since glucose-derived lipids are rate-limiting in myelin production[149], as well as the release of lactate from mature oligodendrocytes that may feed into neurons[152]. Data from *in vitro* differentiation of stem cells into neurons indicating that expression of *CEROX1* rises in neurons during brain development would be consistent with their oxidative catabolism of such transferred nutrients. The idea that *CEROX1* is most highly expressed in a glycolytic environment within oligodendrocytes, however, is highly inconsistent with existing evidence to support its oxidative respiration-promoting role. The argument made above, that the atypical metabolism of the liver might explain the conserved absence of *CEROX1*, could also be applied to mature oligodendrocytes whose role is to channel pre-mitochondrial carbon intermediates into other cells, rather than produce ATP. It is possible that under normal respiratory conditions, oligodendrocytes export primarily pyruvate[145], which also travels through MCT1, leaving glycolysis-derived NADH from which energy must be efficiently harvested by complex I: mature

oligodendrocytes are still energetic, with a large surface area across which a resting membrane potential must be maintained[140]. Perhaps only when respiration is inhibited is this NADH recycled by lactate generation as seen by Lee et al.[151]. Substrates also flow *into* oligodendrocytes, including neuron-derived N-acetylaspartate (NAA), which generates aspartate and acetyl-CoA in oligodendrocytes. However, metabolic labelling studies *in vitro* suggest a slow rate of oxidation of the resulting aspartate, with most acting in the malate-aspartate shuttle, and this suggests that the major role of this NAA transfer is to provide acetyl-CoA for lipid synthesis and regulatory roles[153,154], as does the rise in its synthesis during post-natal development, coinciding with myelin generation[155]. Additionally, glycolytic pyruvate or lactate from astrocytes may feed into oligodendrocytes as well as neurons[142]: genetic ablation of astrocyte-oligodendrocyte coupling causes semi-lethal hypomyelination and glial death[156].

The putative enrichment of *CEROX1* in striosomes is interesting for three reasons: a) striosomes are known to exhibit higher levels of myelination than the striatal matrix[157], and so differential *CEROX1* enrichment on this axis is again consistent with a role in myelinating oligodendrocytes; b) the striosome-matrix system exhibits differential expression of key disease-associated genes such as the LRRK2[158], the gene most prevalently associated with Parkinson's disease[159]; and c) given the putative opposing pattern of *SOX8*, which might be enriched in the matrix, the expression of *CEROX1* in striosomes suggests that these two genes' expression profiles might have limited overlap on an individual cell-by-cell basis. Indeed, examination of a single cell sequencing profile of oligodendrocyte populations across the mouse brain from Marques et al.[160] indicates that *Sox8* expression was detected in only 47% of *Cerox1*-positive oligodendrocytes, while 91% of *Sox8*-positive cells also expressed detectable levels of *Cerox1*. The average expression level of each gene was higher (by 40-48%) when the other was also expressed, resulting from a highly significant association between expression of the two genes (Fisher's exact test $p = 2.4 \times 10^{-13}$). This supports a model of transcriptional co-regulation at the *Cerox1-Sox8* locus but also additional regulation biasing *Cerox1* expression within the oligodendrocyte population. The ~530 nt *CEROX1* ISH probe, designed to target the 3' exon, will not detect every isoform, so there may be an isoform-specific striosome-associated expression pattern. Further exploration of the hypothesis that *CEROX1* is enriched in striosomes was attempted by co-staining of fluorescent ISH (FISH) for *Cerox1* and immunohistochemical (IHC) staining for striosome and matrix markers Oprm1 and Calb1, respectively, in mouse striatal sections (Appendix A7). Unfortunately,

limited resources and time, as well as problems with antibodies cross-reacting with the myelin bundles that punctuate and surround the striatum, marred these tricky experiments.

The metabolism of the neuronal-glial network as a whole is complex, and our understanding incomplete. The full responsibilities of oxidative metabolism, and of *CEROX1* in regulating it, are yet to be uncovered in mature oligodendrocytes, potentially alongside an additional rationale for the evolutionarily conserved expression of *SOX8*. It is also possible that the role of *CEROX1*, as well as the type of cellular metabolism (the ratio between glycolytic and oxidative metabolism), is subject to switching at the termination of differentiation, implicating a second functionality. It is also possible that *CEROX1* expression is simply tied to that of *SOX8*, due to their common promoter, and that its expression is directed by the evolutionary history of *SOX8*, itself now an sub-effectual remnant in this tissue. The identification of the role(s) that *CEROX1* might play in the regulation of mitochondrial energy and development and maintenance of myelination may have wide implications for human diseases, such as multiple sclerosis (MS) and amyotrophic lateral sclerosis (ALS). Reduced MCT1 expression observed in the brain of ALS patients and in the spinal cord of the SOD1 mutant mouse model[151] likely leads to impaired substrate transport from oligodendrocytes and subsequent axonal energy deficits. In MS, on the other hand, MCT1 is upregulated but neuronal MCT2 is reduced[161], presumably also obstructing axonal energy supply. Additionally, mitochondrial dysfunction and reactive oxygen species (ROS) are increasingly implicated in the pathophysiology of MS[162-165], and this casts the ability of *Cerox1* to reduce ROS production when overexpressed *in vitro* (Sirey et al., 2019[20]) into an exciting light. *CEROX1* expression appears decreased in the white matter tract of the optic chiasm in MS patients (63% reduction)[166] but very similar changes in expression of *SOX8*, *SOX10*, and *OLIG2* suggests that this reflects selective loss of oligodendrocytes as part of the pathophysiological progression of the disease, not a specific genetic downregulation. Similarly reduced expression of *Cerox1* as well as *Sox8*, *Sox10*, and *Olig1*, is observed in a oligodendrocyte lineage-specific *Sox2* deletion, which diminishes precursor proliferation, so impairing oligodendrocyte differentiation and myelination[167]. In an experimental autoimmune encephalomyelitis (EAE) mouse, commonly used as an *in vitro* model of MS, expression of *Cerox1*, *Sox10*, and *Mog* are decreased by ~2-fold in mature myelinating oligodendrocytes, but not in oligodendrocyte precursor cells[168].

Cerox1-deficient mouse models, including inducible oligodendrocyte-specific deletions, would be instrumental in identifying the importance of its role in development and maintenance of this cell type, and

of CNS myelination. Based upon the data discussed above, it would be expected that a decrease in oxidative metabolism resulting from *Cerox1* deletion might impede or delay oligodendrocyte specification and differentiation. The effects of *Cerox1*-deletion in terminally differentiated myelinating oligodendrocytes are less predictable, though deletion of *Cerox1* on a respiration deficient background, such as the Cox10 mutant used by Fünfschilling et al.[145] would ascertain whether any phenotype seen is indeed effected via activity of the electron transport chain. In investigating the potential of *CEROX1* as a therapeutic target for demyelinating diseases such as MS, it would first be informative to modulate *Cerox1* expression levels in a model of EAE as well as in the demyelination-remyelination process elicited by dosing and withdrawal of cuprizone[169].

Finally, *Cerox1* (there named *lnc6689*) was recently identified by Delás et al. (preprint)[170] as a putative regulator of the differentiation of haematopoietic stem cells (HSCs). Knockdown of *Cerox1* in an *in vivo* reconstituted bone marrow reporter system resulted in lineage bias towards myeloid cells, indicating that *Cerox1* plays a role in promoting development of the lymphoid lineage[170]. Interestingly, *Cerox1* is expressed in HSCs, but in neither common lymphoid nor myeloid progenitors, and then in pro- and pre-B cells[170]. HSCs are highly glycolytic[171], while upregulated oxidative phosphorylation was shown in common myeloid progenitors, but common lymphoid progenitors were not examined[170]. Several publications have identified mitochondrial activity and ROS production as signals that regulate B cell activation and fate determination[172-175]. It remains to be seen whether this is related to the other proposed *Cerox1*-oxidative phosphorylation axis, but it is interesting to speculate that *Cerox1* might pleotropically regulate disparate paths of differentiation.

REFERENCES

1. Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635 (2014).
2. Haerty, W. & Ponting, C.P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* **14**, R49 (2013).
3. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology* **14**, 103 (2007).
4. Ponjavic, J., Ponting, C.P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-65 (2007).
5. Haerty, W. & Ponting, C.P. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**, 333-46 (2015).
6. Ulitsky, I. & Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26-46 (2013).
7. Hezroni, H., Koppstein, D., Schwartz, Matthew G., Avrutin, A. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11**, 1110-1122 (2015).
8. Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**, 925-38 (2010).
9. Stadler, P.F. Evolution of the Long Non-coding RNAs MALAT1 and MEN beta/epsilon. *Advances in Bioinformatics and Computational Biology* **6268**, 1-12 (2010).
10. Ulitsky, I., Shkumatava, A., Jan, Calvin H., Sive, H. & Bartel, David P. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* **147**, 1537-1550 (2011).
11. Pang, K.C., Frith, M.C. & Mattick, J.S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**, 1-5 (2006).
12. Smith, K.N., Starmer, J., Miller, S.C., Sethupathy, P. & Magnuson, T. Long Noncoding RNA Moderates MicroRNA Activity to Maintain Self-Renewal in Embryonic Stem Cells. *Stem Cell Reports* **9**, 108-121 (2017).
13. Kleaveland, B., Shi, C.Y., Stefano, J. & Bartel, D.P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* **174**, 350-362.e17 (2018).
14. Sarangdhar, M.A., Chaubey, D., Srikakulam, N. & Pillai, B. Parentally inherited long non-coding RNA Cyrano is involved in zebrafish neurodevelopment. *Nucleic Acids Research* **46**, 9726-9735 (2018).
15. Ponjavic, J., Oliver, P.L., Lunter, G. & Ponting, C.P. Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain. *PLOS Genetics* **5**, e1000617 (2009).
16. Vance, K.W., Sansom, S.N., Lee, S., Chalei, V. *et al.* The long non-coding RNA Paupar regulates the expression of both local and distal genes. *EMBO J* **33**, 296-311 (2014).
17. Chalei, V., Sansom, S.N., Kong, L., Lee, S. *et al.* The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. *Elife* **3**, e04530 (2014).
18. Tan, J.Y., Vance, K.W., Varela, M.A., Sirey, T. *et al.* Cross-talking noncoding RNAs contribute to cell-specific neurodegeneration in SCA7. *Nat Struct Mol Biol* **21**, 955-61 (2014).
19. Pavlaki, I., Alammari, F., Sun, B., Clark, N. *et al.* The long non-coding RNA Paupar promotes KAP1-dependent chromatin changes and regulates olfactory bulb neurogenesis. *The EMBO journal* **37**, e98219 (2018).
20. Sirey, T.M., Roberts, K., Haerty, W., Bedoya-Reina, O. *et al.* The long non-coding RNA Cerox1 is a post transcriptional regulator of mitochondrial complex I catalytic activity. *eLife* **8**, e45051 (2019).
21. Péterfy, M., Ben-Zeev, O., Mao, H.Z., Weissglas-Volkov, D. *et al.* Mutations in LMF1 cause combined lipase deficiency and severe hypertriglyceridemia. *Nature Genetics* **39**, 1483 (2007).
22. Péterfy, M., Bedoya, C., Giacobbe, C., Pagano, C. *et al.* Characterization of two novel pathogenic variants at compound heterozygous status in lipase maturation factor 1 gene causing severe hypertriglyceridemia. *Journal of Clinical Lipidology* (2018).
23. Paterniti, J., Brown, W., Ginsberg, H. & Artzt, K. Combined lipase deficiency (cld): a lethal mutation on chromosome 17 of the mouse. *Science* **221**, 167-169 (1983).

24. Davis, R.C., Ben-Zeev, O., Martin, D. & Doolittle, M.H. Combined lipase deficiency in the mouse. Evidence of impaired lipase processing and secretion. *Journal of Biological Chemistry* **265**, 17960-6 (1990).
25. Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Research* **14**, 62-66 (2004).
26. Sturm, R.A., Dalton, S. & Wells, J.R. Conservation of histone H2A/H2B intergene regions: a role for the H2B specific element in divergent transcription. *Nucleic Acids Research* **16**, 8571-8586 (1988).
27. Schmidt, C., Fischer, G., Kadner, H., Genersch, E. *et al.* Differential effects of DNA-binding proteins on bidirectional transcription from the common promoter region of human collagen type IV genes COL4A1 and COL4A2. *Biochim Biophys Acta* **1174**, 1-10 (1993).
28. Sugimoto, M., Oohashi, T. & Ninomiya, Y. The genes COL4A5 and COL4A6, coding for basement membrane collagen chains alpha 5(IV) and alpha 6(IV), are located head-to-head in close proximity on human chromosome Xq22 and COL4A6 is transcribed from two alternative promoters. *Proceedings of the National Academy of Sciences* **91**, 11679-11683 (1994).
29. Momota, R., Sugimoto, M., Oohashi, T., Kigasawa, K. *et al.* Two genes, COL4A3 and COL4A4 coding for the human α3(IV) and α4(IV) collagen chains are arranged head-to-head on chromosome 2q36 1. *FEBS Letters* **424**, 11-16 (1998).
30. Ito, M., Ishikawa, M., Suzuki, S., Takamatsu, N. & Shiba, T. A rainbow trout SRY-type gene expressed in pituitary glands. *Febs Letters* **377**, 37-40 (1995).
31. Schepers, G.E., Bullejos, M., Hosking, B.M. & Koopman, P. Cloning and characterisation of the Sry-related transcription factor gene Sox8. *Nucleic Acids Research* **28**, 1473-1480 (2000).
32. Bowles, J., Schepers, G. & Koopman, P. Phylogeny of the SOX Family of Developmental Transcription Factors Based on Sequence and Structural Indicators. *Developmental Biology* **227**, 239-255 (2000).
33. Kellerer, S., Schreiner, S., Stolt, C.C., Scholz, S. *et al.* Replacement of the Sox10 transcription factor by Sox8 reveals incomplete functional equivalence. *Development* **133**, 2875-2886 (2006).
34. Sock, E., Schmidt, K., Hermanns-Borgmeyer, I., Bösl, M.R. & Wegner, M. Idiopathic Weight Reduction in Mice Deficient in the High-Mobility-Group Transcription Factor Sox8. *Molecular and Cellular Biology* **21**, 6951-6959 (2001).
35. Schmidt, K., Schinke, T., Haberland, M., Priemel, M. *et al.* The high mobility group transcription factor Sox8 is a negative regulator of osteoblast differentiation. *The Journal of Cell Biology* **168**, 899-910 (2005).
36. O'Bryan, M.K., Takada, S., Kennedy, C.L., Scott, G. *et al.* Sox8 is a critical regulator of adult Sertoli cell function and male fertility. *Developmental biology* **316**, 359-370 (2008).
37. Britsch, S., Goerich, D.E., Riethmacher, D., Peirano, R.I. *et al.* The transcription factor Sox10 is a key regulator of peripheral glial development. *Genes & Development* **15**, 66-78 (2001).
38. Stolt, C.C., Rehberg, S., Ader, M., Lommes, P. *et al.* Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. *Genes & Development* **16**, 165-170 (2002).
39. Stolt, C.C., Lommes, P., Sock, E., Chaboissier, M.-C. *et al.* The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes & Development* **17**, 1677-1689 (2003).
40. Wagner, T., Wirth, J., Meyer, J., Zabel, B. *et al.* Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9. *Cell* **79**, 1111-1120 (1994).
41. Chaboissier, M.C., Kobayashi, A., Vidal, V.I.P., Lutzkendorf, S. *et al.* Functional analysis of Sox8 and Sox9 during sex determination in the mouse. *Development* **131**, 1891-1901 (2004).
42. Stolt, C.C., Lommes, P., Friedrich, R.P. & Wegner, M. Transcription factors Sox8 and Sox10 perform non-equivalent roles during oligodendrocyte development despite functional redundancy. *Development* **131**, 2349-2358 (2004).
43. Stolt, C.C., Schmitt, S., Lommes, P., Sock, E. & Wegner, M. Impact of transcription factor Sox8 on oligodendrocyte specification in the mouse embryonic spinal cord. *Developmental Biology* **281**, 309-317 (2005).
44. Schepers, G., Wilson, M., Wilhelm, D. & Koopman, P. SOX8 Is Expressed during Testis Differentiation in Mice and Synergizes with SF1 to Activate the Amh Promoter in Vitro. *Journal of Biological Chemistry* **278**, 28101-28108 (2003).
45. Akiyama, H., Chaboissier, M.-C., Martin, J.F., Schedl, A. & de Crombrugghe, B. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes & Development* **16**, 2813-2828 (2002).

46. Kataruka, S., Akhade, V.S., Kayyar, B. & Rao, M.R.S. Mrhl Long Noncoding RNA Mediates Meiotic Commitment of Mouse Spermatogonial Cells by Regulating Sox8 Expression. *Molecular and Cellular Biology* **37**(2017).
47. Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S. *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biology* **14**, R131-R131 (2013).
48. Babenko, V.N., Chadaeva, I.V. & Orlov, Y.L. Genomic landscape of CpG rich elements in human. *BMC Evolutionary Biology* **17**, 19 (2017).
49. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1412-1417 (2006).
50. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *Journal of Molecular Biology* **196**, 261-282 (1987).
51. International Human Genome Sequencing, C., Lander, E.S., Linton, L.M., Birren, B. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).
52. Elango, N. & Yi, S.V. Functional Relevance of CpG Island Length for Regulation of Gene Expression. *Genetics* **187**, 1077-1083 (2011).
53. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222-228 (2014).
54. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Research* **20**, 1493-1502 (2010).
55. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936 (2010).
56. Hon, C.-C., Ramilowski, J.A., Harshbarger, J., Bertin, N. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199 (2017).
57. Marchler-Bauer, A. & Bryant, S.H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* **32**, W327-W331 (2004).
58. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, D225-D229 (2011).
59. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Research* **43**, D222-D226 (2015).
60. Marchler-Bauer, A., Bo, Y., Han, L., He, J. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* **45**, D200-D203 (2017).
61. Desai, P.J. A null mutation in the UL36 gene of herpes simplex virus type 1 results in accumulation of unenveloped DNA-filled capsids in the cytoplasm of infected cells. *J Virol* **74**, 11608-18 (2000).
62. Smith, C.A., Bates, P., Rivera-Gonzalez, R., Gu, B. & DeLuca, N.A. ICP4, the major transcriptional regulatory protein of herpes simplex virus type 1, forms a tripartite complex with TATA-binding protein and TFIIB. *J Virol* **67**, 4676-87 (1993).
63. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
64. Shiraki, T., Kondo, S., Katayama, S., Waki, K. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100**, 15776-15781 (2003).
65. Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T. *et al.* High-Efficiency Full-Length cDNA Cloning by Biotinylated CAP Trapper. *Genomics* **37**, 327-336 (1996).
66. Carninci, P. & Hayashizaki, Y. [2] High-efficiency full-length cDNA cloning. in *Methods in Enzymology*, Vol. 303 19-44 (Academic Press, 1999).
67. Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res* **21**, 1150-9 (2011).
68. Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
69. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* **16**, 22 (2015).
70. Shi, X. & Jarvis, D.L. A new RACE method for extremely GC-rich genes. *Analytical biochemistry* **356**, 222-228 (2006).

71. Schnoor, M., Voß, P., Cullen, P., Böking, T. *et al.* Characterization of the synthetic compatible solute homoectoine as a potent PCR enhancer. *Biochemical and Biophysical Research Communications* **322**, 867-872 (2004).
72. Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A. *et al.* Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**, 454 (2012).
73. Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H. *et al.* Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biology* **17**, 14 (2016).
74. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120 (2011).
75. Ji, P., Diederichs, S., Wang, W., Böing, S. *et al.* MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031 (2003).
76. William, C., Romano, R. & Yingqun, H. Emerging roles of long noncoding RNAs in neurological diseases and metabolic disorders. 1 electronic resource (76 pages) (Frontiers Media SA, 2015).
77. Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S. *et al.* Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA Biology* **9**, 1076-1087 (2012).
78. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580 (2013).
79. Guo, J.H., Huang, Q., Studholme, D.J., Wu, C.Q. & Zhao, Z. Transcriptomic analyses support the similarity of gene expression between brain and testis in human as well as mouse. *Cytogenet Genome Res* **111**, 107-9 (2005).
80. Guo, J., Zhu, P., Wu, C., Yu, L. *et al.* In silico analysis indicates a similar gene expression pattern between human brain and testis. *Cytogenetic and Genome Research* **103**, 58-62 (2003).
81. Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Research* **16**, 11-19 (2006).
82. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-27 (2011).
83. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).
84. Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-47 (2014).
85. Zywitza, V., Misios, A., Bunatyan, L., Willnow, T.E. & Rajewsky, N. Single-cell transcriptomics characterizes cell types in the subventricular zone and uncovers molecular defects underlying impaired adult neurogenesis. *bioRxiv* (2018).
86. Gokce, O., Stanley, G.M., Treutlein, B., Neff, N.F. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *Cell Reports* **16**, 1126-1137 (2016).
87. Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).
88. Darr, A.J., Danzi, M.C., Brady, L., Emig-Agius, D. *et al.* Identification of genome-wide targets of Olig2 in the adult mouse spinal cord using ChIP-Seq. *PLoS One* **12**, e0186091 (2017).
89. Weng, Q.J., Chen, Y., Wang, H.B., Xu, X.M. *et al.* Dual-Mode Modulation of Smad Signaling by Smad-Interacting Protein Sip1 Is Required for Myelination in the Central Nervous System. *Neuron* **73**, 713-728 (2012).
90. Gong, S.C., Zheng, C., Doughty, M.L., Losos, K. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917-925 (2003).
91. Graybiel, A.M. & Ragsdale, C.W., Jr. Histochemically distinct compartments in the striatum of human, monkeys, and cat demonstrated by acetylthiocholinesterase staining. *Proc Natl Acad Sci U S A* **75**, 5723-6 (1978).
92. Brimblecombe, K.R. & Cragg, S.J. The Striosome and Matrix Compartments of the Striatum: A Path through the Labyrinth from Neurochemistry toward Function. *Acs Chemical Neuroscience* **8**, 235-242 (2017).
93. Wouterlood, F.G., Hartig, W., Groenewegen, H.J. & Voorn, P. Density gradients of vesicular glutamate- and GABA transporter-immunoreactive boutons in calbindin- and mu-opioid receptor-

defined compartments in the rat striatum. *Journal of Comparative Neurology* **520**, 2123-2142 (2012).
94. Tajima, K. & Fukuda, T. Region-Specific Diversity of Striosomes in the Mouse Striatum Revealed by the Differential Immunoreactivities for Mu-Opioid Receptor, Substance P, and Enkephalin. *Neuroscience* **241**, 215-228 (2013).
95. Davis, M.I., Crittenden, J.R., Feng, A.Y., Kupferschmidt, D.A. *et al.* The cannabinoid-1 receptor is abundantly expressed in striatal striosomes and striosome-dendron bouquets of the substantia nigra. *Plos One* **13**(2018).
96. Schmitt, B.M., Rudolph, K.L.M., Karagianni, P., Fonseca, N.A. *et al.* High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA–tRNA interface. *Genome Research* **24**, 1797-1807 (2014).
97. Lindsay, S.J., Xu, Y., Lisgo, S.N., Harkin, L.F. *et al.* HDBR Expression: A Unique Resource for Global and Individual Gene Expression Studies during Early Human Brain Development. *Front Neuroanat* **10**, 86 (2016).
98. Nayler, S.P., Powell, J.E., Vanichkina, D.P., Korn, O. *et al.* Human iPSC-Derived Cerebellar Neurons from a Patient with Ataxia-Telangiectasia Reveal Disrupted Gene Regulatory Networks. *Frontiers in Cellular Neuroscience* **11**(2017).
99. Wells, C.A., Mosbergen, R., Korn, O., Choi, J. *et al.* Stemformatics: Visualisation and sharing of stem cell gene expression. *Stem Cell Research* **10**, 387-395 (2013).
100. Weick, J.P., Held, D.L., Bonadurer, G.F., Doers, M.E. *et al.* Deficits in human trisomy 21 iPSCs and neurons. *Proceedings of the National Academy of Sciences* **110**, 9962-9967 (2013).
101. Mariani, J., Coppola, G., Zhang, P., Abyzov, A. *et al.* FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375-390 (2015).
102. Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A. *et al.* Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science* **334**, 521 (2011).
103. dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J. *et al.* Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences* **279**, 3491 (2012).
104. dos Reis, M., Donoghue, P.C.J. & Yang, Z. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. *Biology Letters* **10**(2014).
105. Takada, S., DiNapoli, L., Capel, B. & Koopman, P. Sox8 is expressed at similar levels in gonads of both sexes during the sex determining period in turtles. *Developmental Dynamics* **231**, 387-395 (2004).
106. Kumar, S. & Hedges, S.B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917-20 (1998).
107. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034-1050 (2005).
108. Hezroni, H., Perry, R.B.T., Meir, Z., Housman, G. *et al.* A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biology* **18**(2017).
109. Hoeppner, M.P., Lundquist, A., Pirun, M., Meadows, J.R.S. *et al.* An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts. *PLOS ONE* **9**, e91172 (2014).
110. Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A. *et al.* Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res* (2018).
111. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).
112. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics* **5**, 1-19 (2004).
113. McWilliam, H., Li, W., Uludag, M., Squizzato, S. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013).
114. Li, W., Cowley, A., Uludag, M., Gur, T. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* **43**, W580-4 (2015).
115. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-4 (2016).
116. Hou, Z.-C., Sterner, K.N., Romero, R., Than, N.G. *et al.* Elephant Transcriptome Provides Insights into the Evolution of Eutherian Placentation. *Genome Biology and Evolution* **4**, 713-725 (2012).
117. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences* **99**, 803-808 (2002).

118. Efron, B., Halloran, E. & Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences* **93**, 13429-13429 (1996).
119. Kutter, C., Watt, S., Stefflova, K., Wilson, M.D. *et al.* Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. *PLOS Genetics* **8**, e1002841 (2012).
120. Rolfe, D.F. & Brown, G.C. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiol Rev* **77**, 731-58 (1997).
121. Wallace, D.C. Mitochondrial DNA in aging and disease. *Sci Am* **277**, 40-7 (1997).
122. Marques, A.C., Tan, J. & Ponting, C.P. Wrangling for microRNAs provokes much crosstalk. *Genome Biol* **12**, 132 (2011).
123. Tay, Y., Rinn, J. & Pandolfi, P.P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344-52 (2014).
124. Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P. & Stoffel, M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell* **54**, 766-76 (2014).
125. Broderick, J.A. & Zamore, P.D. Competitive endogenous RNAs cannot alter microRNA function in vivo. *Mol Cell* **54**, 711-3 (2014).
126. Ala, U., Karreth, F.A., Bosia, C., Pagnani, A. *et al.* Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc Natl Acad Sci U S A* **110**, 7154-9 (2013).
127. Figliuzzi, M., Marinari, E. & De Martino, A. MicroRNAs as a selective channel of communication between competing RNAs: a steady-state theory. *Biophys J* **104**, 1203-13 (2013).
128. Bosson, A.D., Zamudio, J.R. & Sharp, P.A. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Mol Cell* **56**, 347-59 (2014).
129. Schwanhausser, B., Wolf, J., Selbach, M. & Busse, D. Synthesis and degradation jointly determine the responsiveness of the cellular proteome. *Bioessays* **35**, 597-601 (2013).
130. Dinger, M.E., Gascoigne, D.K. & Mattick, J.S. The evolution of RNAs with multiple functions. *Biochimie* **93**, 2013-2018 (2011).
131. Karlsen, A.S. & Pakkenberg, B. Total Numbers of Neurons and Glial Cells in Cortex and Basal Ganglia of Aged Brains with Down Syndrome—A Stereological Study. *Cerebral Cortex* **21**, 2519-2524 (2011).
132. Mercer, T.R., Qureshi, I.A., Gokhan, S., Dinger, M.E. *et al.* Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci* **11**, 14 (2010).
133. Rivers, L.E., Young, K.M., Rizzi, M., Jamen, F. *et al.* PDGFRA/NG2 glia generate myelinating oligodendrocytes and piriform projection neurons in adult mice. *Nature Neuroscience* **11**, 1392 (2008).
134. Schoenfeld, R., Wong, A., Silva, J., Li, M. *et al.* Oligodendroglial differentiation induces mitochondrial genes and inhibition of mitochondrial function represses oligodendroglial differentiation. *Mitochondrion* **10**, 143-150 (2010).
135. Roth, A.D. & Núñez, M.T. Oligodendrocytes: Functioning in a Delicate Balance Between High Metabolic Requirements and Oxidative Damage. in *Glial Cells in Health and Disease of the CNS* (ed. von Bernhardi, R.) 167-181 (Springer International Publishing, Cham, 2016).
136. Back, S.A., Han, B.H., Luo, N.L., Chricton, C.A. *et al.* Selective Vulnerability of Late Oligodendrocyte Progenitors to Hypoxia–Ischemia. **22**, 455-463 (2002).
137. French, H.M., Reid, M., Mamontov, P., Simmons, R.A. & Grinspan, J.B. Oxidative stress disrupts oligodendrocyte maturation. *Journal of Neuroscience Research* **87**, 3076-3087 (2009).
138. Thorburne, S.K. & Juurlink, B.H.J. Low Glutathione and High Iron Govern the Susceptibility of Oligodendroglial Precursors to Oxidative Stress. *Journal of Neurochemistry* **67**, 1014-1022 (1996).
139. Giacci, M.K., Bartlett, C.A., Smith, N.M., Iyer, K.S. *et al.* Oligodendroglia Are Particularly Vulnerable to Oxidative Damage after Neurotrauma In Vivo. **38**, 6491-6504 (2018).
140. Harris, J.J. & Attwell, D. The Energetics of CNS White Matter. *The Journal of Neuroscience* **32**, 356-371 (2012).
141. Butts, B.D., Houde, C. & Mehmet, H. Maturation-dependent sensitivity of oligodendrocyte lineage cells to apoptosis: implications for normal development and disease. *Cell Death And Differentiation* **15**, 1178 (2008).
142. Saab, A.S., Tzvetanova, I.D. & Nave, K.-A. The role of myelin and oligodendrocytes in axonal energy metabolism. *Current Opinion in Neurobiology* **23**, 1065-1072 (2013).
143. Salzer, J.L. & Zalc, B. Myelination. *Current Biology* **26**, R971-R975 (2016).
144. Yeung, Maggie S.Y., Zdunek, S., Bergmann, O., Bernard, S. *et al.* Dynamics of Oligodendrocyte Generation and Myelination in the Human Brain. *Cell* **159**, 766-774 (2014).

145. Fünfschilling, U., Supplie, L.M., Mahad, D., Boretius, S. *et al.* Glycolytic oligodendrocytes maintain myelin and long-term axonal integrity. *Nature* **485**, 517 (2012).
146. Philips, T. & Rothstein, J.D. Oligodendroglia: metabolic supporters of neurons. *Journal of Clinical Investigation* **127**, 3277-3286 (2017).
147. Kane, D.A. Lactate oxidation at the mitochondria: a lactate-malate-aspartate shuttle at work. *Frontiers in Neuroscience* **8**(2014).
148. Pellerin, L. & Magistretti, P.J. Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 10625-10629 (1994).
149. Rinholm, J.E., Hamilton, N.B., Kessaris, N., Richardson, W.D. *et al.* Regulation of Oligodendrocyte Development and Myelination by Glucose and Lactate. *The Journal of Neuroscience* **31**, 538-548 (2011).
150. Hanu, R., McKenna, M., O'Neill, A., Resneck, W.G. & Bloch, R.J. Monocarboxylic acid transporters, MCT1 and MCT2, in cortical astrocytes in vitro and in vivo. *American Journal of Physiology-Cell Physiology* **278**, C921-C930 (2000).
151. Lee, Y., Morrison, B.M., Li, Y., Lengacher, S. *et al.* Oligodendroglia metabolically support axons and contribute to neurodegeneration. *Nature* **487**, 443 (2012).
152. Saab, Aiman S., Tzvetavona, Iva D., Trevisiol, A., Baltan, S. *et al.* Oligodendroglial NMDA Receptors Regulate Glucose Import and Axonal Energy Metabolism. *Neuron* **91**, 119-132 (2016).
153. Moffett, J.R., Arun, P., Ariyannur, P.S. & Namboodiri, A.M.A. N-Acetylaspartate reductions in brain injury: impact on post-injury neuroenergetics, lipid synthesis, and protein acetylation. *Frontiers in Neuroenergetics* **5**, 11 (2013).
154. Amaral, A.I., Hadera, M.G., Kotter, M. & Sonnewald, U. Oligodendrocytes Do Not Export NAA-Derived Aspartate In Vitro. *Neurochemical Research* **42**, 827-837 (2017).
155. Burri, R., Steffen, C. & Herschkowitz, N. N-Acetyl-L-Aspartate is a Major Source of Acetyl Groups for Lipid Synthesis during Rat Brain Development. *Developmental Neuroscience* **13**, 403-411 (1991).
156. Tress, O., Maglione, M., May, D., Pivneva, T. *et al.* Panglial Gap Junctional Communication is Essential for Maintenance of Myelin in the CNS. *The Journal of Neuroscience* **32**, 7499-7518 (2012).
157. Grigoletto, J., Pukaß, K., Gamliel, A., Davidi, D. *et al.* Higher levels of myelin phospholipids in brains of neuronal α-Synuclein transgenic mice precede myelin loss. *Acta Neuropathologica Communications* **5**, 37 (2017).
158. Mandemakers, W., Snellinx, A., O'Neill, M.J. & de Strooper, B. LRRK2 expression is enriched in the striosomal compartment of mouse striatum. *Neurobiology of Disease* **48**, 582-593 (2012).
159. Singh, A., Zhi, L. & Zhang, H. LRRK2 and mitochondria: Recent advances and current views. *Brain Research* (2018).
160. Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326-1329 (2016).
161. Nijland, P.G., Michailidou, I., Witte, M.E., Mizee, M.R. *et al.* Cellular distribution of glucose and monocarboxylate transporters in human brain white matter and multiple sclerosis lesions. *Glia* **62**, 1125-41 (2014).
162. Mao, P.Z. & Reddy, P.H. Is multiple sclerosis a mitochondrial disease? *Biochimica Et Biophysica Acta-Molecular Basis of Disease* **1802**, 66-79 (2010).
163. Su, K., Bourdette, D. & Forte, M. Mitochondrial dysfunction and neurodegeneration in multiple sclerosis. *Frontiers in Physiology* **4**, 169 (2013).
164. Witte, M.E., Mahad, D.J., Lassmann, H. & van Horssen, J. Mitochondrial dysfunction contributes to neurodegeneration in multiple sclerosis. *Trends in Molecular Medicine* **20**, 179-187 (2014).
165. Campbell, G. & Mahad, D.J. Mitochondrial dysfunction and axon degeneration in progressive multiple sclerosis. *Febs Letters* **592**, 1113-1121 (2018).
166. Itoh, N., Itoh, Y., Tassoni, A., Ren, E. *et al.* Cell-specific and region-specific transcriptomics in the multiple sclerosis model: Focus on astrocytes. *Proceedings of the National Academy of Sciences* **115**, E302-E309 (2018).
167. Zhang, S., Zhu, X., Gui, X., Croteau, C. *et al.* Sox2 Is Essential for Oligodendroglial Proliferation and Differentiation during Postnatal Brain Myelination and CNS Remyelination. *Journal of Neuroscience* **38**, 1802-1820 (2018).
168. Falcão, A.M., van Bruggen, D., Marques, S., Meijer, M. *et al.* Disease-specific oligodendrocyte lineage cells arise in multiple sclerosis. *Nature Medicine* (2018).

169. Torkildsen, Ø., Brunborg, L.A., Myhr, K.M. & Bø, L. The cuprizone model for demyelination. *Acta Neurologica Scandinavica* **117**, 72-76 (2008).
170. Delás, M.J., Jackson, B.T., Kovacevic, T., Vangelisti, S. *et al.* LncRNA Spehd regulates hematopoietic stem cells and progenitors and is required for multilineage differentiation. *bioRxiv* (2018).
171. Takubo, K., Nagamatsu, G., Kobayashi, Chiharu I., Nakamura-Ishizu, A. *et al.* Regulation of Glycolysis by Pdk Functions as a Metabolic Checkpoint for Cell Cycle Quiescence in Hematopoietic Stem Cells. *Cell Stem Cell* **12**, 49-61 (2013).
172. Jang, K.-J., Mano, H., Aoki, K., Hayashi, T. *et al.* Mitochondrial function provides instructive signals for activation-induced B-cell fates. *Nature Communications* **6**, 6750 (2015).
173. Waters, L.R., Ahsan, F.M., Wolf, D.M., Shirihai, O. & Teitell, M.A. Initial B Cell Activation Induces Metabolic Reprogramming and Mitochondrial Remodeling. *iScience* **5**, 99-109 (2018).
174. Akkaya, M., Traba, J., Roesler, A.S., Miozzo, P. *et al.* Second signals rescue B cells from activation-induced mitochondrial dysfunction and death. *Nature Immunology* **19**, 871-884 (2018).
175. Stein, M., Dütting, S., Mougiakakos, D., Bösl, M. *et al.* A defined metabolic state in pre B cells governs B-cell development and is counterbalanced by Swiprosin-2/EFhd1. *Cell Death And Differentiation* **24**, 1239 (2017).

CHAPTER 4: A HIDDEN PROTEIN – IS *CEROX1* TRULY A NON-CODING RNA?

INTRODUCTION

4.1 Hiding in plain sight: not all 'lncRNAs' are entirely non-coding

During the discovery of small RNAs (e.g. miRNAs) and particularly lncRNAs, huge swathes of the genome hitherto deemed 'junk' were reassigned to the class of potentially functional non-coding RNAs. The two apparently opposing categories – protein-coding and non-coding – have since formed the basis for an almost entirely dichotomous system of gene organisation.

In this introduction, several maxims will be discussed that have redefined the boundary between coding and non-coding genes as well as the interpretation of the current forms of evidence utilised in classifying genes into these, and other, categories. Thereon, this chapter will focus on *CEROX1* as an example of an RNA whose protein-coding capacity has not previously been thoroughly interrogated, beginning with an appraisal of the data regarding the bioinformatic and experimental data informing the current classification of *CEROX1* as a lncRNA.

4.1.1 Some 'non-coding' RNAs have been shown to produce microproteins from small ORFs

Many genes classified as lncRNAs contain substantial open reading frames (ORFs). Aside from a lack of basic physical evidence for an encoded protein, these ORFs were initially dismissed from further consideration as protein-coding units on the basis of a lack of evolutionary conservation, the lack of a canonical AUG start codon[1], or their small sizes compared to the somewhat arbitrary traditional 100 amino acid threshold[2].

However, literature from the past 15 years has identified hundreds of small ORFs (smORFs; less than 100 aa but some as small as 11 aa[3]) that are translated in all eukaryotic species examined, from yeast[4-6] and soybean[7] to *Drosophila*[3,8,9], zebrafish[10], and human[10,11]. Some smORF-encoded proteins are conserved throughout mammals, vertebrates, or metazoa, and those microproteins whose expression and function have been thoroughly characterised have been associated with diverse functions. For example, the vertebrate-conserved smORF-derived Toddler is a necessary regulator of cell migration during gastrulation

in zebrafish embryogenesis[12], while the human-mouse-conserved protein SPAR modulates the activity of mTORC1 during muscle regeneration[13]. Interestingly, smORFs are not found only in the nuclear genome: for example, the mitochondrially-encoded smORF humanin regulates apoptosis by interaction with Bax[14].

In exception to two other 'rules' of eukaryotic translation, some microproteins are derived from ORFs without canonical AUG start codons, and this can permit multicistronic translation. For example, the protein-coding mRNA *FRAT2* possesses a small translated ACG-initiated upstream ORF in addition to the primary coding region, which is a typical AUG-initiated ORF[11]. Given that mutation of the upstream start codon to AUG ablates translation of the primary ORF, this suggests that bicistronic translation is permitted by 'leaky' translation outside of the canonical AUG-scanning process.

Although *CEROX1* contains several predicted small ORFs, they will not be the focus of the experimental work in this chapter. However, smORFs serve as excellent illustrations of the discovery of exceptions to traditional theories and criteria of translation.

4.1.2 'Coding non-coding RNAs' illustrate that coding potential need not diminish the functional importance of non-coding RNAs

All data examined in this study thus far indicate that *CEROX1* exhibits functionality as an RNA molecule and that this forms the basis for its evolutionary conservation (Chapter 3). Therefore, any protein-coding capacity that might yet be identified would be an addition of function. As discussed in Chapter 1, lncRNAs have been evidenced to play diverse roles throughout the cell, via a variety of mechanisms, and there exists a subset of lncRNAs that have been shown to be bifunctional. Among these are examples of "coding non-coding RNAs" (cncRNAs)[15], which appear to effect their roles via both RNA and protein forms. As well as rebutting the traditionally dichotomous model of coding and non-coding RNAs, these cncRNAs raise interesting questions about the evolution of gene function. One corollary of the widely commended RNA world hypothesis, which utilises the powerful molecular diversity of RNA to posit that it was the main self-sustaining component of early life prior to protein and DNA[16], is that protein-coding genes must at some point have evolved from non-coding ones[17].

4.1.2.1 The coding non-coding RNA *SRA* is not only bifunctional but self-interacting

The first example of these cncRNAs was steroid receptor RNA activator (*SRA*), which was initially identified as an RNA regulator of nuclear steroid receptor signalling such as that of progesterone, oestrogen, androgen, and glucocorticoid receptors[18]. Additionally, it was then shown to influence the activities of other transcription factors including the developmental transcription factor GATA3 and the myogenic differentiation regulator MyoD[18]. Originally, despite the presence of a substantial ORF, the failure to detect any translation products led to the classification of *SRA* as a functional lncRNA. However, the subsequent identification of additional isoforms with an extended 5' end led to the annotation of a slightly longer ORF with an initiating start codon that results in translation of a protein, SRAP. *SRA*/SRAP thereby became the first example of a characterised gene with function at both the RNA and protein level[19]. Both sequences are well conserved throughout vertebrates. Most interestingly, it was observed that the two molecular species are not entirely independent. While SRAP does not physically interact with MyoD or *SRA* RNA directly[20], it is able to modulate the MyoD-regulating ability of SRA[21]. The perturbed relative expression levels of *SRA* RNA and SRAP during myogenic differentiation and also breast cancer tumours[22] has led to a model whereby the function of the gene depends upon the balance of *SRA* and SRAP, whereby there must be a sufficient excess of *SRA* RNA in order to activate MyoD and myogenesis[23]. The production of multiple molecular species from one gene therefore allows the development of complex and highly tuned regulatory programs, in which the balance of activities allows for incorporation of wider molecular cues.

4.1.2.2 RNA-mediated function adds a layer of complexity to existing protein-protein interactions

SRA/SRAP is not the only bifunctional gene whose products act in close proximity to one another. The multifaceted interaction between the tumour suppressor protein p53 and its regulator protein Mdm2 is extensively characterised[24]. The ubiquitin ligase activity of Mdm2 targets p53 protein for ubiquitin-targeted proteasome degradation, but Mdm2 also promotes translation of the *TP53* mRNA into two proteins, one of which is insensitive to Mdm2-induced degradation[25]. In concert with the transcriptional activation of Mdm2 expression by p53, this leads to a tight circuit of feedback regulation in which Mdm2 balances the synthesis and degradation of p53. However, it has emerged that this regulatory loop is not mediated solely by the protein species. *TP53* mRNA also interacts with Mdm2 protein, increasing its p53 translation-promoting role, while inhibiting its p53-degrading ubiquitination activity[26]. Remarkably, the domain of the *TP53* mRNA

that interacts with Mdm2 protein is the same sequence that does so when translated, suggesting that the two levels of function might have evolved in parallel[24].

4.1.2.3 Commonalities between characterised cncRNAs

These studies, one in which protein-coding activity was ascribed to an already functional RNA, and another in which an mRNA gained new relevance beyond the transfer of information between DNA and protein, have acted as proofs of the principle that RNAs already characterised, be they protein-coding or non-coding, may have uncharacterised functions mediated by a protein. In theory, any RNA, coding or not, can evolve additional functions[17]. Both the above examples illustrate the apparent evolution of bifunctionality involving two species acting in the same pathway, but this is not always the case.

IRS1 encodes the insulin receptor substrate 1, a target of insulin signalling that promotes growth and proliferation. As well as encoding this protein, the *Irs1* mRNA 5'-UTR comprises a sequence complementary to that of the cell cycle G_1/S-checkpoint regulator retinoblastoma protein, *Rb1*[27]. The complementarity of these sequences induces a reciprocal relationship between their expression levels in myoblasts, contributing to the balance of growth and differentiation. Therefore, even though Irs1 protein and *Irs1* mRNA act on different molecular targets, they seemingly contribute to a common cellular goal. This likely reflects the conflict of selective evolutionary pressures associated with evolving multiple functions in the same locus[17].

Another effect of the added complexity of multiple molecular instigators from one gene, especially where both may have potent effects on the same target or pathway, is the requirement for further regulation, isolating or balancing the species temporally, spatially, or according to environment. Several mechanisms of segregation are demonstrated by the maternal effect gene *osk*, perhaps the best-characterised bifunctional RNA in *Drosophila*. Though the exact mechanism is unknown, mutation analysis shows that *osk* RNA is important for oogenesis (and distinctly so from Osk (Oskar) protein)[28], during which time precocious translation of *osk* RNA is inhibited. Binding of Bru (Bruno) to the 3'UTR of *osk* recruits Cup, which binds to the 5' cap-binding protein eIF4E, and prevents recruitment of additional translational machinery[29]. Within oocytes, however, increased polyadenylation of *osk* promotes translation[30]. Furthermore, in the oocyte and subsequent zygote, *osk* RNA is highly localised, being transported to the posterior pole[31], where it is

exclusively translated and plays a crucial role in anteroposterior patterning by de-repressing translation of Nanos[32].

As well as endogenous segregation of molecular function, this has ramifications for exogenous gene modulation, in identifying cncRNAs and targeting them therapeutically. Both *SRA*/SRAP and *TP53*/p53 exemplify an important facet of bifunctional RNAs: apparent exogenous or endogenous perturbation of a molecule derived from a gene does not necessarily limit the perturbation to that molecular target of interest; all molecular species must be considered[23]; i.e. depending on the function of the exact site within a gene, mutations might affect the sequence or structure of the RNA, the protein, or both. This will be further discussed later in reference to *CEROX1*, its putative translation, and the therapeutic potential of targeting RNAs.

4.1.3 Translation is also associated with regulation of RNA stability, and possibly antigen production

In spite of the examples above of lncRNAs that have been shown to conduct additional functionality via the production of a *bona fide* protein, many lncRNAs with (sm)ORFs that have been shown reproducibly to associate with translational machinery[33] (see below) do not appear to produce detectable proteins. This raises the question: does the apparent translation of non-coding RNAs hold importance outside of protein production in a traditional sense?

One of the other major pathways in which the ribosome participates is nonsense-mediated decay (NMD), a mRNA quality-control mechanism during which transcripts are subjected to a 'pioneer round' of translation. Mediated initially by the prolonged binding of the ATP-dependent RNA helicase UPF1 (up-frameshift 1) to an RNA target, a complex of NMD factors forms at a translational termination site. Promiscuous binding of UPF1 to the 5'UTR or coding sequence is removed by the scanning of the ribosome. Under normal circumstances, the assembly of the NMD does not lead to accelerated degradation, since UPF1 remains inactive, and the complex disassembles with the release of the ribosome. On mutated, erroneous transcripts, however, several parameters can lead to activation. The best characterised and most efficient applies to mis-spliced or prematurely terminated transcripts. The exon junction complex (EJC), placed ~20-24 nucleotides upstream of an exon-exon junction, activates UPF1 activity via EJC-associated UPF2 and

UPF3A/B. Normally the EJC would be removed by the procession of the ribosome, but termination of translation more than ~50-55 nucleotides upstream of the EJC (~70-80 nucleotides upstream of the exon-exon junction) leads to its retention, and the subsequent activation of NMD. This results in degradation of many transcripts with introns in their 3'UTRs. Additionally, there is an association between long 3'-UTRs and a less efficient NMD pathway, facilitated by the inhibition of NMD by the polyA-binding protein PABPC1 and its associated factor eIF4G.

Given that many lncRNA-ORFs fall upstream of an exon-exon junction, it might therefore be expected that any association between these transcripts and the translational machinery is due to the NMD process. The well-known multifunctional lncRNA *GAS5*, from the introns of which are produced 9-10 snoRNAs, is rapidly degraded by translation-dependent NMD[34]. Growth arrest resulting from agents such as rapamycin and cycloheximide sharply inhibits the pioneer translation of *GAS5* and results in RNA accumulation[35]. However, examples of transcripts containing ORFs followed by multiple exon-exon junctions have been associated with stable translation, such as that of the human microprotein NoBody, which is translated upstream of two exon-exon junctions[36]. Like many of the criteria for RNA and protein behaviour, therefore, there are exceptions to the recruitment of NMD by upstream ORFs.

Interestingly, translating ribosome affinity purification (TRAP)[37,38] data in *Arabidopsis* suggest that association with ribosomes, and translation of smORFs, may *increase* stability and longevity of lncRNAs[39].

It has also been posited that the production of rare proteins from lncRNAs or cryptic ORFs in protein-coding genes might play a role in cellular antigenicity. Ribosome profiling of human cytomegalovirus (HMCV) transcripts identified many novel transcripts and variants containing ORFs, the translation of which was demonstrated by epitope tagging[40]. Interestingly, some of these proteins elicit cellular immune responses in T cells of HMCV-positive, but not HMCV-negative, donors[41], suggesting that the immune systems of the former have been exposed to these proteins as natural antigens, despite their low abundance. Indeed, it has been shown that a large portion of nascent polypeptides are degraded by the proteasome before even being released from the ribosome (co-translational protein degradation), including ~33% of ribosome-bound polypeptides in yeast[42]. Similarly, ~35% of proteins in mammalian cells (human HeLa and hamster fibroblasts) are degraded within 2 hours[43] of synthesis, and ~30% of newly-synthesised proteins in HeLa are

"defective ribosomal products", DRiPs. It has been suggested to varying degrees[44] that these DRiPs contribute to the pool of peptides presented by MHC class I complexes[45,46], and therefore that even transcripts that produce rare polypeptides in an abortive pioneering round of translation that elicits their degradation via NMD, might still contribute to the cellular antigen profile.

4.1.4 Thorough case-by-case examination is needed to classify any RNA as non-coding

Although these examples are limited, and are not intended to suggest that all RNAs are in fact coding, they do illustrate that initial categorisations of some genes as lncRNAs are likely to be incorrect. As shown in Chapter 3, *CEROX1* has been conserved throughout Eutheria in its sequence and in its tissue-specific transcription, and this supports its being ascribed a putative role conveyed by its functionality as an RNA molecule, by orthology to mouse *Cerox1*. However, this does not preclude possible coding-potential and an associated extension of functionality. Especially in light of the incomplete expression-function correlation highlighted in the discussion of Chapter 3, it is important to consider all possible functionalities of *CEROX1*.

The approaches that may be applied to investigating the coding potential of an RNA fall into two categories: a) examination of its characteristics with respect to the known trends of protein-coding and non-coding transcriptomes, which begins, but by no means should end, with the identification of possible ORFs; and b) experimental investigation seeking to identify physical evidence of transcription or a resulting protein. Both of these approaches will be exemplified in this chapter by application to *CEROX1*.

4.2 Is *CEROX1* simply a long-non coding RNA?

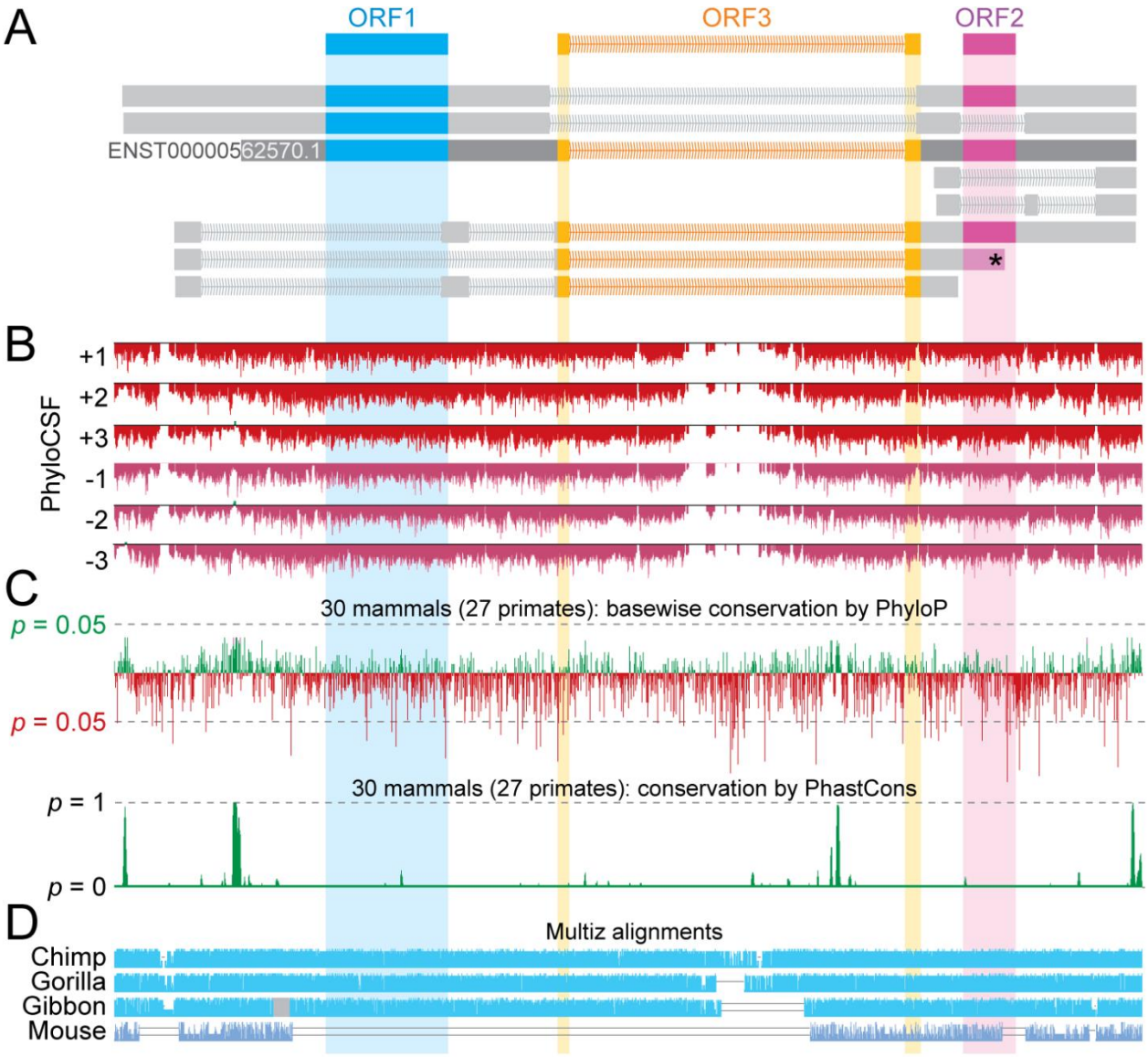
Here, the current approaches to identifying likely protein-coding transcripts, on both a whole-genome and individual basis, will be explored, with especial reference to *CEROX1*.

4.2.1 *CEROX1* contains a number of potential ORFs, including one of substantial length

CEROX1 was originally identified as a long non-coding RNA and has thus far been treated as such. Nevertheless, all *CEROX1* isoform sequences, when translated *in silico* with ATGpr (atgpr.dbcls.jp)[47], have the potential to encode at least several polypeptides, including three ORFs larger than 50 amino acids (Figure 4.1A). The longest CEROX1-ORF is the 702 nucleotide-long CEROX1-ORF1, which would encode a protein of 234 amino acids. This is substantially greater than the customary 100 amino acid threshold for overlooked proteins. As demonstrated by Dinger et al.[48] using a randomly generated population of 20,000 sequences, the expected maximum length of random ORFs within a sequence increases in a logarithmic fashion as the sequence lengthens. For *CEROX1*, at 3.2 kb, the distribution of random ORF lengths has a mean length of ~260 nucleotides, and a standard deviation of ~73 nucleotides. This places CEROX1-ORF1 over six standard deviations from the mean ($p = 7.9 \times 10^{-8}$), suggesting that the existence of an ORF of its size is highly unlikely to be coincidental. However, this simple metric does not account for any other factors that refute the traditional coding potential of the sequence, chiefly among them a lack of evolutionary conservation.

4.2.2 Existing bioinformatic data inconclusively support a non-coding classification of *CEROX1*

The categorisation of genes as coding or non-coding, while readily attempted, is not trivial, and a number of metrics and algorithms have been developed to achieve this. Their application in a preliminary assessment of *CEROX1* will be discussed hereon.



E

Open reading frame	Identity to Kozak rule (A/GXXATGG)	ORF length (aa)	Predicted size (kDa)
CEROX1-ORF1	cXXATGc	234	23.4
CEROX1-ORF2	cXXATGt	100	10.5
CEROX1-ORF3	AXXATGG	52	5.5

Figure 4.1: The three largest open reading frames in *CEROX1* are not well conserved. (A) The three ORFs larger than 50 amino acids in length (predicted by ATGpr, atgpr.dbcls.jp[47]) are shown here annotated across eight *CEROX1* isoform transcripts. Only one isoform, ENST00000562570.1, contains all three ORF sequences. One isoform (*) contains a truncated CEROX1-ORF2 sequence, with no STOP codon. PhyloCSF indicates negative scores for the alignment of 30 mammals (27 primates) across the putative ORF elements **(B)**, identifying them as unconserved (on a protein-coding scale) and likely non-coding. Likewise, PhyloP **(C)** indicates a low probability of conserved elements. MultiZ alignments **(D)** across the *CEROX1* locus show that the ORF nucleotide sequences are broadly conserved in some primates (see also Figure 4.2 for polypeptide alignments), but that none of the three CEROX1-ORF sequences are present in mouse. **(E)** Only CEROX1-ORF3 has a consensus Kozak sequence.

4.2.2.2 The CEROX1-ORFs are not well conserved evolutionarily

PhyloCSF, which calculates a likelihood that a sequence within a multi-species alignment represents a protein-coding element, based on a statistical analysis of codon-synonymous and -non-synonymous mutations[49], indicates that *CEROX1* contains no conserved ORFs (Figure 4.1B). In contrast to the conservation of RNA sequence across the alignable portions of human *CEROX1* and mouse *Cerox1*, none of the sequences encoding the three largest CEROX1-ORFs are conserved in mouse (Figure 4.1D). However, the CEROX1-ORF1 start codon and amino acid sequence are conserved in recent primate ancestors (the catarrhine clade of apes and 'old world' monkeys, especially the hominids including chimpanzee and gorilla; Figure 4.2) – this will be expanded upon later in this study. By the standard of most protein-coding genes, this low level of conservation is supportive of a non-coding classification, though there are at least three human-only protein-coding genes known: *CLLU1*, *C22orf45*, and *DNAH10OS*[50]. The authors whose work recently identified these three 'orphans' also estimate by extrapolation that there may be around another 15 human-specific protein-coding genes in the human genome[50], these having evolved from ancestrally non-coding sequences since the human-chimpanzee divergence. Even in spite of those proteins whose essentiality has made them integral parts of the genomes of species for billions of years, the estimated 7-15 million years since the human-gorilla divergence[51-53] is sufficient time for a sequence to gain an independent function or additional relevance[52], and *CEROX1* might represent another example of human- or hominid clade-specific protein-coding capacity.

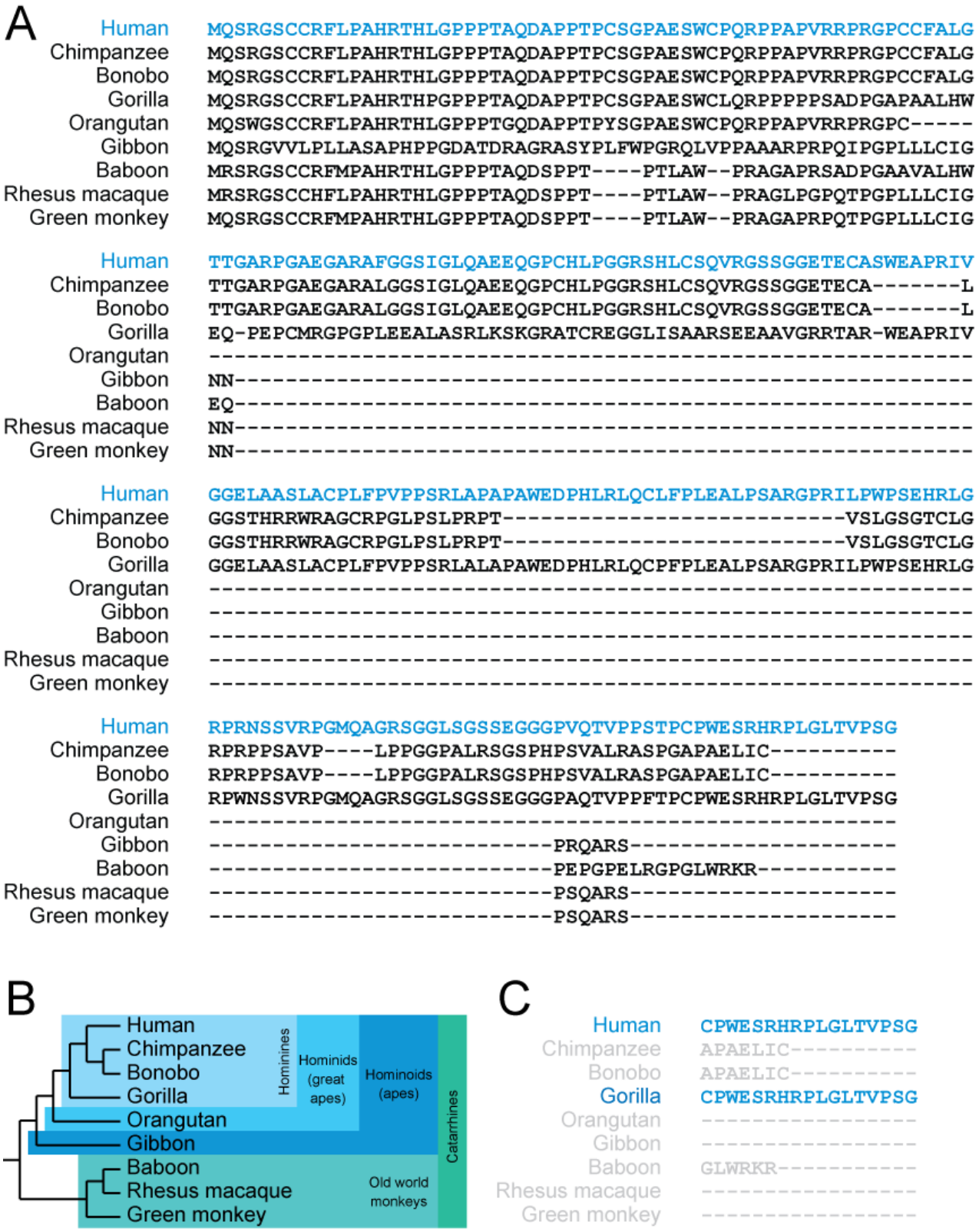


Figure 4.2: The CEROX1-ORF1 sequence is partially conserved within recent primate ancestors. (A) Alignment of putative CEROX1-ORF1 protein sequences in catarrhine species shows a high level of conservation at the N-terminal end and in the middle of the ORF, particularly in the hominid clade (great apes) **(B)**. **(C)** At the C-terminus, the sequence is perfectly conserved only in gorilla.

4.2.2.3 Coding potential programs indicate that *CEROX1* may encode a poorly conserved protein

CEROX1 is annotated as a lncRNA in GenBank and Ensembl, but six bioinformatic analyses of the coding potential of the *CEROX1* transcript provide different conclusions (Table 4.1). The original Coding Potential Calculator (CPC, cpc.cbi.pku.edu.cn)[54] classified *CEROX1* as non-coding, albeit "weakly". This is likely due to the lack of evolutionary conservation, since three of the six metrics taken into account are derived from a BLASTX search of any ORFs identified[54,55], and CPC therefore has a tendency to misclassify novel protein-coding transcripts as lncRNAs[56]. However, the most recently updated CPC2 (cpc2.cbi.pku.edu.cn)[57], which "exhibits superior accuracy… especially for long non-coding transcripts" now assigns a much higher likelihood of coding potential, at 86%. CPC2 retains the ORF length and integrity metrics from CPC, while adding two others: the Fickett score[58], which is a statistical analysis of the codon bias in light of nucleotide composition; and the isoelectric point (pI) of the putative protein, thereby incorporating metrics from the DNA, RNA, and protein level. The Coding Potential Assessment Tool (CPAT, lilab.research.bcm.edu/cpat)[59] is another program that utilises the Fickett score, in combination with a hexamer score that discriminates protein-coding likelihood based upon the characterised non-independent relationship between consecutive amino acid codons[60]. Like CPC2, CPAT indicates a high probability of coding potential in *CEROX1* (76%). Neither incorporates a measure of evolutionary conservation, supporting the notion that the lack thereof is the main barrier to *CEROX1* being a protein-coding gene: CEROX1-ORF1 itself appears to satisfy conditions of a true protein-coding ORF.

Tool	CODING or NON-CODING	Coding potential probability, if given
Coding Potential Calculator	"Weakly" NON-CODING	-
Coding Potential Calculator 2	CODING	86.2%
Coding Potential Assessment Tool	CODING	75.9%
PORTRAIT	NON-CODING	4.8%
iSeeRNA	CODING	-
RNAcon	CODING	-

Table 4.1: Estimations of the coding potential of *CEROX1* as produced by coding potential programs. The sequence of the *CEROX1* isoform ENST00000562570, which includes the three largest ORFs, was submitted to six web-based prediction servers: Coding Potential Calculator (CPC, cpc.cbi.pku.edu.cn)[54]; CPC2 (cpc2.cbi.pku.edu.cn)[57]; Coding Potential Assessment Tool (CPAT, lilab.research.bcm.edu/cpat)[59]; PORTRAIT (bioinformatics.org/portrait)[61]; iSeeRNA (sunlab.cpy.cuhk.edu.hk/iSeeRNA)[56], and RNAcon (crdd.osdd.net/raghava/rnacon)[62]. Two-thirds of the programs suggested that *CEROX1* is in fact a protein-coding gene.

PORTRAIT (bioinformatics.org/portrait)[61] is another algorithm that uses metrics derived *ab initio* from the sequence of interest and not conservation or homology. While this program indicates a highly contrasting result to CPC2 and CPAT (5% chance of protein-coding potential), it was specifically designed with a view to identifying ORFs in poorly characterised genomes such as fungi, by tolerating errors and small sequencing fragments[61].

The final two tools whose algorithms suggested that *CEROX1* is a coding RNA were iSeeRNA (sunlab.cpy.cuhk.edu.hk/iSeeRNA)[56] and RNAcon (crdd.osdd.net/raghava/rnacon)[62]. Both use nucleotide composition metrics. iSeeRNA also accounts for overall conservation using PhyloCSF and ORF quality scores similar to those of CPC2 and CPAT.

4.2.3 CEROX1-ORF1 is exceptional for a lncRNA-ORF

Despite the fact that *CEROX1* fits the specification of a typical lncRNA in terms of its relatively short length and low number of introns[63,64], CEROX1-ORF1 defies many of the criteria identified on a genome-wide scale that distinguish between protein-coding ORFs and non-coding ORFs[65]: as noted above, CEROX1-ORF1 is extremely long for a 'non-coding' ORF, falling within the range of protein-coding ORFs as opposed to the longest ORFs in most lncRNAs, the size distribution of which is closer to that of other non-primary ORFs in mRNAs[65]. Likewise, the fact that CEROX1-ORF1 begins with the first occasion of an AUG codon within the sequence is also reminiscent of protein-coding ORFs, not lncRNA-ORFs, which are often many within a transcript. This would contribute to a more positive translational context of CEROX1-ORF1, since it is believed that the association of a ribosome with numerous upstream ORFs (uORFs) can lead to abortive translation events, reduction in downstream translation efficiency, and mRNA destabilisation[65]. The human microprotein NoBody is also the first 5' ORF within a highly expressed transcript, and the existence of two introns after the ORF does not impede translation, showing that it is not degraded by NMD[36].

Differential nucleotide composition, including GC content, and codon bias are also observed between protein-coding mRNAs and lncRNAs[65], but the inclusion of nucleotide composition metrics and the Fickett score in the above prediction algorithms that identified coding potential in CEROX1-ORF1 suggests that this ORF is not lncRNA-like in these respects. Indeed, the GC content of *CEROX1*, as mentioned in Chapter 3, is extremely high for a lncRNA.

4.2.4 Experimental approaches to determining protein-coding capacity *in vitro* and *in vivo*

Ultimately, these tools are trained with current knowledge, and can only model and predict based upon limited parameters. As such, several reviews of coding sequence identification have noted the importance of supplementing *in silico* data, however well-informed, with comprehensive experimental studies *in vitro* and *in vivo*[23,66]. Not least due to the fact that, as shown in the first section of this introduction, there are exceptions to most if not all of the rules predicting the features and behaviour of protein-coding transcripts.

Considering current mass spectrometry-derived proteomics databases[67,68], including catalogues of smORFs[10], there is no evidence of any *CEROX1* polypeptides. However, of the 20,377 proteins annotated in the human proteome (neXtProt[69], November 2018), 2,243 are "missing proteins"[70], being supported by only limited or ambiguous proteomics evidence, and over 500 are considered merely hypothetical[71,72]. These data clearly indicate that unambiguous identification of translated polypeptides by wide-scale high-throughput approaches is not guaranteed with the current state of the art. And, thereby, that interpreting an absence of existing high-stringency proteomic evidence as a protein simply not existing is a foregone conclusion. Aside from the current 'presumed' and 'hypothetical' proteins, there are many factors that might mask experimental discovery of novel proteins outside traditional criteria. Peptides derived from low abundance proteins within very complex mixtures such as whole cell lysates are vulnerable to undersampling during spectrometry[73]. Instability and rapid degradation, and extensive post-translational modification can all physically impede detection[41,73], and the possibility of highly specific spatiotemporal translation[71] increases the difficulty exponentially.

4.2.4.1 Ribosome profiling data suggest a possible tissue or cell type-specific translation environment

Proteomic approaches rely upon detection of the proteins themselves, but it is possible to infer information about the physical translation process based upon the recruitment of translation machinery to a target transcript. The human ribosome comprises 80 proteins and 4 ribosomal RNAs arranged into two major parts, the large 60S subunit and the small 40S subunit[74]. Following recruitment of the 40S subunit by the 5' cap, this subunits scans towards the 3' end of the RNA until a start codon (AUG) is found. Binding of the associated initiator methionine tRNA ($tRNA_i^{Met}$) to the start codon and subsequent GTP hydrolysis leads to the binding of the 60S subunit and the formation of the full 80S ribosome initiation complex. The binding of ribosomes to transcripts may be analysed by two parallel techniques: polysome fractionation and ribosome

footprinting or profiling. Polysome fractionation exploits the association between effective translation and tandem ribosome binding, whereby transcripts bound by multiple ribosomes are considered more likely to be truly coding. Sucrose gradient fractionation of cellular lysates, in order to separate unbound RNA, monosomal ribosome-bound RNA, and polysomal RNA associated with six to ten ribosomes, is followed by RT-qPCR or RNA-seq. Ribosome profiling is a higher resolution and more accurate mapping strategy[75], in which ribosomal-bound and therefore protected RNA[76] is trimmed by digestion with RNase, and isolated. RNA-seq of the resulting footprints reveals ribosome binding sites at near-single nucleotide resolution[77].

Contrary to initial expectations, it was found that many lncRNAs associate with cytosolic ribosomes[33], including the polyribosome fraction[78]. Furthermore, the degree of association (translation efficiency or ribosome coverage) between a transcript and the ribosome population is insufficient to distinguish between non-coding and protein-coding RNAs, or indeed the 5'-UTRs of protein-coding genes. Indeed, early studies identified some classical non-coding RNAs, such as telomerase RNA, and spliceosome RNAs as highly represented in the poly-ribosome fraction and densely populated with ribosomes[79]. However, this is largely a background effect caused by the association of these RNAs in other RNA-protein (ribonucleoprotein, RNP) complexes. These non-ribosome-bound, yet RNase protected, fragments may be filtered out of datasets based upon the characteristic size of ribosome footprints[41,79].

The metric that was found to best distinguish protein-coding capacity is quantified by the ribosome release score, a measure of the tendency of the RNA-ribosome complex to dissociate following a stop codon[79]. In protein-coding mRNAs, ribosomes are released rapidly following the incorporation of the C-terminal amino acid, at the stop codon. This results in a much lower fold density of ribosome binding in the 3'-UTR (over 100-fold lower, on average). In classical lncRNAs, however, the relative binding of ribosomes to the ORF and its respective 3'-UTR is much less dependent upon position, with <2-fold differences on either side of the stop codon. A more recently developed indicator of ribosome profile veracity is ribosome position periodicity: it is expected that coding regions will show a strong bias towards ribosome-protected fragments that align to a single reading frame consistent with the associated ORF[10,77].

CEROX1 fragments do appear in a number of ribosome profiling datasets (HEK293T cells[80], data not shown; human brain[81], Figure 4.3). As explained above, this is not direct evidence for translation or protein-coding activity. However, it is notable that, within samples of human brain tissue, the great majority of ribosome profiling reads are situated within the three ORFs identified previously (Figure 4.1; Figure 4.3A-B). Although

the lack of coverage within the 3'-UTRs of the respective ORFs is possibly exacerbated by the modest total number of footprint sequences, there is a visually apparent enrichment of reads within the ORFs (which raises the ribosome release score). The exception to this is the 3' end of CEROX1-ORF1, which appears to demonstrate some read-through to a second stop codon.

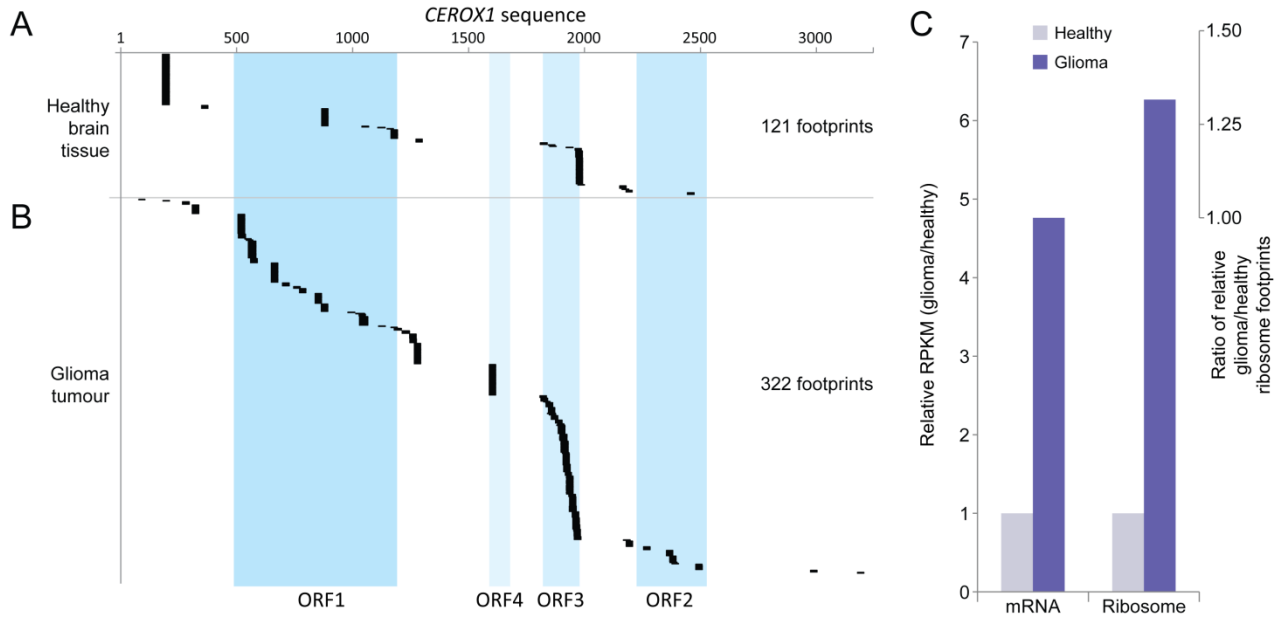


Figure 4.3: *CEROX1* is bound by ribosomes. Ribosome profiling data[81] reveal that translational machinery binds to *CEROX1* in both healthy brain tissue **(A)** and glioma tumour tissue **(B)**. Particularly in the latter, ribosome footprints are concentrated in the three ORFs identified above (ORF1, ORF2, and ORF3), as well as the fourth largest CEROX1-ORF, ORF4. **(C)** Overall expression of *CEROX1* at the mRNA level is higher in glioma than healthy tissue, by 4.8-fold, while the number of ribo-seq reads is 6.3-fold higher. The relative enrichment of ribosome reads normalised to mRNA expression is therefore 1.3-fold.

Additionally, the density of ribosome footprints is altered in the glioma tumour samples as opposed to healthy brain tissue (Figure 4.3C). Expression of *CEROX1* at the RNA level is 4.8-fold higher in the tumour, but the number of ribosome footprints is 6.3-fold higher, resulting in a relative enrichment of ribosomal binding of 1.3 fold. If this single experiment is representative, which cannot be concluded here, this suggests that *CEROX1* translation, if truly present, may be dysregulated in disease states.

While the study did not utilise ribosome immunoprecipitation to minimise the RNase-protection footprinting effect of other RNA-RNA or RNA-protein complexes, the size distribution of protected *CEROX1* fragments is similar to that of the classical protein-coding mRNAs *SOX8* and *TLCD1*, peaking at 28 nucleotides[77], while the distribution of fragments from the RNaseP-associated RNA *RPPH1* is notably different (Figure 4.4).

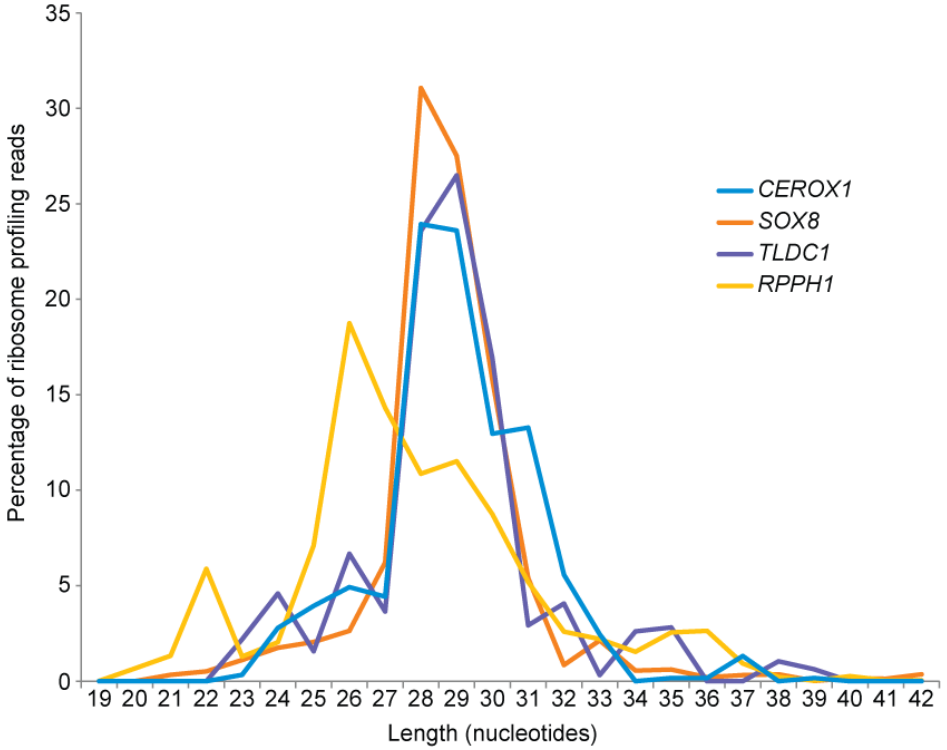


Figure 4.4: The distribution of length in RNase-protected reads following ribosome profiling of human brain tissue[81] indicates that reads ascribed to *CEROX1* are indeed associated with ribosomes. Unlike that of the classical non-coding RNaseP-associated RNA RPPH1 (yellow), the distribution of fragment sizes for *CEROX1* (blue) is more similar to the protein-coding genes *SOX8* (orange) and *TLDC1* (purple), peaking at ~28 nt.

Interestingly, despite lacking the large ORF investigated here, TRAP analysis of mouse CNS cell types[37] reveals that mouse *Cerox1* strongly associates with the ribosome in some cell types, being enriched by 15- to 40-fold ($p < 0.001$) in the ribosome-bound fraction in cortical and cerebellar oligodendrocytes. A smaller level of enrichment (2.5- to 7-fold; $p < 0.0003$) is observed in astrocytes and Bergmann glia, and a 4-fold depletion in cortical nociceptive neurons ($p < 0.003$). No changes were observed in other cell types[37]. This suggests that any *CEROX1* association with, or regulation by, the ribosome, might be highly cell type-specific.

4.2.5 CEROX1-ORF1 is a candidate for further investigation

The preliminary data presented here demonstrate that CEROX1-ORF1 is a barrier to the classification of *CEROX1* as a classical lncRNA, from the perspective of both bioinformatic and simple ribosome profiling analyses. While none of the data confirm protein-coding capacity, there is sufficient evidence to question the assumption that *CEROX1* is simply a lncRNA. Further experiments were therefore designed to investigate CEROX1-ORFs *in vitro* and *in vivo*, with the aim of identifying whether any is expressed as a protein in cells or human tissue.

4.3 A strategy was devised to investigate the putative CEROX1-ORFs in vitro

As shown in Figure 4.1, only one *CEROX1* isoform transcript, ENST00000562570.1 (RefSeq BC098409), contains the full sequence of all three of the largest potential CEROX1-ORFs. As such, this transcript was chosen as the simplest for investigating the possible translation of the three larger CEROX1-ORFs *in vitro*. While the above discussion of small ORFs and microproteins illustrates that the limit imposed in this study of 50 amino acids for proteins of interest is somewhat arbitrary, it was also informed by the experimental limits in reproducibly detecting very small proteins.

Potential approaches for this study included inserting the putative ORF sequences upstream of a fluorescent reporter protein expression sequence; inserting a fluorescent protein sequence downstream of the ORF sequences within the *CEROX1* overexpression construct; or adding a small epitope tag at either end of the ORF sequences within the *CEROX1* overexpression construct. Ultimately, the third strategy was implemented, tagging the ORFs at their 3' C-termini, for the reason of attempting to maintain the ORFs' putative N-terminal start codons in the most physiologically relevant context possible, and minimising disruption to the overall *CEROX1* sequence.

Expression constructs were generated in which the 3' sequences of the three *CEROX1*-ORFs were mutated so as to insert a detectable epitope tag prior to the STOP codon. The C-terminal ends of potential polypeptides were thereby tagged with a human influenza haemagglutinin (HA) tag (YPYDVPDYA). The EYFP expression control vector was equivalently tagged and used as a known translated protein positive control. These constructs were then transiently transfected into HEK293T cells, and overexpression of the mutant *CEROX1* transcripts was confirmed by RT-qPCR. Transfected cell lysates were then analysed by western blotting using an anti-HA antibody.

4.3.1 Detection of very small putative proteins requires additional considerations in detection by western blotting

The two smaller CEROX1-ORFs investigated (CEROX1-ORF2 and CEROX1-ORF3) are very small, by the standards of most physiological proteins, being predicted to have molecular weights of 10.5 kDa and 5.5 kDa, respectively. In most PAGE and blotting systems, such small putative polypeptides would typically be lost from the end of the gel, or during transfer. Conditions were therefore optimised for small protein

retention and detection, and success was illustrated by co-detection of a small control protein: the polypeptide hormone insulin. A sample of protein from the rat pancreatic β-cell line INS-1 was run alongside CEROX1-ORF samples, and probed with an anti-insulin antibody. A clear band was detected at around 6 kDa, representing the 5.8 kDa insulin protein, as well as a fainter 10 kDa band representing pro-insulin (Appendix A8). From the results of these experiments, it was thus concluded: a) that small proteins were detectable; and b) that any observed absence of a protein in this western blotting system, even in the single kDa range, was therefore consistent with its being absent from the sample.

4.4 At least two ORFs in *CEROX1* are subject to translation *in vitro*

4.4.1 *CEROX1*-ORF1 induces translation of polypeptides when overexpressed in HEK293T cells

Overexpression of C-terminally HA-tagged CEROX1-ORF1, hereon denoted as CEROX1-ORF1$^{C\text{-}HA}$, in HEK293T cells resulted in the detection by western blotting of a number of bands not present in the non-tagged *CEROX1* control sample (Figure 4.5). The primary band indicated a protein with an apparent molecular weight of around 30 kDa, potentially consistent with the *in silico* translated CEROX1-ORF1$^{C\text{-}HA}$ sequence predicted size (25.3 kDa). This is suggestive of translation of CEROX1-ORF1 *in vitro*. Expression of neither CEROX1-ORF2$^{C\text{-}HA}$ nor CEROX1-ORF3$^{C\text{-}HA}$ seemed to induce translation of any detectable products.

It is unclear why there are multiple higher molecular weight bands present in the *CEROX1*-ORF1$^{C\text{-}HA}$ lane of the anti-HA western blot, in the 35 to 80 kDa range (Figure 4.5). While the highly denaturing conditions of the experimental protocol (cells were lysed in radioimmunoprecipitation assay (RIPA) buffer containing sodium dodecyl sulphate (SDS), sodium deoxycholate, and IGEPAL® CA-360; samples were denatured and loaded in lithium dodecyl sulphate (LDS) buffer; and electrophoresis was conducted in SDS running buffer) ought to result in denaturation and dissociation of any protein complexes, it may be that the protein is involved in a strong heteromeric protein-protein interaction(s). Alternatively, it may be subject to a post-translational modification(s), such as glycosylation. The smaller bands, on the other hand, might represent cleavage products.

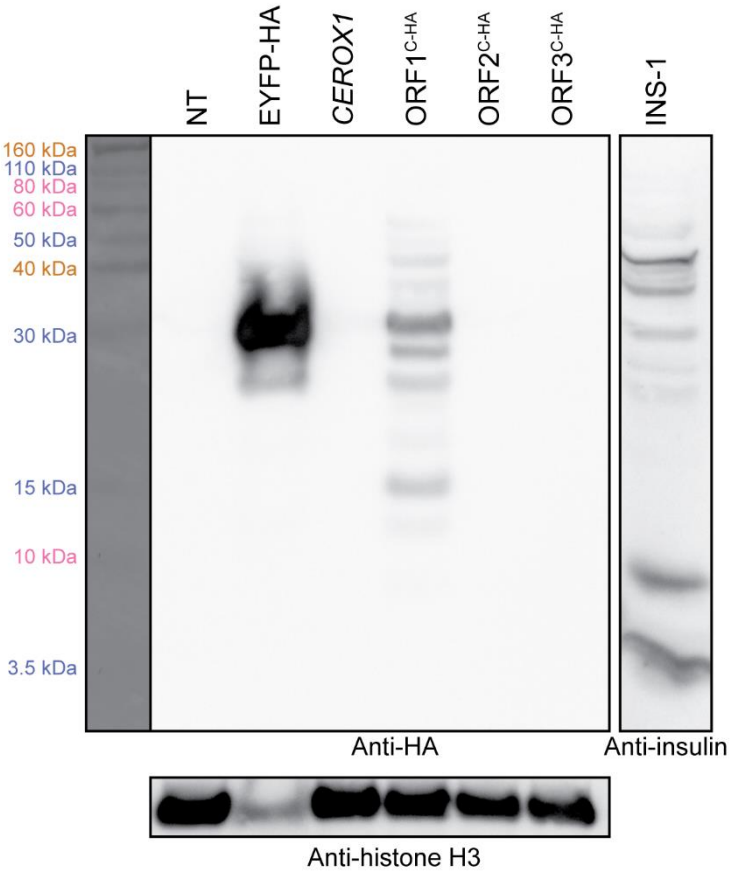


Figure 4.5: CEROX1-ORF1 may be translated *in vitro*. Lysates from HEK293T cells overexpressing HA-tagged EYFP and *CEROX1* constructs were analysed by western blotting. 40 µg of protein was loaded into each lane (except EYFP$^{C\text{-}HA}$, for which only 5 µg was loaded due to its extraordinary expression), and then probed with an anti-HA antibody. While the non-transfected control, non-tagged EYFP, and non-tagged *CEROX1* lanes were blank, the positive control EYFP$^{C\text{-}HA}$, and *CEROX1*-ORF1$^{C\text{-}HA}$ lanes indicated translation of products. Histone H3 (observed band size 17 kDa) was used as a loading control, indicating equal protein loading across the samples, except for the EYFP$^{C\text{-}HA}$ sample. Blotting of protein from the rat pancreatic β-cell line INS-1, electrophoresed on the same gel and transferred to the same membrane, with anti-insulin indicated that any proteins smaller than those detected – including putative *CEROX1*-ORF2$^{C\text{-}HA}$ and *CEROX1*-ORF3$^{C\text{-}HA}$ – were not obscured by the experimental conditions. NT – non-transfected.

4.4.2 Translation of CEROX1-ORF2, but not CEROX1-ORF3, is detectable under degradation-inhibiting conditions

4.4.2.1 A combination of drug treatments was used to inhibit cellular pathways of degradation

It is possible that these two ORFs are also translated, but at a much lower level than CEROX1-ORF1. Alternatively, if products of CEROX1-ORF2 or CEROX1-ORF3 were translated but immediately targeted for degradation, their steady-state concentration may be undetectable using the previous approach. In order to address these possibilities, transfected cells were treated with two drugs that interfere with pathways of protein degradation (MG132 and bafilomycin A1) and then subjected to anti-HA immunoprecipitation.

The ubiquitin-proteasome pathway is one of two major pathways of protein degradation in the cell, the other being autophagic lysosomal proteolysis. During the former, attachment of the small protein ubiquitin to proteins, including damaged, misfolded, or oxidised proteins as well as those whose concentration undergoes a necessitated temporal concentration, such as cyclin-dependent kinases, targets them to the proteasome, by which they are degraded. MG132 is a potent inhibitor of the proteasome[82,83]. Treatment of cells with MG132 halts this process, inducing a build-up of ubiquitinated proteins targeted for degradation.

Bafilomycin A1, on the other hand, is an inhibitor of vacuolar H^+-ATPase (V-ATPase) and thereby impedes lysosome acidification[84]; through this lysosome alkalisation as well as a pH-independent effect[85,86], bafilomycin A1 prevents the fusion of autophagosomes with lysosomes to form autolysosomes[87,88]. This results in accumulation of autophagosomes and their degradation-targeted cargo.

If the hypothetical CEROX1-ORF2 or CEROX1-ORF3 polypeptides were to be translated and immediately degraded by either the ubiquitin-proteasome or autophagic pathways, treatment of cells with MG132 or bafilomycin A1, respectively, would result in their accumulation and subsequently increased detectability. Cells were grown for 48 hours following transfection, and then treated for six hours prior to rapid lysis and protein extraction. To control for the success of the drugs in eliciting the desired effects, the cell lysates were first probed for ubiquitin and LC3. Ubiquitin is conjugated to proteins by a myriad of ligase enzymes, and can form long chains that extend a target's molecular weight considerably. Blotting of ubiquitin following MG132 treatment demonstrates this, with a ladder or smear of darker bands extending into the hundreds of kDa, representing proteins targeted for proteasome degradation (Figure 4.6A).

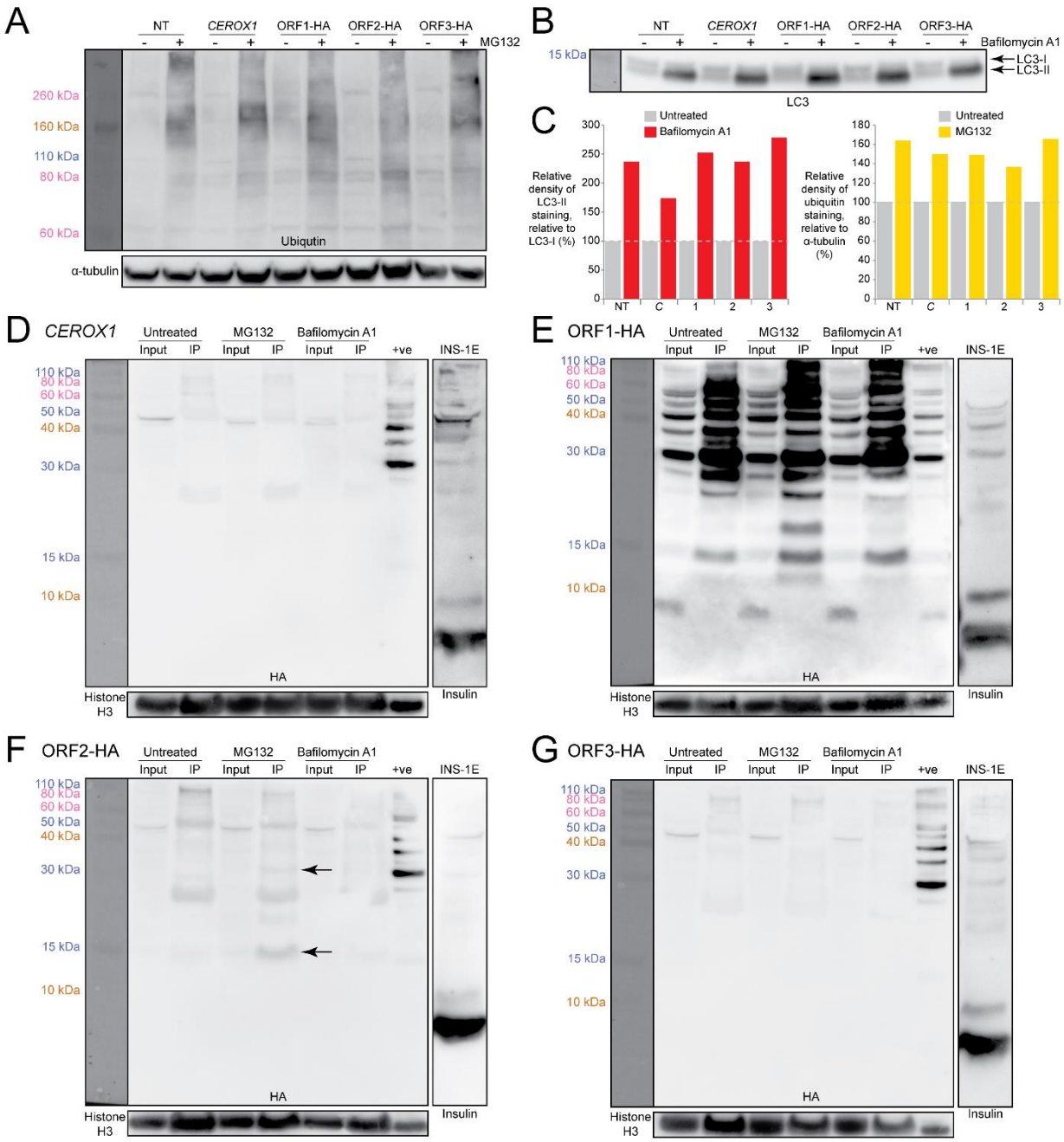


Figure 4.6: Expression of CEROX1-ORF protein products following inhibition of cellular degradation pathways. Following transfection with HA-tagged *CEROX1* constructs, HEK293T cells were treated with MG132 (50 μM) or bafilomycin A1 (100 nM) for six hours, to inhibit the proteasome and lysosome pathways of degradation, respectively. Blotting of **(A)** ubiquitin and **(B)** LC3-II indicated successful inhibition of these pathways, with each accumulating following MG132 or bafilomycin A1 treatment, respectively. **(C)** Densitometry indicated that the effects of the drugs occurred independently of transfection, with non-transfected cells (NT) showing similar levels of inhibition markers to those transfected with untagged *CEROX1* (*C*), or tagged CEROX1-ORF constructs (1, 2, 3). Cell lysates (1 mg total protein) were then subjected to anti-HA immunoprecipitation and the eluates analysed by western blotting, alongside input lysates. Histone H3 was used as a loading control, indicating equal protein loading across the samples, and a lane of CEROX1-ORF1$^{C\text{-}HA}$ transfected protein was used as a positive control for HA blotting on each membrane (+ve). **(D)** Untagged *CEROX1* negative control samples indicated no bands, and the previously observed bands were present in the CEROX1-ORF1$^{C\text{-}HA}$ samples **(E)**, with the immunoprecipitation protocol enriching the key bands. Two unique bands were observed in CEROX1-ORF2$^{C\text{-}HA}$ transfected cells **(F)**, but only after exposure to MG132 (arrows). No CEROX1-ORF3$^{C\text{-}HA}$ proteins were observed under any conditions **(G)**. Successful blotting of the 6 kDa insulin protein within a sample of INS-1 lysate on the same membranes indicates that detection of any putative proteins in the 3.5-10 kDa range was not precluded by the experimental conditions **(D-G)**. NT – non-transfected; IP – anti-HA immunoprecipitation; +ve – CEROX1-ORF1$^{C\text{-}HA}$ lysate was included on each blot as a positive control for the anti-HA antibody.

The protein LC3 is found in two forms: LC3-I is a soluble cytosolic protein, which is converted to a lipidated form LC3-II by conjugation to phosphatidylethanolamine[89], whereby it is recruited to the membrane of autophagosomes. Following fusion with a lysosome, the LC3-II is degraded along with the contents of the organelle[89]. Under bafilomycin A1 treatment, and subsequent inhibition of lysosome activity, autophagosomes, and so levels of associated LC3-II, accumulate (Figure 4.6B). The level of inhibition, as shown by the accumulation of these degradation markers, was independent of the sample (Figure 4.6C).

4.4.2.2 *CEROX1*-ORF2 is lowly expressed or subject to rapid degradation

As expected, transfection of non-HA-tagged *CEROX1* resulted in the detection of no polypeptides by anti-HA blotting (Figure 4.6D). CEROX1-ORF1$^{C\text{-}HA}$ was clearly expressed as before, this being unchanged by exposure to either MG132 or bafilomycin A1 (Figure 4.6E). Indeed, CEROX1-ORF1$^{C\text{-}HA}$ samples were used as positive controls for anti-HA blotting in each experiment. Under untreated conditions, no polypeptides were detected for CEROX1-ORF2$^{C\text{-}HA}$. However, following treatment with MG132, but not bafilomycin A, a faint band was visible at around 15 kDa (Figure 4.6F). This band was enhanced by anti-HA immunoprecipitation, which also revealed a second band at about 30 kDa. Given the predicted molecular weight for ORF2 of 10.5 kDa (plus an additional 1 kDa for the HA tag), the 15 kDa band possibly represents a translated protein. If so, it is presumably subject to rapid turnover via the ubiquitin-proteasome pathway, which prevents accumulation to detectable levels in the absence of MG132.

4.4.2.3 *CEROX1*-ORF3 remains undetected and is likely not translated

In spite of the inhibition of degradation through the proteasome and autophagy pathways, no HA signal was detected following transfection of CEROX1-ORF3$^{C\text{-}HA}$ (Figure 4.6G). As before, INS-1 samples run on the same gels and transferred onto the same membranes were used to blot for insulin, demonstrating the ability to detect very small proteins (Figure 4.6D-G). Absence of any polypeptides even under these degradation-inhibiting conditions leads to the conclusion that CEROX1-ORF3 is not translated, in this system at least, and it was thus excluded from further investigation here. This does not necessarily preclude its being translated in other systems, however.

4.4.2.4 Mass spectrometry confirms the translation of ORF1 *in vitro*

In order to verify the identity of the putative CEROX1-ORF1$^{C\text{-}HA}$ bands visible on these blots, cell lysates were subjected to anti-HA immunoprecipitation and subsequent SDS-PAGE as before (Figure 4.7A). The most prominent bands, being the 30 kDa primary band and the fainter band at around 27 kDa for ORF1, were then analysed by mass spectrometry (Advanced Proteomics Service, University of Oxford). Interestingly, this confirmed that both bands represented species derived from CEROX1-ORF1$^{C\text{-}HA}$ (Figure 4.7B). Band 1 exhibited 84% coverage from 14 individual peptides, while the lower density of band 2 resulted in lower but substantial coverage (56% from 11 peptides). In light of this, it is possible that the multiplicity of CEROX1-ORF1$^{C\text{-}HA}$ species results from a cleavage *in vitro*, generating the smaller of the two species probed here, which lacks coverage of over 50 residues at the N-terminus. These data demonstrate unequivocally that CEROX1-ORF1 is translated *in vitro* following overexpression.

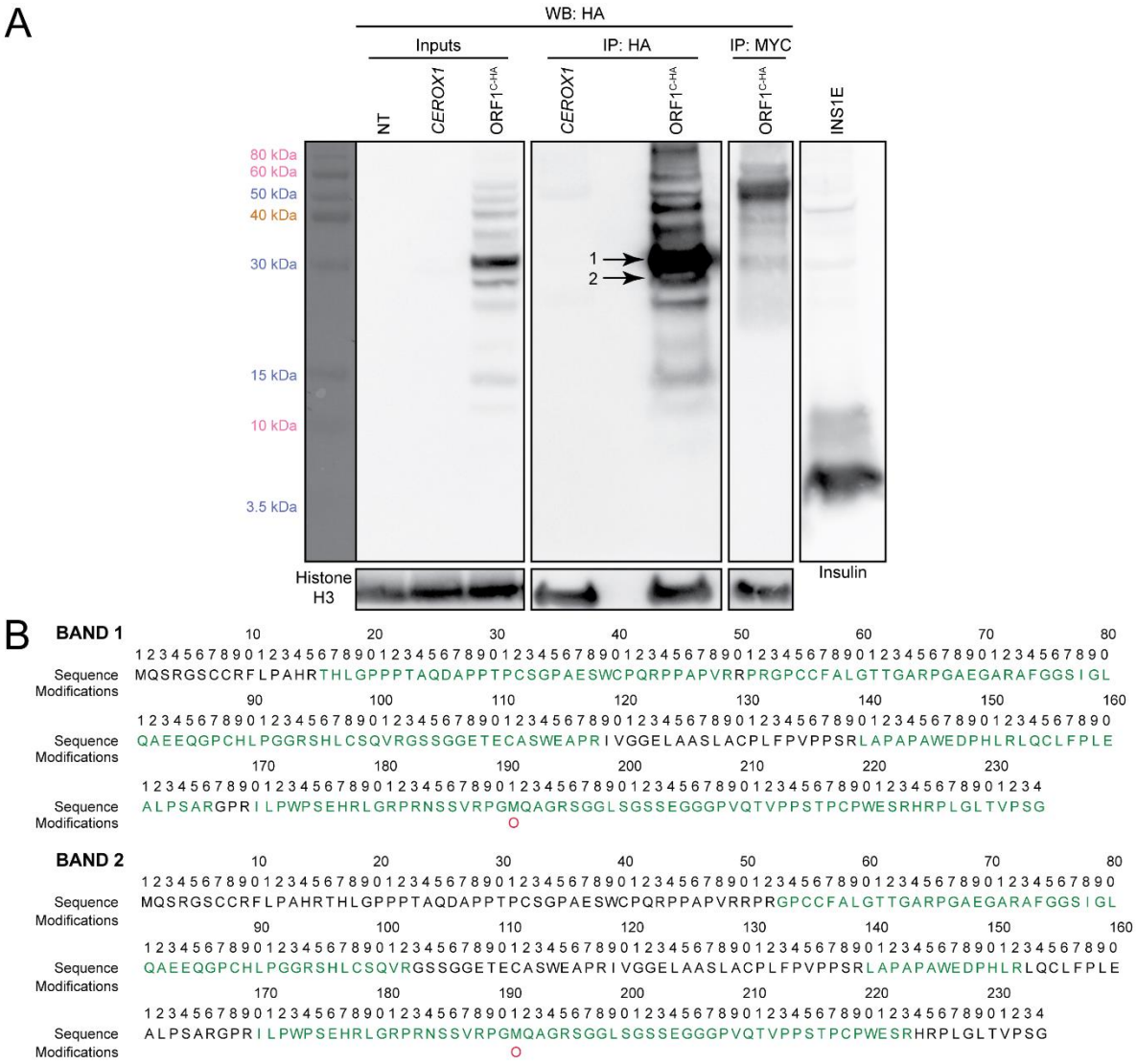


Figure 4.7: Mass spectrometry sequencing of CEROX1-ORF1$^{C\text{-}HA}$ translation products. Band 1 is the primary band of 30 kDa; band 2 is the secondary band of 27 kDa. Green letters indicate alignment of identified fragments with the 234 amino acid CEROX1-ORF1 sequence. In red are modifications to amino acids identified within the stretches of polypeptide identified, namely a single oxidised methionine (M191) in each sequence.

4.4.3 Expression of CEROX1-ORF1 is visible by immunocytochemistry following overexpression

Given the convergence of function of the protein-coding and non-coding functions of cncRNAs such as *SRA*/SRAP and Irs1, and the putative oxidative phosphorylation-regulating function of *CEROX1*, it might be expected that CEROX1-ORF1 would be associated with mitochondria. Submission of the protein sequence to the localisation prediction program TargetP (cbs.dtu.dk/services/TargetP)[90,91] in fact results in an albeit low "reliability class" prediction of mitochondrial localisation. In order to investigate this *in vitro*, CEROX1-ORF1$^{C\text{-}HA}$-overexpressing HEK293T cells were probed by fluorescent immunocytochemistry (ICC) using the same anti-HA antibody used previously. However, this experiment indicated a broadly cytoplasmic localisation profile for the exogenous protein, with no overt enrichment in a pattern resembling that of any particular organelle, including mitochondria (Figure 4.8).

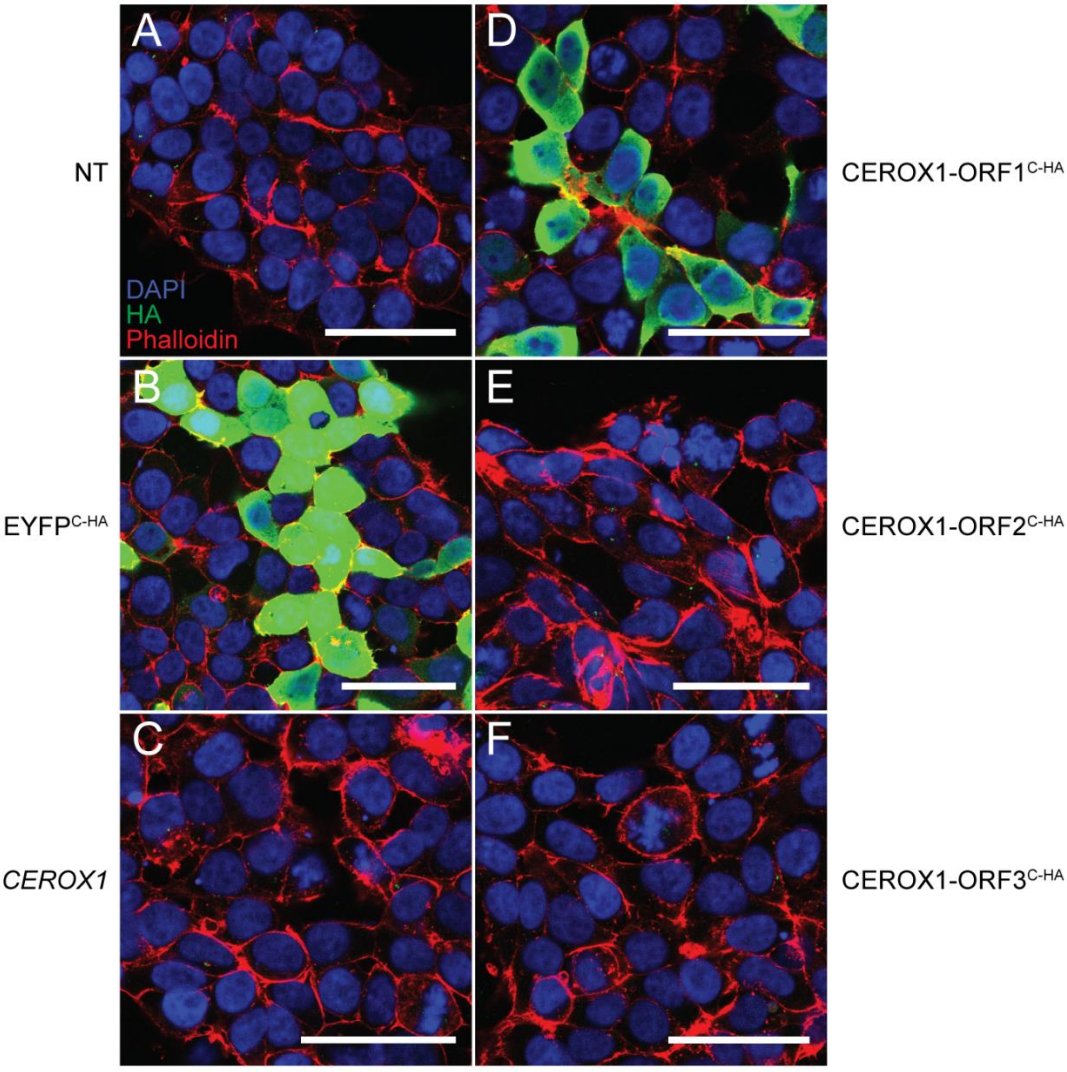


Figure 4.8: Expression of CEROX1-ORF1$^{C\text{-}HA}$ protein in HEK293T cells. Cells overexpressing CEROX1-ORF1$^{C\text{-}HA}$ were visualised by immunocytochemistry with an anti-HA antibody (red) and stained with DAPI (blue) to mark nuclei and phalloidin (red) to mark the actin cytoskeleton at the cell membrane. HA signal was observed in neither the non-transfected (NT) cells, nor those transfected with untagged CEROX1, CEROX1-ORF2$^{C\text{-}HA}$, or CEROX1-ORF3$^{C\text{-}HA}$. Scale bars represent 50 μm.

4.5 Does *CEROX1* encode a protein endogenously *in vitro* and *in vivo*?

It is unknown whether the exogenous CEROX1-ORF1 protein seen here *in vitro* is expressed endogenously, since the overexpression of *CEROX1* in transfected HEK293T cells is extremely high, resulting in supra-physiological concentrations of the transcript. Conversely, there may be ORFs within *CEROX1* that demonstrate translation *in vivo* but not in these *in vitro* systems. As mentioned previously, there is no evidence of the CEROX1-ORF1 polypeptide, nor any other *CEROX1* polypeptides, in existing proteomics databases[67,68], including catalogues of smORFs[10].

4.5.1 CEROX1-ORF1 translation was not detected endogenously *in vitro*

A custom antibody to the CEROX1-ORF1 protein, was successfully derived and shown to be capable of detecting both CEROX1-ORF1$^{C\text{-}HA}$ and non-tagged CEROX1-ORF by western blotting (Figure 4.9A) and ICC (Figure 4.9B). Mutation of the CEROX1 or CEROX1-ORF1$^{C\text{-}HA}$ start codon (^{ATGmut}ORF1 and ^{ATGmut}ORF1$^{C\text{-}HA}$) was shown to ablate the major band in western blotting (Figure 4.9A), and ICC (Appendix A9), confirming the target of the antibody. Intriguingly, other bands remain in the ATG-mutant lanes, and fluorescence is present in ATG-mutant cells (Appendix A9), suggesting the possibility of additional cryptic start sites.

Non-transfected cell lysates were then probed in search of an endogenously expressed protein. It has previously been observed that HEK293T cells, which express low levels of *CEROX1* endogenously (RT-qPCR, data not shown), are capable of translating CEROX1-ORF1 when overexpressed exogenously (Figures 4.5-8). While western blotting of non-transfected HEK293T cell lysates revealed no obviously isolated band indicative of CEROX1-ORF1 expression, a number of bands were detected across a full range of sizes, including several in the 25-30 kDa range expected of CEROX1-ORF1 (Figure 4.9C). In order to account for low abundance owing to the relatively low expression of *CEROX1* RNA in HEK293T cells, as well as to increase assay specificity, the custom antibody was used to immunoprecipitate putative CEROX1-ORF1 protein from non-transfected and *CEROX1*-overexpressing HEK293T cell lysates (Figure 4.9D). While the latter illustrates that the antibody may be successfully applied to immunoprecipitation, no difference was observed in the non-transfected immunoprecipitate from the antibody relative to that of non-immunised serum IgG. Additional ambiguity was provided, however, by the highly similar size of CEROX1-ORF1 to the small IgG band from the denatured IP antibody.

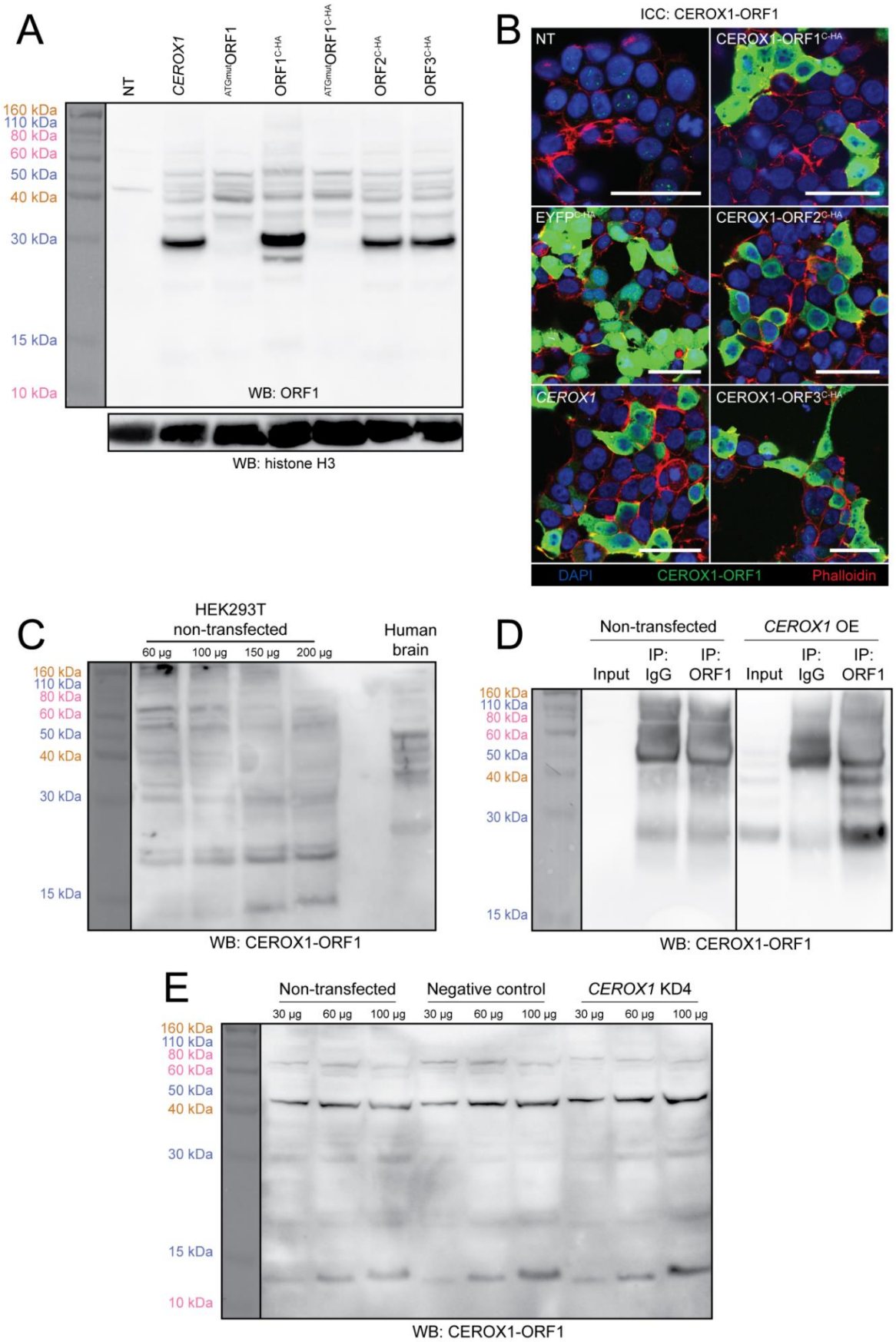


Figure 4.9: A custom rabbit antibody successfully detects exogenous, but not endogenous, CEROX1-ORF1 *in vitro*. (A) Western blotting and **(B)** immunocytochemistry in overexpressing cells illustrates detection of both tagged and untagged CEROX1-ORF1. This antibody was then used to probe HEK293T cell and whole human brain lysates by **(C)** western blotting and **(D)** immunoprecipitation, in search of endogenously expressed CEROX1-ORF1 protein. **(E)** Following identification of likely non-specific bands and possible ambiguity of the small IgG band from IP, knockdown of endogenous *CEROX1* RNA using GapmeR technology was used to validate the absence of endogenous CEROX1-ORF1. NT – non-transfected; IgG – non-immunised rabbit serum; ORF1 – CEROX1-ORF antibody.

The non-specificity of any detected bands to, and therefore absence of, CEROX1-ORF1 in HEK293T cells was confirmed further by the knockdown of *CEROX1* using a LNA™ GapmeR (Exiqon) (Figure 4.9E) (50% knockdown by RT-qPCR, data not shown). No alteration in detected bands was observed with knockdown of *CEROX1*. A range of loading quantities was implemented in order to pre-emptively account for non-linear detection.

In order to confirm that HEK293T cells are not unique in translation of CEROX1-ORF1 *in vitro*, it was shown that SK-N-SH neuroblastoma cells also generate CEROX1-ORF1 proteins when overexpressed exogenously (Figure 4.10).

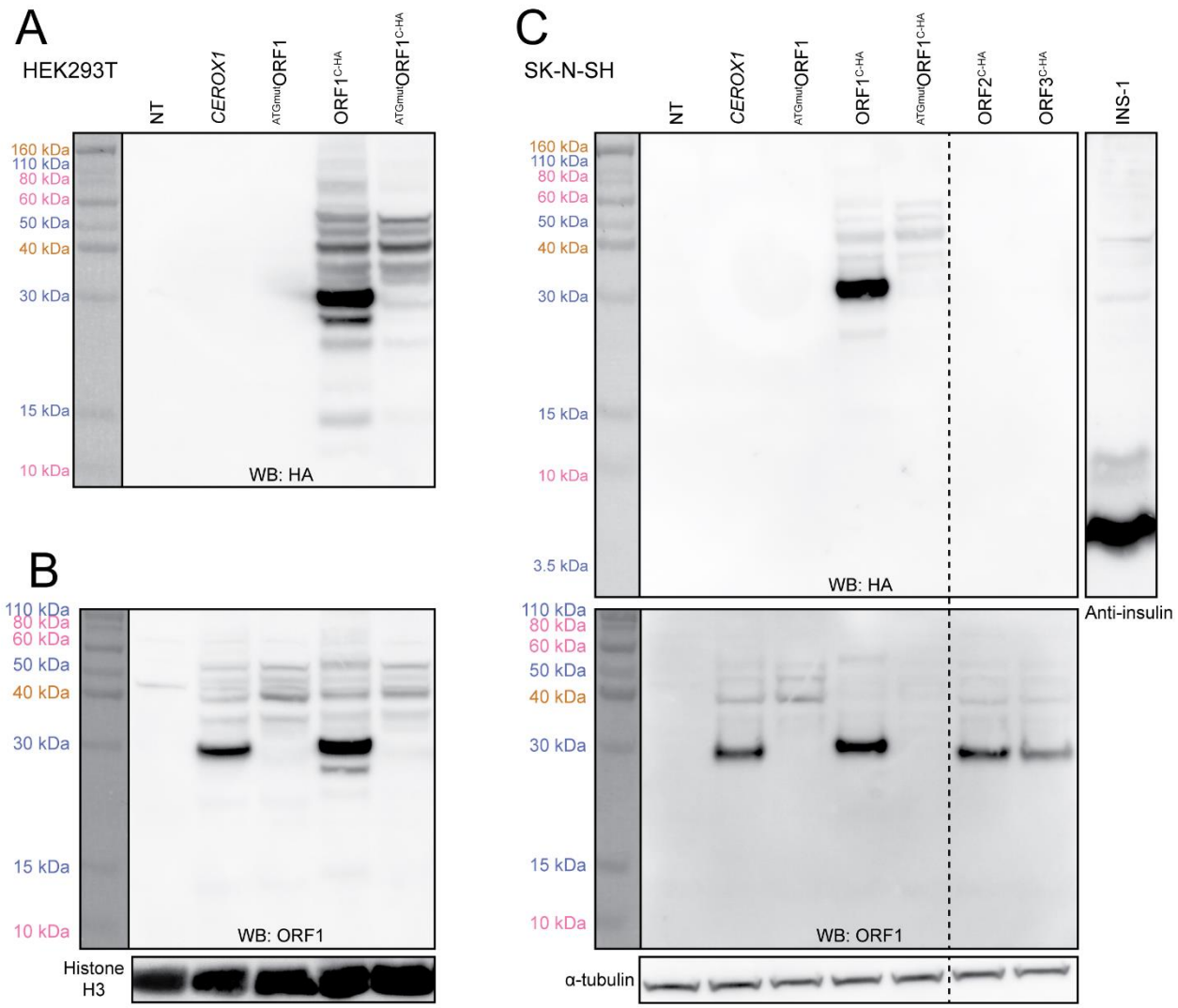


Figure 4.10: SK-N-SH neuroblastoma cells also translate exogenous CEROX1-ORF1 *in vitro*. As seen with HEK293T cells probed with anti-HA **(A)** and anti-CEROX1-ORF1 **(B)** antibodies, SK-N-SH cells were shown to produce ORF1 proteins when transfected with untagged or HA-tagged CEROX1-ORF1 **(C)**. Incidentally, neither CEROX1-ORF2 nor CEROX1-ORF3 was stably translated in SK-N-SH cells (right of dashed line). ATG mutants (^{ATGmut}ORF1 and ^{ATGmut}ORF1^{C-HA}) ablated expression of the major band in both HEK293T and SK-N-SH cells **(A-C)**.

4.5.2 Probing for CEROX1-ORF1 *in vivo*

Following the failure to detect expression of endogenous CEROX1-ORF1 protein in HEK293T cells, other human tissues were selected for sampling, based upon conclusions garnered from the expression pattern of *CEROX1* (Chapter 3). The brain is the tissue in which *CEROX1* is most highly expressed, and this pattern is conserved throughout the vertebrates in which *CEROX1* orthologues are expressed. Since protein abundance is proportional to RNA expression[92], it was hypothesised that any endogenous translation of a CEROX1-ORF resulting in a protein was most likely to take place in the brain. Although this does omit the possibility of spatial isolation of *CEROX1* RNA and a CEROX1-ORF protein product, the scope of this study meant that this consideration was outweighed by the likelihood of low protein abundance, and associated problems with detection[73], in a more lowly expressing tissue. As shown by western blotting, whole brain protein did yield several unconfirmed CEROX1-ORF1 bands (Figure 4.9C).

Within the brain, *CEROX1* is most highly expressed in the nucleus accumbens, the caudate nucleus, and the putamen. Together with the olfactory tubercle, these regions comprise the region of the basal ganglia known as the striatum (Figure 4.11A). Notably, the striatum was not represented in the most comprehensive human proteome project, which sampled only frontal cortex to represent the brain[68]. Meanwhile, the highest *in vivo* expression of *CEROX1* observed during the course of this ongoing *CEROX1* study was in a sample of proprietary oligodendroglioma RNA, in which *CEROX1* was expressed at 28-fold higher concentrations than in healthy brain, according to RT-qPCR (Figure 4.11B). Given the elevated expression of *CEROX1* in these tissues, they were chosen as the samples in which to investigate human CEROX1-ORF1 protein expression. Samples of healthy human striatum tissue and glioma tumour tissue were obtained *post mortem* from the Oxford Brain Bank (Dr Olaf Ansorge).

In order to address not only the expression of the putative CEROX1-ORF1 protein in humans, but also the potential evolutionary conservation of this molecule, the investigation was simultaneously extended to one of our closest living relatives, the gorilla. Chimpanzee, while considered even more evolutionarily similar to humans, has undergone a mutation at the *CEROX1* locus, which has frame shifted the C-terminus of the putative CEROX1-ORF1 and thereby precludes the detection of any putative CEROX1-ORF1 peptides with the existing custom antibody (Figure 4.2). The gorilla sequence, on the other hand, has been subject to several mutations throughout, but the N- and C-termini remain highly homologous, including the complete C-terminal epitope sequence (Figure 4.11C).

A fresh-frozen sample of cortex tissue from a western lowland gorilla (*Gorilla gorilla gorilla*) was obtained from the International Zoo Veterinary Group pathology laboratory in Keighley, UK (Dr Mark Stidworthy). The sample originated in Twycross Zoo, Warwickshire, UK, who kindly gave permission for a third-party material transfer for this study. In order to verify the integrity of the gorilla sample, aliquots of protein were extracted and subjected to SDS-PAGE and western blotting. Staining of the PVDF membrane with Ponceau S indicated discrete bands of protein, and probing with several antibodies against highly conserved targets such as cytochrome c yielded clean bands of correct sizes, indicating minimal degradation of the sample during ten years of storage (Appendix A10).

4.5.2.1 Immunoprecipitation-mass spectrometry indicates that CEROX1-ORF1 is stably translated in gorilla cortex, but not human striatum or glioma

Mass spectrometry is the gold standard for targeted protein identification. The high sensitivity and resolution afforded by this technique allows for thorough categorisation of the proteins within a refined sample, using *de novo* peptide sequencing and extensive database searches. This is especially important in validating the existence of a novel, potentially rare protein.

In order to maximise the chance of identifying CEROX1-ORF1 peptides, protein extracts derived from the three human and gorilla tissues were first subjected to immunoprecipitation using the custom CEROX1-ORF1 antibody (Figure 4.11D). Pulldown eluates, enriched for CEROX1-ORF1 if present, were then analysed by tandem mass spectrometry (LC MS/MS; Dr Roman Fischer and Dr Sarah Bonham, Target Discovery Institute, University of Oxford) (Figure 4.11E). All mass spectrometry data analyses were subject to a false discovery rate (FDR) threshold of 1%.

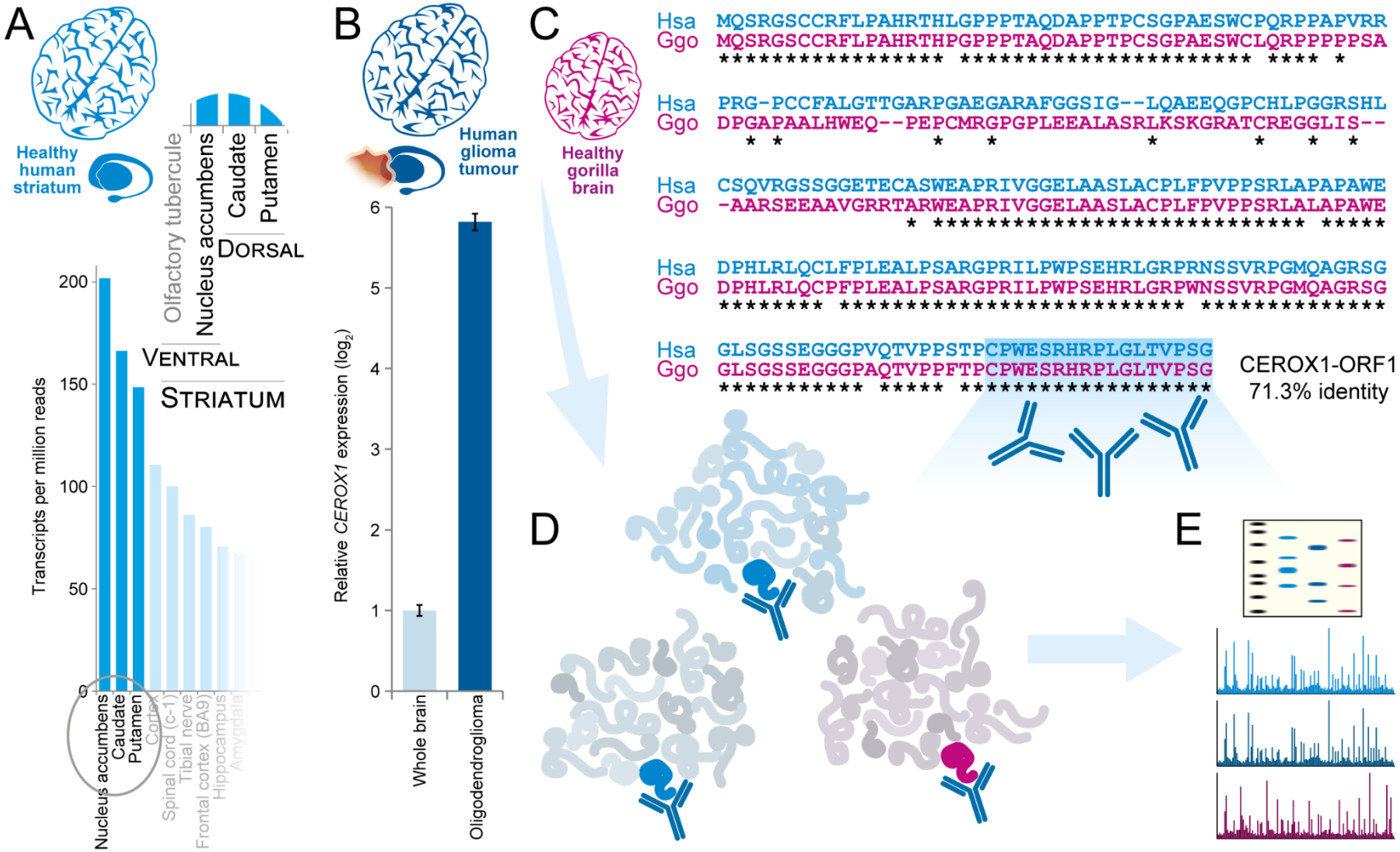


Figure 4.11: Possible conserved translation of CEROX1-ORF1 *in vivo* was investigated in parallel in human and gorilla brain. (A) Human *CEROX1* is most highly expressed in the striatum. **(B)** Its expression also highly elevated in oligodendroglioma tumour tissue. **(C)** Meanwhile, the human CEROX1-ORF1 sequence in most similar to that in the western lowland gorilla (*Gorilla gorilla gorilla*), with 70% identity across the ORF, and near-exact similarity at the N- and C-termini. This includes the epitope for the CEROX1-ORF1 antibody in the final 17 amino acids. N.B. occasional residue agreements in between the two highly conserved blocks are incidental, since two 2 bp indels have frame-shifted the otherwise well-conserved nucleotide sequence. **(D)** This antibody was used to immunoprecipitate putative CEROX1-ORF1 molecules from homogenates of these samples, and the resulting eluates were analysed by western blotting and mass spectrometry **(E)**.

While no peptides representing a human CEROX1-ORF1 protein were identified, evidence was uncovered to support translation of gorilla CEROX1-ORF1 in cortex. An 11 amino acid peptide HRPLGLTVPSG (5% total coverage) representing the C-terminal end of the gorilla sequence was identified from 13 collision-induced dissociation fragment ions (Figure 4.12).

The peptide was uniquely matched to CEROX1-ORF1 ($p = 4.7 \times 10^{-39}$), and no other complete matches were identified in BLASTP and TBLASTN searches against protein and nucleotide databases, respectively, for gorilla or human, or any other species. In the absence of alternative potential sources of this peptide, it was concluded that CEROX1-ORF1 is translated, to some extent, in gorilla cortex.

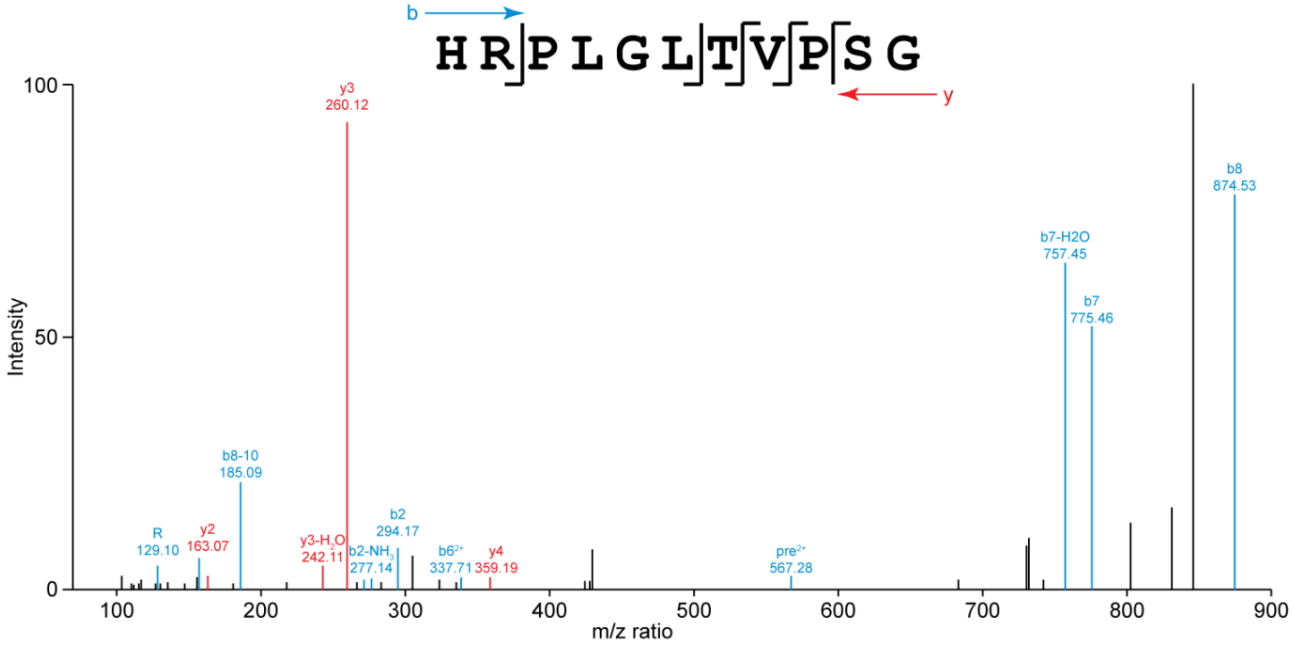


Figure 4.12: MS/MS reveals the presence of a CEROX1-ORF1 peptide in gorilla cortex. Immunoprecipitation eluates were analysed by tandem mass spectrometry. Shown above is the MS2 spectrum for the fragments of the collision-induced dissociation of the HRPLGLTVPSG peptide, resulting in 13 identifiable fragment ions (9 b ions from the N-terminus, blue; 4 y ions from the C-terminus, red).

4.5.2.2 IHC suggests that CEROX1-ORF1 is translated in a highly cell type-specific manner

The extremely low level of CEROX1-ORF1 protein expression in gorilla cortex and lack of detection in human striatum, while anticipated due to the previous absence of identification of this protein, suggests that, if present, a) expression might be very low, and b) the translation of this ORF might be highly restricted. Its expression might be limited to a specific cell type or sub-population thereof, or in a temporal or developmental manner. To complement the above proteomic strategy used to probe CEROX1-ORF1 expression *in vivo*, the putative protein was also investigated by immunohistochemistry (IHC), which carries the advantage of examining individual cells and cell type populations, interchanging specificity for

sensitivity. Within sections of the same human striatum sample used for the IP-MS strategy above, a number of cells were stained using the custom CEROX1-ORF1 antibody (Figure 4.13A). Widespread punctate non-specific staining was seen in the no primary control, suggesting cross-reactivity of the secondary antibody (Figure 4.13B), but the specific staining of the antibody was also distinguished morphologically. Cells fell into two categories: relatively sparse cells with very large somata but few to no projections (Figure 4.13C); and clusters of cells with small somata and extensive projections (Figure 4.13D).

Unlike the cerebral cortex and cerebellum, for example, the striatum is a highly heterogeneous non-laminar tissue. Aside from the striosome-matrix distinction discussed in Chapter 3, there is little arrangement of cells. The great majority (around 95%) of the neurons in the striatum are GABA-ergic medium spiny projection neurons. The remaining neurons are GABA-ergic or cholinergic interneurons, representing around 4% and 1% of the neuronal population, respectively. Additionally, the glial population includes astrocytes, myelinating oligodendrocytes, and microglia. Based upon the diameters of their somata, it was hypothesised that the larger CEROX1-ORF1-positive cells were cholinergic interneurons, but the identity of the smaller, spinier cells was less predictable. In order to clarify the cell types involved, sections were co-stained with a number of cell type markers: FOXP1 stains medium spiny projection neurons; glutamate decarboxylase (GAD65) would be expected to stain all GABA-ergic cells, including projection neurons and GABA-ergic interneurons; choline acetyltransferase (CHAT, Figure 4.13E) is required for acetylcholine production in cholinergic neurons; glial fibrillary acidic protein (GFAP, Figure 4.13F) is an intermediate filament protein present in astrocytes; OLIG2 is a transcription factor involved in oligodendrocyte development and maintenance; and ionized calcium-binding adapter molecule (IBA1) is a marker of microglia and macrophages. Co-localisation of CHAT and GFAP staining in the two populations of CEROX1-ORF1-positive cells indicates that these are cholinergic interneurons (Figure 4.13G) and astrocytes (Figure 4.13H), respectively. Interestingly, only a subset of the GFAP-positive astrocyte population was stained by the CEROX1-ORF1 antibody, suggesting that the putative protein is restricted to a very specific sub-population of cells, as hypothesised above.

Unfortunately, attempts to stain for CEROX1-ORF1 in gorilla cortex, following successful identification of translation in this tissue, were hampered by its preparation: the chunk of frozen tissue was not amenable to cryosectioning and all morphology was lost.

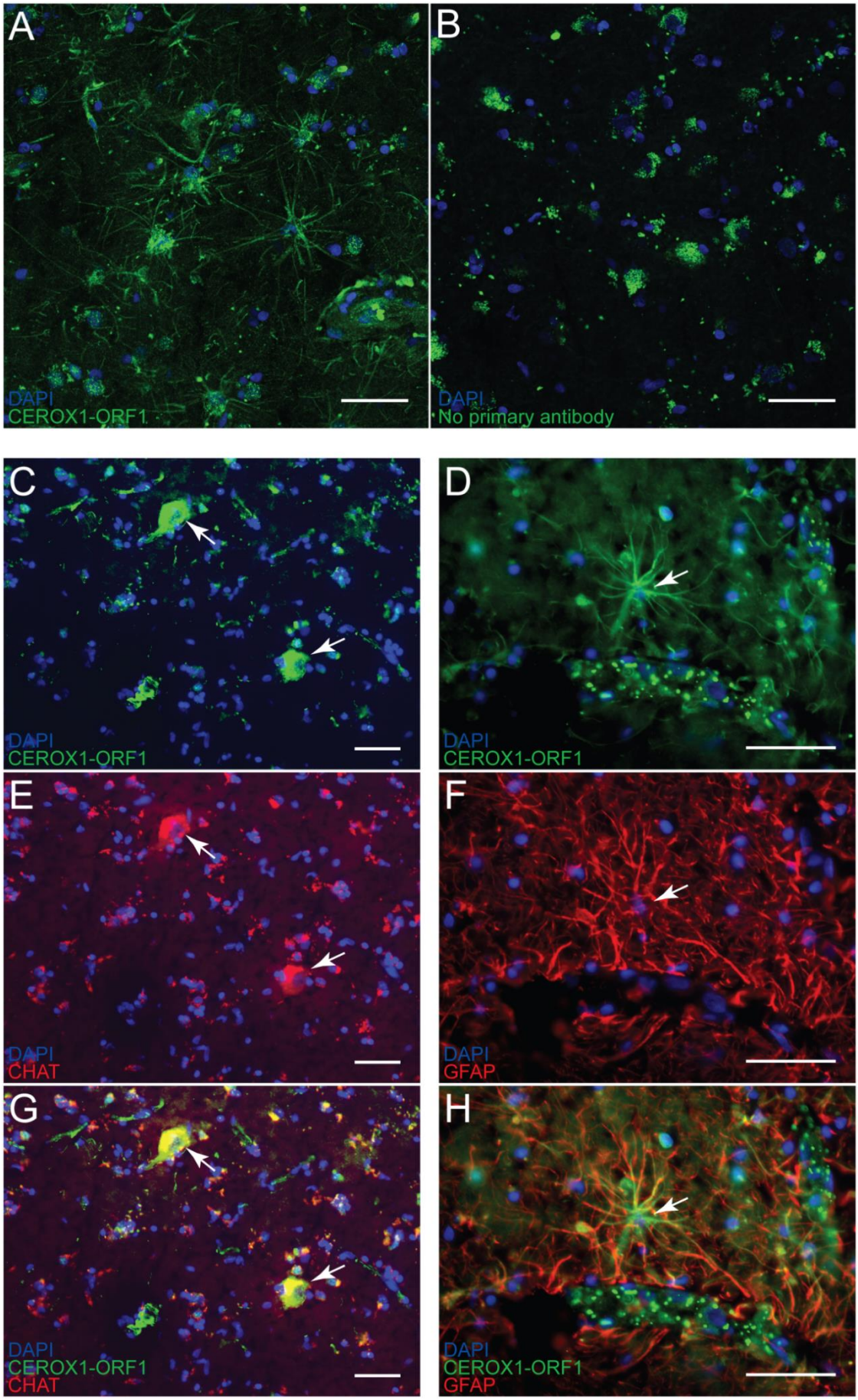


Figure 4.13: Immunostaining of CEROX1-ORF1 *in vivo* reveals cell type specificity. (A) In the human striatum, the CEROX1-ORF1 antibody stains a number of cells with extensive projections. Other cells are subject to non-specific staining, as illustrated by the no primary control **(B)**. Staining of CEROX1-ORF1 **(C-D)** alongside with the cell type markers CHAT **(E)** and GFAP **(F)** illustrate co-localisation of the antibody target with these proteins in cholinergic interneurons **(G)** and astrocytes **(H)**, respectively (arrows indicate exemplar cells). Scale bars represent 50 µm.

4.5.2.3 Antibody specificity *in vivo* is difficult to determine

Despite the fact that the immunohistochemical staining results are consistent with the translation of a rare protein from CEROX1-ORF1, it is difficult to validate that the staining seen as a result of antibody binding is in fact due to the existence of the target protein. To address this, a test of specificity was sought in the form of a peptide competition assay, in which the immunogenic peptide used to produce the polyclonal CEROX1-ORF1 antibody was co-incubated along with the antibody during primary incubation. It would be expected that the presence of the antigen in large molar excess would block the paratope of the antibody responsible for antigen-specific staining, and remove *bona fide* signal. Unusually, the staining signal was massively *increased* by the presence of the antigenic peptide (Figure 4.14A-C). This suggests that the occlusion by the peptide antigen of the specific antibody paratope leads to more extensive non-specific binding through other sites on the antibody molecule. This is consistent with the results of a peptide competition assay in western blotting on human and tissue lysates, in which inclusion of the peptide along with the antibody results in enhanced staining of, and additional, presumably-non-specific bands (Figure 4.14D). As a result of these data, it is impossible to draw conclusions regarding the specificity of the human IHC staining experiment, though this does not mean that the staining observed is non-specific.

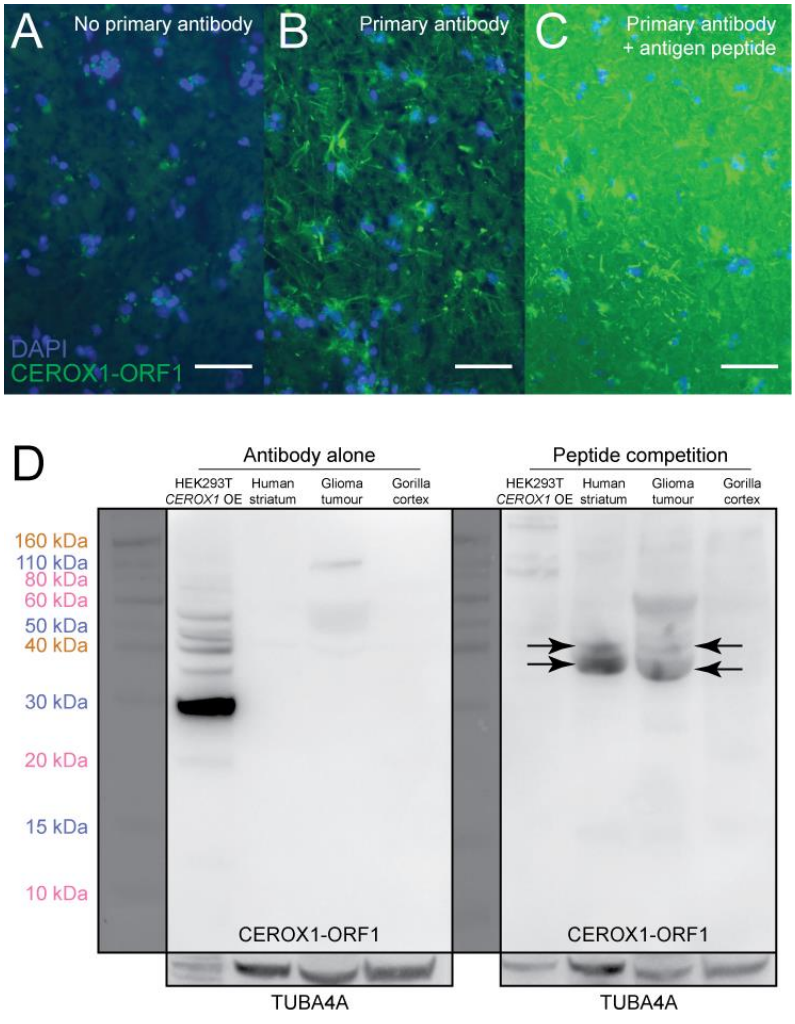


Figure 4.14: Specificity of binding of the CEROX1-ORF1 antibody was not clarified by a peptide competition assay. Incubation with secondary antibody but no primary antibody resulted in a homogeneous low level of background in immunocytochemistry **(A)**, while incubation with the custom CEROX1-ORF1 antibody achieved reproducible staining of a proportion of cells in the human striatum **(B)**. Addition of the antigenic peptide during primary incubation massively increased staining density in ICC **(C)** and in western blotting **(D)**, in which additional bands (arrows) were also stained in samples of human striatum, as well as glioma tumour, with the addition of antigenic peptide to the primary antibody incubation.

DISCUSSION

Since the advent of long non-coding RNAs, there have been many attempts to classify all genes into coding and non-coding classes, largely based upon broad patterns in whole-genome comparisons. Understanding of the complexity of the relationships between protein-coding, non-coding, and indeed 'coding non-coding RNAs', has been furthered by synergistic analyses of bioinformatic and wide-scale experimental approaches, but has also identified the need for well-characterised case studies *in vivo*.

CEROX1 was immediately classified as a lncRNA upon its annotation in 2012, but is also an example of a gene that does not sit cleanly within an exclusive ncRNA compartment, hosting an ORF of significant length that is robustly translated *in vitro*. The crucial question of *in vivo* expression of CEROX1-ORF1 remains incompletely answered in human basal ganglia, though detection of a unique peptide in gorilla cortex is suggestive of conserved translation. This study is the first targeted experimental investigation towards, and detection of, a novel protein-coding ORF in a non-human primate, and therefore the first primate-lineage comparative study of the protein-coding quality of a lncRNA.

There are still many reasons that *bona fide* CEROX1-ORF1 translation might not have been detected *in vivo*. Expression might be restricted to a tissue or developmental stage not addressed in this study, which preliminarily focused on the site of maximal expression; the first extension to this study would be application of the IP-MS approach developed to a wider range of human tissues. Context-specific translation efficiency may not result in a simple positive linear relationship between levels of *CEROX1* RNA and of translation. As shown for CEROX1-ORF2 *in vitro*, CEROX1-ORF1 may have a relatively short half-life *in vivo*. Given the typically long post-mortem delay for frozen brain tissue (~48 hours for the gorilla cortex and human striatum samples used in this study), this would heavily impede detection. Modification of proteins *in vivo*, especially with more extensive and repetitive PTMs such as glycosylation, can make MS identification difficult, owing to the ambiguity of peptide sizes. As mentioned previously, exogenous expression of epitope tagged CEROX1-ORF1 *in vitro* results in translation of multiple detectable bands, which might represent modifications that could be present *in vivo*. The clarity of the bands, and the fact that they are detected by both the anti-HA and the anti-CEROX1-ORF1 antibodies suggests that they are not products of degradation or aggregation as a consequence of sample preparation. Work was undertaken to elucidate the identities of the additional bands, questioning hypotheses: that the protein is ubiquitinated

(by ubiquitin immunoprecipitation); that the protein is glycosylated (using Schiff staining); and that the N-terminus is cleaved *in vitro* to produce smaller species (by additional epitope tagging of the N-terminal end). None of these routes of investigation were fruitful: no ubiquitination or glycosylation was detected, and limited time prevented a thorough appraisal of the behaviour of doubly tagged CEROX1-ORF1 proteins.

As mentioned above, the identification of protein-coding activity in *CEROX1* would not subtract from the importance of its non-coding activities, and may simply be an addition to its repertoire. Indeed, while conservation of transcription at the *CEROX1* locus is almost certainly due to its activities as a lncRNA, in light of the conserved expression levels and pattern of *Cerox1* in rodents despite the loss of the CEROX1-ORF1-containing sequence, one might posit that this stretch of sequence previously had no essential function, and that the evolution of a protein-coding element is therefore an addition to the other, non-coding, functions of the gene. Alternatively, a protein-coding ORF could have co-evolved into the locus with another element related to the non-coding function. Indeed the protein-coding mRNA-like features of *CEROX1*, such as its high GC content, lack of other uORFs (from the perspective of CEROX1-ORF1), and high expression level may be the very features that have permitted it to evolve a protein-coding ORF independently of its conserved expression as a functional lncRNA. The aforementioned presence of human-only protein-coding genes[50] illustrates that new coding regions have arisen during recent evolution. Additionally, the accelerated evolution of several genes associated with brain development throughout the hominine lineage (human, chimpanzee, and gorilla) illustrates a role for adaptive modification in recent evolution[52]. The evolution of CEROX1-ORF1 within *CEROX1* may be consistent with the 'proto-gene' model of protein evolution demonstrated by Carvunis et al.[93]: the translation of an initially 'non-genic' ORF that arose by neutral evolution may provide material that becomes subject of further functional relevance in the future, via the incidental arrival of adaptive advantage within the product of the ORF[94,95]. It seems sensible that a sequence that is already stably transcribed is a more likely candidate for the evolution of protein-coding function than a random sequence[96], though the existing functionality of the RNA may constrain the necessary changes[17].

Assuming that CEROX1-ORF1 is already stably translated and functional, it is challenging to speculate upon its possible role, due to its recent evolution and subsequent lack of homologues. The protein contains an unusually high proportion of proline (17%). The orthologous sequence in chimpanzee is annotated as a putative collagen-like protein, presumably owing to this fact. Based upon the commonality of the roles of

currently characterised bifunctional RNAs, and the putative electron transport-modulating role of *CEROX1*, a CEROX1-ORF1 protein might be expected to modulate cellular or mitochondrial energy metabolism. The small peptide MOTS-c, which is encoded within a presumed-non-coding mitochondrial rRNA gene, modulates energy metabolism by interaction with the folate cycle and subsequent activation of AMPK[97].

The apparent lack of subcellular localisation in HEK293T cells overexpressing CEROX1-ORF1 does not necessarily imply that endogenous CEROX1-ORF1 would be distributed in the same way. The previously mentioned *Drosophila* RNA *osk* is strictly spatially regulated to the posterior pole of oocytes under wild type expression conditions, but overexpression using the GAL4-UAS system results in erroneous clustering of a large portion of *osk* RNA in the centre of the oocyte[98].

Assuming that the results of the *in vivo* search for CEROX1-ORF1 in human striatum are representative of other tissues, and therefore that the protein is not present in human *in vivo*, then the difference in behaviour of *CEROX1 in vitro* and *in vivo* must be addressed. Based upon analogy to the lncRNA *GAS5*, it might be suggested that *CEROX1* is subject to nonsense-mediated decay owing to its downstream exon-exon junction or long 3'-UTR, thereby invoking degradation of *CEROX1* following association with the ribosome, which ribosome profiling indicates does take place. The exogenously *CEROX1* overexpression construct contains a cDNA transcript that does not require splicing, thereby precluding the activation of NMD by a downstream exon-exon junction complex. However, the extremely high steady state expression of *CEROX1* in human tissues suggests that it is not subject to such rapid degradation. Perhaps association with the ribosome in fact stabilises *CEROX1*[39]. Both possibilities could be tested in human cells by inhibiting transcription with actinomycin D and then assessing whether inhibition of NMD by cycloheximide results in an altered transcript half-life. It would be informative to: a) use a genome editing approach such as CRISPR to epitope tag the endogenous CEROX1-ORF1 locus *in vitro*, and b) examine whether exogenous overexpression of *CEROX1* in a more physiological model, such as primary cell culture, or indeed *in vivo*, supports robust translation. The differential protein expression of CEROX1-ORF1 endogenously *in vivo* and exogenously *in vitro* has ramifications for the study of *CEROX1* as well as genetic therapies targeting existing genetic materials as well as utilising siRNAs and other oligonucleotides. While off-target and downstream effects of attempted perturbation of a 'characterised' sequence or gene are often considered, it is commonly assumed that such effects will be mediated via the recognised product of the gene only. Even if a gene is not classified as a coding non-coding RNA, overexpression of an apparent lncRNA might produce

protein species. Conversely, this is a reminder that altering expression of a known protein-coding gene might elicit effects through a non-coding function, endogenous or erroneous. A non-coding variant of the *Plscr4* (phospholipid scramblase 4) gene was recently shown to negatively regulate cardiac hypertrophy via decoy of miR-214 and subsequent co-ordinate upregulation of Mfn2 (mitofusin 2)[99], leading the authors to suggest that "*Plscr4* might act as a therapeutic target for the treatment of cardiac hypertrophy and heart failure". While this isoform lacks the ORF encoding the traditional Plscr4 protein, it does contain multiple ORFs and care should be taken to characterise any protein-coding activity that exogenous induction of apparently non-coding ceRNA activity might elicit.

Ultimately, it is still unknown whether CEROX1-ORF1 is translated *in vivo* in humans. Evidence of CEROX1-ORF1 translation in gorilla cortex, and under exogenous conditions in human cells *in vitro* is supportive of this hypothesis, despite a lack of literature describing any other ORFs expressed *in vitro* and not *in vivo* – it may be that CEROX1-ORF1 is the first cautionary tale of this category. Only identification of CEROX1-ORF1 protein, as was attempted using mass spectrometry, can categorically state whether *CEROX1* is translated *in vivo*. The investigation of a greater range of human tissues would more thoroughly probe whether translation products are present *in vivo*, though the absence of a protein is very hard to prove, in the knowledge that it may be short-lived or, though some biochemical or biophysical characteristic, be refractory to identification by techniques such as MS. Given that immunohistochemical staining of striatum revealed sparse cells positive for anti-CEROX1-ORF1 antibody staining, it would be prudent to purify these specific cells, or other anti-CEROX1-ORF1-positive populations that might be identified in other tissues, by laser capture microdissection, prior to proteomic analysis.

Without knowledge of stable translation *in vivo*, analyses of possible function are speculative and of little relevance. However, future functional studies of *CEROX1*, even those assuming that it is a non-coding RNA *in vivo*, should account for the possible roles, physiological or not, of CEROX1-ORF1 translation under overexpression conditions, since the translation of CEROX1-ORF1 in HEK293T and SH-SY5Y cells is robust, and the production of even an artificial protein may elicit effects. This may be controlled for using an ORF1 start codon mutant to preclude translation. This investigation of the coding potential of *CEROX1* is therefore informative even though its final conclusion remains to be seen.

REFERENCES

1. Kochetov, A.V. AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. *Bioinformatics* **21**, 837-840 (2005).
2. Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C. *et al.* The Abundance of Short Proteins in the Mammalian Proteome. *PLOS Genetics* **2**, e52 (2006).
3. Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A. & Couso, J.P. Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family. *PLoS Biology* **5**, e106 (2007).
4. Kessler, M.M., Zeng, Q., Hogan, S., Cook, R. *et al.* Systematic Discovery of New Genes in the Saccharomyces cerevisiae Genome. *Genome Research* **13**, 264-271 (2003).
5. Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E. *et al.* Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Research* **16**, 365-373 (2006).
6. He, C., Jia, C., Zhang, Y. & Xu, P. Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel smORFs in Saccharomyces cerevisiae. *J Proteome Res* (2018).
7. Röhrig, H., Schmidt, J., Miklashevichs, E., Schell, J. & John, M. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proceedings of the National Academy of Sciences* **99**, 1915-1920 (2002).
8. Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S. *et al.* Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology* **9**, 660 (2007).
9. Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U. *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3**, e03528 (2014).
10. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**, 981-93 (2014).
11. Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology* **9**, 59 (2012).
12. Pauli, A., Norris, M.L., Valen, E., Chew, G.-L. *et al.* Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **343**(2014).
13. Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**, 228 (2016).
14. Guo, B., Zhai, D., Cabezas, E., Welsh, K. *et al.* Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**, 456-61 (2003).
15. Kumari, P. & Sampath, K. cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. *Seminars in Cell & Developmental Biology* **47-48**, 40-51 (2015).
16. Joyce, G.F. The antiquity of RNA-based evolution. *Nature* **418**, 214-221 (2002).
17. Dinger, M.E., Gascoigne, D.K. & Mattick, J.S. The evolution of RNAs with multiple functions. *Biochimie* **93**, 2013-2018 (2011).
18. Lanz, R.B., McKenna, N.J., Onate, S.A., Albrecht, U. *et al.* A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell* **97**, 17-27 (1999).
19. Chooniedass-Kothari, S., Emberley, E., Hamedani, M.K., Troup, S. *et al.* The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett* **566**, 43-7 (2004).
20. McKay, D.B., Xi, L., Barthel, K.K.B. & Cech, T.R. Structure and Function of Steroid Receptor RNA Activator Protein, the Proposed Partner of SRA Noncoding RNA. *Journal of Molecular Biology* **426**, 1766-1785 (2014).
21. Hubé, F., Velasco, G., Rollin, J., Furling, D. & Francastel, C. Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation. *Nucleic Acids Research* **39**, 513-525 (2011).
22. Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C. *et al.* Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. *DNA Cell Biol* **25**, 418-28 (2006).
23. Ulveling, D., Francastel, C. & Hubé, F. When one is better than two: RNA with dual functions. *Biochimie* **93**, 633-644 (2011).
24. Naski, N., Gajjar, M., Bourougaa, K., Malbert-Colas, L. *et al.* The p53 mRNA-Mdm2 interaction. *Cell Cycle* **8**, 31-4 (2009).
25. Yin, Y., Stephen, C.W., Luciani, M.G. & Fåhraeus, R. p53 stability and activity is regulated by Mdm2-mediated induction of alternative p53 translation products. *Nature Cell Biology* **4**, 462 (2002).

26. Candeias, M.M., Malbert-Colas, L., Powell, D.J., Daskalogianni, C. *et al.* p53 mRNA controls p53 activity by managing Mdm2 functions. *Nature Cell Biology* **10**, 1098 (2008).
27. Nagano, H., Yamagishi, N., Tomida, C., Yano, C. *et al.* A novel myogenic function residing in the 5' non-coding region of Insulin receptor substrate-1 (Irs-1) transcript. *BMC Cell Biol* **16**, 8 (2015).
28. Jenny, A., Hachet, O., Zavorszky, P., Cyrklaff, A. *et al.* A translation-independent role of oskar RNA in early Drosophila oogenesis. *Development* **133**, 2827-33 (2006).
29. Nakamura, A., Sato, K. & Hanyu-Nakamura, K. Drosophila Cup Is an eIF4E Binding Protein that Associates with Bruno and Regulates oskar mRNA Translation in Oogenesis. *Developmental Cell* **6**, 69-78 (2004).
30. Castagnetti, S. & Ephrussi, A. Orb and a long poly(A) tail are required for efficient oskar translation at the posterior pole of the Drosophila oocyte. *Development* **130**, 835-843 (2003).
31. Kim-Ha, J., Smith, J.L. & Macdonald, P.M. oskar mRNA is localized to the posterior pole of the Drosophila oocyte. *Cell* **66**, 23-35 (1991).
32. Ephrussi, A., Dickinson, L.K. & Lehmann, R. oskar organizes the germ plasm and directs localization of the posterior determinant nanos. *Cell* **66**, 37-50 (1991).
33. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes. *Cell* **147**, 789-802 (2011).
34. Tani, H., Torimura, M. & Akimitsu, N. The RNA Degradation Pathway Regulates the Function of GAS5 a Non-Coding RNA in Mammalian Cells. *PLOS ONE* **8**, e55684 (2013).
35. Smith, C.M. & Steitz, J.A. Classification of gas5 as a Multi-Small-Nucleolar-RNA (snoRNA) Host Gene and a Member of the 5'-Terminal Oligopyrimidine Gene Family Reveals Common Features of snoRNA Host Genes. *Molecular and Cellular Biology* **18**, 6897-6909 (1998).
36. D'Lima, N.G., Ma, J., Winkler, L., Chu, Q. *et al.* A human microprotein that interacts with the mRNA decapping complex. *Nature Chemical Biology* **13**, 174-180 (2017).
37. Doyle, J.P., Dougherty, J.D., Heiman, M., Schmidt, E.F. *et al.* Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types. *Cell* **135**, 749-762 (2008).
38. Heiman, M., Kulicke, R., Fenster, R.J., Greengard, P. & Heintz, N. Cell type–specific mRNA purification by translating ribosome affinity purification (TRAP). *Nature Protocols* **9**, 1282 (2014).
39. Bazin, J., Baerenfaller, K., Gosai, S.J., Gregory, B.D. *et al.* Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. **114**, E10018-E10027 (2017).
40. Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K. *et al.* Decoding Human Cytomegalovirus. *Science* **338**, 1088-1093 (2012).
41. Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell reports* **8**, 1365-1379 (2014).
42. Ha, S.-W., Ju, D. & Xie, Y. Nuclear Import Factor Srp1 and Its Associated Protein Sts1 Couple Ribosome-bound Nascent Polypeptides to Proteasomes for Cotranslational Degradation. *Journal of Biological Chemistry* **289**, 2701-2710 (2014).
43. Wheatley, D.N., Giddings, M.R. & Inglis, M.S. Kinetics of degradation of 'short-' and 'long-lived' proteins in cultured mammalian cells. *Cell Biology International Reports* **4**, 1081-1090 (1980).
44. Rock, K.L., Farfán-Arribas, D.J., Colbert, J.D. & Goldberg, A.L. MHC class I-presented peptides and the DRiP hypothesis. *Trends in immunology* **35**, 144-152 (2014).
45. Yewdell, J.W., Antón, L.C. & Bennink, J.R. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *The Journal of Immunology* **157**, 1823-1826 (1996).
46. Yewdell, J.W. & Nicchitta, C.V. The DRiP hypothesis decennial: support, controversy, refinement and extension. *Trends in Immunology* **27**, 368-373 (2006).
47. Nishikawa, T., Ota, T. & Isogai, T. Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics* **16**, 960-7 (2000).
48. Dinger, M.E., Pang, K.C., Mercer, T.R. & Mattick, J.S. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLOS Computational Biology* **4**, e1000176 (2008).
49. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-82 (2011).
50. Knowles, D.G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res* **19**, 1752-9 (2009).
51. Wilkinson, R.D., Steiper, M.E., Soligo, C., Martin, R.D. *et al.* Dating Primate Divergences through an Integrated Analysis of Palaeontological and Molecular Data. *Systematic Biology* **60**, 16-31 (2011).

52. Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169 (2012).
53. Moorjani, P., Amorim, C.E.G., Arndt, P.F. & Przeworski, M. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences* **113**, 10607-10612 (2016).
54. Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345-9 (2007).
55. Han, S., Liang, Y., Li, Y. & Du, W. Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination. *Biomed Res Int* **2016**, 8496165 (2016).
56. Sun, K., Chen, X., Jiang, P., Song, X. *et al.* iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* **14**, S7-S7 (2013).
57. Kang, Y.J., Yang, D.C., Kong, L., Hou, M. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* **45**, W12-w16 (2017).
58. Fickett, J.W. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research* **10**, 5303-5318 (1982).
59. Wang, L., Park, H.J., Dasari, S., Wang, S. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41**, e74-e74 (2013).
60. Fickett, J.W. & Tung, C.S. Assessment of protein coding measures. *Nucleic Acids Research* **20**, 6441-6450 (1992).
61. Arrial, R.T., Togawa, R.C. & Brigido Mde, M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics* **10**, 239 (2009).
62. Panwar, B., Arora, A. & Raghava, G.P. Prediction and classification of ncRNAs using structural information. *BMC Genomics* **15**, 127 (2014).
63. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-27 (2011).
64. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).
65. Niazi, F. & Valadkhan, S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA* **18**, 825-43 (2012).
66. Ziegler, C. & Kretz, M. The More the Merrier—Complexity in Long Non-Coding RNA Loci. *Frontiers in Endocrinology* **8**, 90 (2017).
67. Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-7 (2014).
68. Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-81 (2014).
69. Gaudet, P., Michel, P.-A., Zahn-Zabal, M., Britan, A. *et al.* The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Research* **45**, D177-D182 (2017).
70. Paik, Y.-K., Jeong, S.-K., Omenn, G.S., Uhlen, M. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature Biotechnology* **30**, 221 (2012).
71. Baker, M.S., Ahn, S.B., Mohamedali, A., Islam, M.T. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat Commun* **8**, 14271 (2017).
72. Elguoshy, A., Hirao, Y., Xu, B., Saito, S. *et al.* Identification and Validation of Human Missing Proteins and Peptides in Public Proteome Databases: Data Mining Strategy. *J Proteome Res* (2017).
73. Fonslow, B.R., Carvalho, P.C., Academia, K., Freeby, S. *et al.* Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *J Proteome Res* **10**, 3690-700 (2011).
74. Khatter, H., Myasnikov, A.G., Natchiar, S.K. & Klaholz, B.P. Structure of the human 80S ribosome. *Nature* **520**, 640 (2015).
75. Calviello, L., Mukherjee, N., Wyler, E., Zauber, H. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **13**, 165-70 (2016).
76. Steitz, J.A. Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. *Nature* **224**, 957 (1969).
77. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-23 (2009).
78. van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S. *et al.* Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol* **15**, R6 (2014).

79. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. & Lander, E.S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154**, 240-251 (2013).
80. Lee, S., Liu, B., Lee, S., Huang, S.X. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **109**, E2424-32 (2012).
81. Gonzalez, C., Sims, J.S., Hornstein, N., Mela, A. *et al.* Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J Neurosci* **34**, 10924-36 (2014).
82. Rock, K.L., Gramm, C., Rothstein, L., Clark, K. *et al.* Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell* **78**, 761-71 (1994).
83. Palombella, V.J., Rando, O.J., Goldberg, A.L. & Maniatis, T. The ubiquitin-proteasome pathway is required for processing the NF-κB1 precursor protein and the activation of NF-κB. *Cell* **78**, 773-785 (1994).
84. Yoshimori, T., Yamamoto, A., Moriyama, Y., Futai, M. & Tashiro, Y. Bafilomycin A1, a specific inhibitor of vacuolar-type H(+)-ATPase, inhibits acidification and protein degradation in lysosomes of cultured cells. *Journal of Biological Chemistry* **266**, 17707-17712 (1991).
85. Mauvezin, C., Nagy, P., Juhász, G. & Neufeld, T.P. Autophagosome–lysosome fusion is independent of V-ATPase-mediated acidification. *Nat Commun* **6**, 7007 (2015).
86. Mauvezin, C. & Neufeld, T.P. Bafilomycin A1 disrupts autophagic flux by inhibiting both V-ATPase-dependent acidification and Ca-P60A/SERCA-dependent autophagosome-lysosome fusion. *Autophagy* **11**, 1437-1438 (2015).
87. Yamamoto, A., Tagawa, Y., Yoshimori, T., Moriyama, Y. *et al.* Bafilomycin A1 Prevents Maturation of Autophagic Vacuoles by Inhibiting Fusion between Autophagosomes and Lysosomes in Rat Hepatoma Cell Line, H-4-II-E Cells. *Cell Structure and Function* **23**, 33-42 (1998).
88. Klionsky, D.J., Elazar, Z., Seglen, P.O. & Rubinsztein, D.C. Does bafilomycin A1 block the fusion of autophagosomes with lysosomes? *Autophagy* **4**, 849-50 (2008).
89. Tanida, I., Ueno, T. & Kominami, E. LC3 and Autophagy. in *Autophagosome and Phagosome* (ed. Deretic, V.) 77-88 (Humana Press, Totowa, NJ, 2008).
90. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *Journal of Molecular Biology* **300**, 1005-1016 (2000).
91. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**, 953 (2007).
92. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337 (2011).
93. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370 (2012).
94. Ruiz-Orera, J., Messeguer, X., Subirana, J.A. & Alba, M.M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).
95. Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genetics* **11**, e1005721 (2016).
96. Weisman, C.M. & Eddy, S.R. Gene Evolution: Getting Something from Nothing. *Current Biology* **27**, R661-R663 (2017).
97. Lee, C., Zeng, J., Drew, Brian G., Sallam, T. *et al.* The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance. *Cell Metabolism* **21**, 443-454 (2015).
98. Zimyanin, V., Lowe, N. & St Johnston, D. An Oskar-Dependent Positive Feedback Loop Maintains the Polarity of the Drosophila Oocyte. *Current Biology* **17**, 353-359 (2007).
99. Lv, L., Li, T., Li, X., Xu, C. *et al.* The lncRNA Plscr4 Controls Cardiac Hypertrophy by Regulating miR-214. *Molecular Therapy - Nucleic Acids* **10**, 387-397 (2018).

CHAPTER 5: *CEROX1* IS A PUTATIVE REGULATOR OF THE ELECTRON TRANSPORT CHAIN

INTRODUCTION

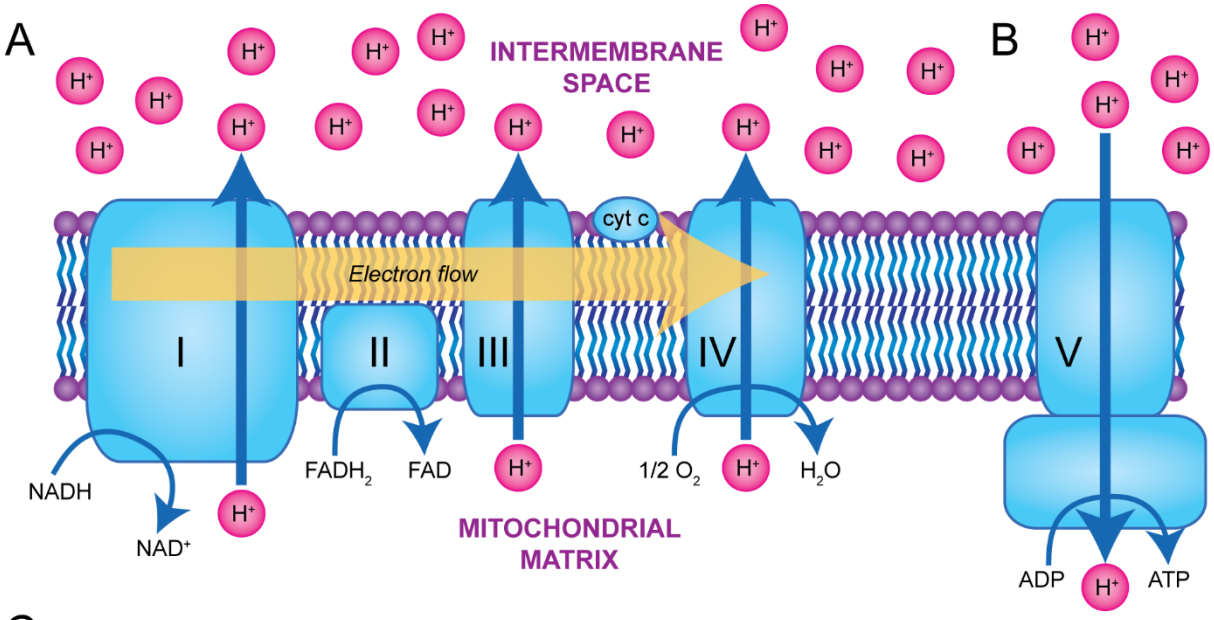
The *CEROX1* locus exhibits a substantial level of evolutionary constraint, having being conserved at the DNA and RNA levels in a tissue-specific manner for at least 90 million years. In light of this, this chapter will aim to identify whether the previously described functionality of mouse *Cerox1*, a putative regulator of mitochondrial energy metabolism, is also conserved in human *CEROX1*.

5.1 The mitochondrial electron transport chain is at the centre of eukaryotic life

Within the inner mitochondrial membrane, the mitochondrial electron transport chain (ETC) acts to meet around 90% of cellular ATP demand[1]. A pathway of four complexes I to IV couple the transport of electrons from the reduced electron carriers NADH and $FADH_2$ to the maintenance of a chemiosmotic proton gradient across the inner membrane (Figure 5.1A). The proton motive force generated is transduced into ATP synthesis by the proton-translocating complex V, the F_OF_1-ATP synthase (Figure 5.1B).

The complexes of the ETC are among the largest and most sophisticated proteins in the cell: human complex I comprises 45 subunits[2] (44 different[3]) and is almost 1 MDa in size (Figure 5.1C). Subunits of the ETC complexes are encoded by two genomes: nuclear DNA and the mitochondrial genome (mtDNA) (Figure 5.1D). The resulting requirement for regulation of production and assembly of products across multiple sub-cellular compartments adds an unparalleled level of regulatory complexity. The structure and assembly of mammalian complex I have only recently been elucidated[2,4-7].

Figure 5.1: The mitochondrial electron transport chain is a complex system (overleaf). (A) Electrons flow from NADH and $FADH_2$ along the ETC complexes I to IV, and are eventually used in the reduction of oxygen to water. Meanwhile, protons are pumped into the intermembrane space by complexes I, III, and IV, forming a proton motive force. **(B)** Protons flow down their chemiosmotic gradient via complex V, ATP synthase, by which ADP is phosphorylated to form ATP. **(C)** Each complex is a large multimer of subunits, with all except complex II incorporating genes from the nuclear and mitochondrial genomes **(D)**. N.B. Until recently, the unique subunit count for complex I was 45, but NDUFA4 has recently been assigned to complex IV[8]. NDUFAB1/SDAP is present twice in crystal structures, bringing the total subunit count per complex to 45[2]. Complex III is active as a dimer.



C

Complex	Size (kDa)	Electron transport	Subunits	
			Nuclear	mtDNA
I - NADH dehydrogenase	980	NADH to coenzyme Q	37	7
II - succinate dehydrogenase	260	$FADH_2$ to coenzyme Q	4	-
III - ubiquinol: cytochrome c oxidoreductase	500	Coenzyme Q to soluble cytochrome c, via the Q-cycle	10	1
IV - cytochrome c oxidase	200	Coenzyme Q to molecular oxygen, reducing it to water	11	3
V - ATP synthase	650	-	15	2

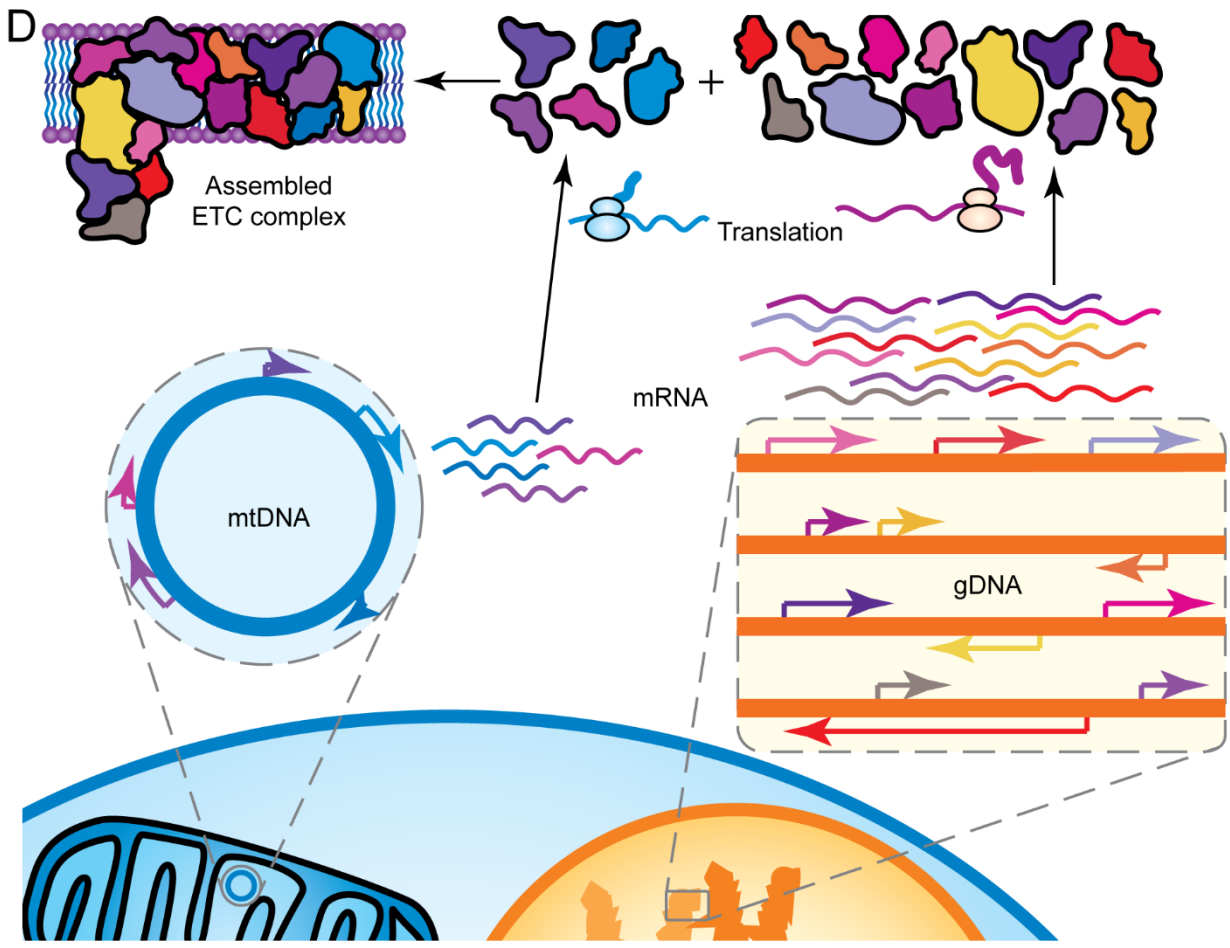


Figure 5.1: The mitochondrial electron transport chain is a complex system.

The ETC, and its dysfunction and deficiency, are implicated in a huge array of diseases, ranging from those specific to mutations in ETC subunits[9] to wider-reaching, aetiologically-complex conditions[10] including Parkinson's[11-13], Alzheimer's[14-19], and Huntington's[20-23] diseases. Leigh syndrome is a fatal congenital disease caused by mutation of complex I, III, and IV subunits[24-27].

5.1.2 The ETC is also subject to higher organisation into super-complexes

Contrary to the traditional depiction of the ETC as a linear pathway of physically independent protein complexes, these complexes are in fact subject to inter-complex interactions that promote formation of super-complexes[28,29]. The best studied super-complex is the "respirasome", $SCI_1III_2IV_1$ (super-complex with one CI, two CIIIs, and one CIV), which is capable of NADH:O_2 oxidoreduction in isolation *in vitro*[29]. Single particle electron microscopy has been used to describe a "mega-complex" $MCI_2III_2IV_2$[3], rivalling the ribosome at over 3 MDa. Though this was also used as a basis upon which to speculate that complex II might also associate with the mega-complex to form a full complement of complexes in an "electron transport chain super-complex"[3], no involvement of complex II within super-complexes has yet been demonstrated. The presence of some specific super-complexes is dependent upon cell type-specific subunit isoforms, with the long isoform of complex IV isoform Cox7a2l being required for the formation of $SCIII_2IV_1$, for example[30], and this contributes to differences in super-complex formation between mouse strains[31]. But as a general phenomenon these super-complexes are present in all mitochondrial systems yet studied, and elements of organisation appear to be evolutionarily conserved, with a conserved core of CI:CIII interactions being present in mammals, plants, and yeast[32]. The function of super-complexes is not yet understood, with hypotheses of substrate channelling and reduction of ROS production inconsistent[30,33] or insufficiently addressed, respectively, by the current state of the field[30]. The complexity of the ETC system makes unambiguously identifying a regulatory role of super-complex formation or modulation thereof by cellular parameters difficult[30], but it remains possible.

5.1.3 Post-transcriptional regulation of the mitochondrial ETC

Owing to the importance of the ETC within the cell, and it genetic complexity, it would be expected that the genes encoding its subunits are carefully co-regulated. Indeed, numerous mechanisms that have been

elucidated that modulate the levels of specific ETC complexes or their constituent subunits. For example, the protein YB-1 (p50), previously characterised as a multi-functional transcription and translation factor, has been shown to regulate translation of several nuclear-encoded ETC subunit proteins, such as NDUFA9 and SDHB[34]. The relative amount of YB-1 bound to mRNAs, which is partially regulated by growth factor pathway-induced phosphorylation that alters its affinity for mRNA, modulates its activity as a translational regulator and its recruitment to either silenced ribonucleoprotein complexes or active poly-ribosomes[35]. YB-1 may therefore represent an important growth-modulated regulator of ETC translation; depletion of YB-1 was shown to upregulate ETC activity via both complex I and complex II[34].

Another RNA binding protein factor that has a translational influence upon ETC complex subunits is En1/2 (collectively, Engrailed), better known for its transcriptional regulator activity. In mouse dopaminergic neurons in culture and in the midbrain *in vivo*, exogenously applied Engrailed upregulates translation of the complex I subunits Ndufs1 and Ndufs3, increases complex I activity, and protects neurons against the complex I inhibitors rotenone and MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine)[36], an MPP^+ precursor used to elicit selective neuronal death in Parkinson's disease models. Increased striatal dopamine levels were shown to be dependent on upregulation of Ndufs1 translation[36], though the mechanistic connection between mitochondrial activity and dopamine secretion has not been elucidated[37]. Impressively, translation rates of Ndufs1 and Ndufs3 were increased by 43% and 68%, respectively, within only an hour of Engrailed treatment, indicating the potency of translation-based modulation[36].

5.1.3.1 Co-ordinate regulation of mitochondria and ETC complexes by miRNAs

Several miRNAs have been shown to target transcripts encoding ETC complex subunits and thereby effect regulation of respiratory function. miR-210, which is upregulated during hypoxia[38], has been shown to regulate the levels of several mitochondrial proteins, including ETC subunits: the complex II subunit *SDHD*[39,40], and the complex IV subunit and assembly factor *NDUFA4*[39] and *COX10*[41], respectively. Furthermore, the iron-sulphur cluster assembly enzymes *ISCU1* and *ISCU2* are directly downregulated by miR-210[41,42], and this consequently reduces the activity of complex I, which incorporates a number of iron-sulphur centres. By simultaneously targeting components within, or involved in the maturation of, multiple ETC complexes, miR-210 thereby effects a co-ordinated regulation across the ETC pathway[43]. *In vivo*, human placentas manifesting pre-eclampsia, which involves formation of a hypoxic state, demonstrate reduced

activity of complex III, and reduced levels of complexes I and IV[44], likely in part due to reduced ISCU expression[44,45]. Recapitulation of these findings in primary trophoblasts also led to the demonstration that the dysfunction may be reversed using a miR-210 inhibitor[44], suggesting that the role of this miRNA as a master regulator of the hypoxic response[46] incorporates post-transcriptional modulation of the ETC. Interestingly, miR-147b shares its seed with miR-210, and also interacts with *SDHD* and *NDUFA4* RNAs[47]; miR-147b is induced during inflammation, suggesting overlap between the post-transcriptional network responses to different states, and confirming the role of miRNAs in a general cellular stress response.

miR-338 has been shown to bind and downregulate the complex IV subunit *Cox4* in primary cultures of rat superior cervical ganglia neurons, with Cox4 mRNA and protein levels reciprocally responsive to modulation of miR-388 expression[48]. Interestingly, it was shown that the machinery required to generate and process miRNAs is present in axons independently of the cell soma[48]. Inhibition of endogenous miR-388 led to increased oxygen consumption and ATP synthesis, as well as increased neurotransmitter uptake, suggesting increased gross neuronal function; conversely, increased miR-338 led to reduced activity[48]. Other mitochondrial genes were predicted to be targets of miR-338, including *Atp5g1*. The regulation of this ATP synthase subunit at the mRNA and protein level by miR-338 was subsequently validated in superior cervical ganglia neurons[49], indicating a pleotropic regulatory influence of miR-338 upon the activity of the ETC.

In high-fat diet-fed mice, elevation of miR-33a was shown to reduce complex I activity by inhibiting Ndufa5 expression; this was rescued with an anti-miR-331 inhibitor, which also restored ATP production[50].

As well as directly impacting ETC subunit expression, miRNAs may also target: a) other molecules that post-transcriptionally regulate the ETC, such as those RNA-binding proteins described above (e.g. Engrailed and YB-1); and b) key transcription factors that regulate mitochondrial expression programs. For example, in response to changes in physical activity, miR-696 has been shown to directly regulate the expression of PGC-1α (peroxisome proliferator-activated receptor gamma coactivator)[51], a key transcriptional regulator of mitochondrial biogenesis, and of fatty acid oxidation, which feeds NADH and $FADH_2$ directly into the ETC.

5.1.3.2 MitomiRs extend the regulatory influence of miRNAs directly into the mitochondria

Interestingly, there is a growing body of evidence supporting the localisation of some miRNAs directly to the mitochondrion *in vivo*[52,53] and *in vitro*[54-56], though very few of these so-called "mitomiRs"[57] have been

characterised. Das et al.[58] showed that miR-181c is enriched within the mitochondria of rat hearts, where it targets the 3'-UTR of the mtDNA-encoded *mt-Cox1*, an important complex IV subunit, and reduces its translation[58]. Further study by the same group showed that exogenous miR-181c administration in rat hearts *in vivo* results in reduced RNA and protein expression of all three mtDNA-encoded complex IV subunits mt-Cox1, mt-Cox2, and mt-Cox3 and at least one nuclear-encoded subunit, Cox7a, suggesting gross downregulation of complex IV[59]. Interestingly, although complex IV activity was reduced, ETC oxygen consumption was increased, and this was attributed to increased ROS production by electron transport through complexes I and II[59]. Interestingly, miR-181c overexpression caused increased expression of the mitochondrial calcium uniporter, and increased basal calcium concentrations in the mitochondrial matrix. It was hypothesised that this increased the provision of ETC substrates by its positive modulation of pyruvate dehydrogenase and the TCA cycle enzyme α-ketoglutarate dehydrogenase[59]. Additionally, this reduces the viable calcium uptake potential of the mitochondria, leading to a more readily activated mitochondrial permeability transition pore[59], a marker that precedes apoptosis. Together with the complex IV block, this widespread mitochondria dysregulation led to heart failure in rats with only a 2-fold upregulation of miR-181c[59], illustrating the potential potency of mitomiRs in regulating cell viability and processes.

miR-151a-5p is upregulated in the seminal plasma of patients with asthenozoospermia, and has been shown *in vitro* in a spermatogonial cell line to negatively regulate *CYTB*, a crucial mtDNA-encoded subunit of complex III, resulting in reduced ETC activity and ATP production[60]. This illustrates the diverse physiological localities in which miRNA-mediated regulation of the ETC may be observed, and another instance of putatively causative dysregulated miRNA expression in disease. Interestingly, the muscle-specific miR-1 has been found to localise directly to mitochondria (~40%), where it is associated with Ago, but not other RISC components, and *increases* translation of mitochondrial proteins including mt-Cox1[61], suggesting further complexity.

5.1.3.3 Other RNAs may regulate the ETC by miRNA-dependent mechanisms

Given the apparent role of miRNAs in regulation of the ETC and mitochondrial metabolism, of which only a small portion may has been characterised, there would seem to be opportunity for an additional layer of ceRNA regulation, whereby other mRNAs or lncRNAs are able to regulate multiple complexes in concert via

the modulation of miRNA levels[43]. Mouse *Cerox1*, as will be described further below, is the first RNA described with such a role.

Interestingly, regulation of mitochondrial complex I by RNA is not limited to eukaryotes. Reeves et al.[62] demonstrated that infection of cells with human cytomegalovirus (HCMV) results in protection of cells against apoptosis stimulated by the complex I inhibitor rotenone. A viral-encoded lncRNA β2.7 interacts with the essential complex I subunit NDUFA13 (GRIM19) and prevents its re-location to the nucleus, where it acts as a pro-apoptotic signal[63]. This stabilises the integrity of complex I, the mitochondrial membrane potential and therefore the production of ATP, maintaining cell viability for HCMV's relatively long life cycle[62].

5.2 Previous work by the Ponting lab has characterised mouse *Cerox1* as a ceRNA regulating mitochondrial activity

The data illustrated in the following sections (5.2.1-5.2.2) were generated by Sirey et al. of the Ponting lab, including some assistance from me prior to the work of this doctoral thesis, as stated previously in the Declaration of Authorship. A manuscript[64] presenting these data may be found at the end of this thesis.

5.2.1 *Cerox1* regulates the mitochondrial electron transport chain

As a preliminary investigation of putative function, overexpression of *Cerox1* in N2A cells was analysed by microarray technology, and resulted in significant differential expression of 286 genes, of which 83% were upregulated, a highly significant skew towards concerted regulation (binomial test $p = 4.9 \times 10^{-31}$). Gene ontology analysis indicated a substantial and significant enrichment (>20-fold) for genes of the mitochondrial respiratory chain within those upregulated by *Cerox1* overexpression. RT-qPCR demonstrated that 12 of the 35 nuclear-encoded transcripts encoding complex I subunits are upregulated by more than 1.4-fold following *Cerox1* expression, and that reciprocal changes are seen upon *Cerox1* knockdown by shRNAs. The increases seen in RNA abundance are transduced into differential protein expression, with western blotting for two complex I subunits Ndufs1 and Ndufs3 indicating ~2-fold increases. Finally, it was shown that these changes at the RNA and protein level correspond to a 22% increase in enzymatic activity of complex I. Additionally, the activity of the terminal ETC complex IV was also

increased by 30%, and the overall activity of the ETC was raised accordingly, with 52% greater ATP-linked oxygen consumption. Conversely, knockdown of *Cerox1* elicited a decrease in complex I activity (11%), complex IV activity (19%), and ATP-linked oxygen consumption (52%). Importantly, the *Cerox1*-mediated changes in ETC activities were found to influence both ROS production and sensitivity to mitochondrial inhibitors: ROS production was 20% increased by *Cerox1* knockdown, and reduced by 50% upon overexpression, which elicited a 35% decrease in oxidative protein damage. Overexpression of *Cerox1* reduces the sensitivity of N2A cells to the complex IV inhibitor azide, and specifically reduces ROS production in the presence of the complex I inhibitor rotenone.

5.2.2 *Cerox1* acts by interacting with mmu-miR-488-3p

The ability of *Cerox1* to regulate complex I was demonstrated in mouse ES cells, and it was then shown that the phenotype was ablated by deletion of the miRNA-processing enzyme Dicer, suggesting that *Cerox1* acts via miRNAs. Sequential mutation of miRNA binding sites revealed that the activity of *Cerox1* is mediated by a single site responsible for binding mmu-miR-488-3p. *Cerox1* and of mmu-miR-488-3p reciprocally regulate one another's expression when overexpressed or knocked down, and physical interaction between the two RNAs was demonstrated by pulldown of a biotinylated mmu-miR-488-3p mimic. Furthermore, modulation of mmu-miR-488-3p levels resulted in reciprocal changes in RNA expression of all 12 *Cerox1*-sensitive biomarkers, most of which were also shown to be enriched by biotinylated pulldown.

Cerox1 is thereby a regulator of mitochondrial energy metabolism that elicits control over activity of the electron transport chain by via modulation of complex I activity mediated by mmu-miR-488-3p. Expression of miR-488 in mouse is low-to-moderate and extremely tissue-specific, being near-exclusively enriched in the nervous system in a conserved manner, akin to *CEROX1* orthologues, suggesting possible co-mechanistic-conservation of the two RNAs.

PRELIMINARY RESULTS

The data illustrated in the following sections (5.3.1-5.3.5) were generated prior to the work of this doctoral thesis, and have been incorporated into another degree submission, as stated previously in the Declaration of Authorship.

5.3 Initial studies of human *CEROX1* indicated that it is a conserved regulator of electron transport chain complexes

5.3.1 *CEROX1* upregulated the activities of complexes I and III of the human ETC *in vitro*

In order to investigate whether human *CEROX1* has conserved the ability held by the mouse transcript to regulate the ETC in human cells, *CEROX1* was overexpressed in HEK293T cells, and the activities of complexes I to IV were assayed by the same spectrophotometric assays used in the studies of mouse *Cerox1*.

HEK293T cells were initially chosen as a model in which to manipulate *CEROX1* expression for several reasons: they are of neural crest ectodermal origin[65], and thus possess a number of neural cell line characteristics, including expression of neuronal markers[66]; HEK293T cells are commonly used as an *in vitro* model for exploration of neurodegenerative disorders, including Parkinson's disease[67,68]; despite their transformed status as a cancerous cell line, HEK293T cells exhibit complex I-dependent respiration and have been used to interrogate mitochondrial function[69-71]; finally, HEK293T cells are readily grown and genetically manipulated with a very high transfection efficiency, beneficial for transduction of overexpression and knock-down genetic material as well as production of the large quantities of protein required for multiple enzyme assays.

Relative to overexpression of the EYFP control, it was found that overexpression of *CEROX1* resulted in increased specific activities of complex I (31.5%, $p = 6.4 \times 10^{-5}$) and of complex III (16.6%, $p = 2.7 \times 10^{-7}$) of the human ETC (Figure 5.2). No significant changes were observed in the activities of complexes II or IV, or in that of citrate synthase, a tricarboxylic acid cycle enzyme that is routinely used a control of basal mitochondrial number and activity, due to its correlation with mitochondrial content[72]. This suggested a specific effect upon complex I and complex III that is independent of cellular mitochondrial content, consistent with regulation at the level of subunit RNA or protein.

5.3.2 Preliminary gene expression data supported increased activity of complexes I and III

As a preliminary study of gene expression changes elicited by *CEROX1* overexpression in HEK293T cells, RT-qPCR was used to assay expression of the transcripts encoding subunits 25 of the 48 subunits of complex I, and all eight subunits of complex III. It was found that many complex I and complex III genes were subject to small but consistent increases in expression at the RNA level (Figure 5.3). While only one expression change was statistically significant (complex I *NDUFA7*, 30.0% increase, $p = 0.021$), other genes also showed expression changes of magnitudes consistent with those seen at the enzyme activity level. The largest gene expression changes were observed in the transcripts of three of the complex I assembly factors *NDUFAF1*, *NDUFAF2*, and *NDUFAF4*. These changes suggested that the effect of *CEROX1* upon complex I might be mediated by enhanced assembly.

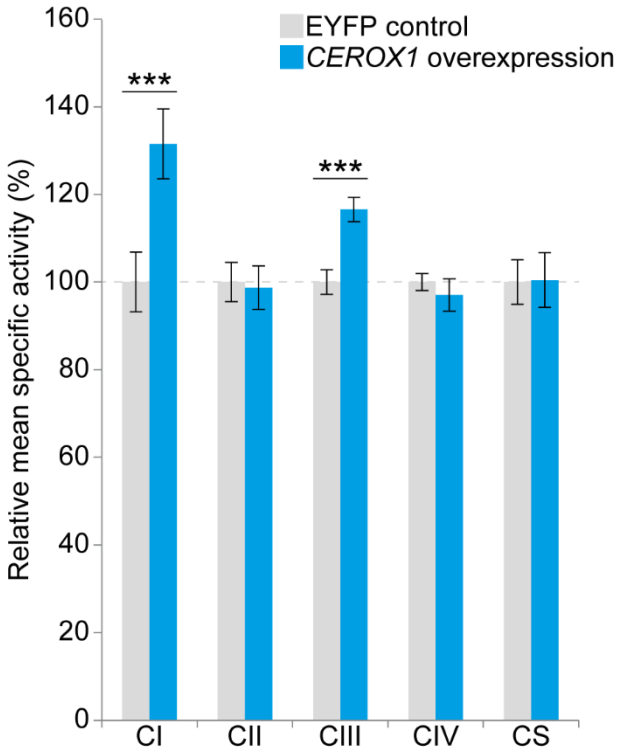


Figure 5.2: Overexpression of human *CEROX1* upregulated activities of ETC complexes I and III *in vitro*. Following overexpression of *CEROX1* in HEK293T cells, protein samples were assayed by spectrophotometric assays for enzyme activities of complexes I to IV (CI-IV) and citrate synthase (CS). Activities were normalised to total protein concentration, measured by BCA assay, in order to calculate specific enzyme activities.

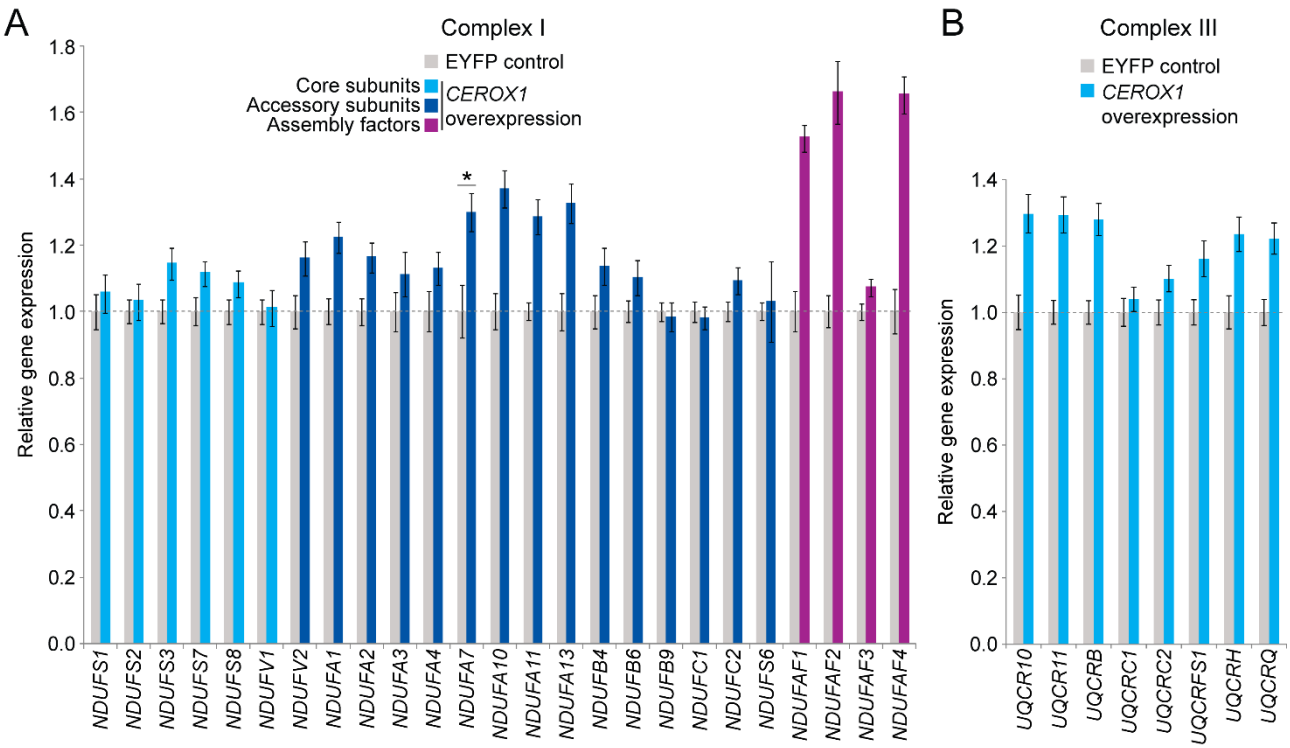


Figure 5.3: Gene expression changes underlie the changes in enzymatic activities of complexes I and III. Expression levels of **(A)** 25 complex I-related and **(B)** eight complex III subunit transcripts were assayed by RT-qPCR in HEK293T cells overexpressing *CEROX1*, relative to those overexpressing an EYFP control. Expression was normalised to that of α-tubulin *TUBA1A* and RNA polymerase II *POLR2A* genes.

5.3.3 The molecular mechanism of complex I upregulation appeared to be conserved between mouse *Cerox1* and human *CEROX1*

Both mouse *Cerox1* and human *CEROX1* appeared to upregulate complex I activity when overexpressed in cell lines of their own respective species backgrounds. It was hypothesised that the mechanism via which these phenotypes were achieved was also conserved at the molecular level, and that the two transcripts might therefore be biologically active in cells of the other species. In order to investigate this, human *CEROX1* was overexpressed in mouse N2A cells, while mouse *Cerox1* was overexpressed in human HEK293T cells, and the same specific enzyme activities were assayed as before. In accordance with the above hypothesis, human *CEROX1* increased mouse complex I activity by 12.0% ($p = 0.034$), while human complex I activity was upregulated by mouse *Cerox1* by 29.6% ($p = 3.8 \times 10^{-5}$) (Figure 5.4). With statistically significant regulation of complex I by human *CEROX1* and mouse *Cerox1* in both human and mouse cells, these data were indicative of *CEROX1* having an evolutionarily conserved role in the regulation of complex I, across the 100-90 My of divergence between human and mouse[73-75].

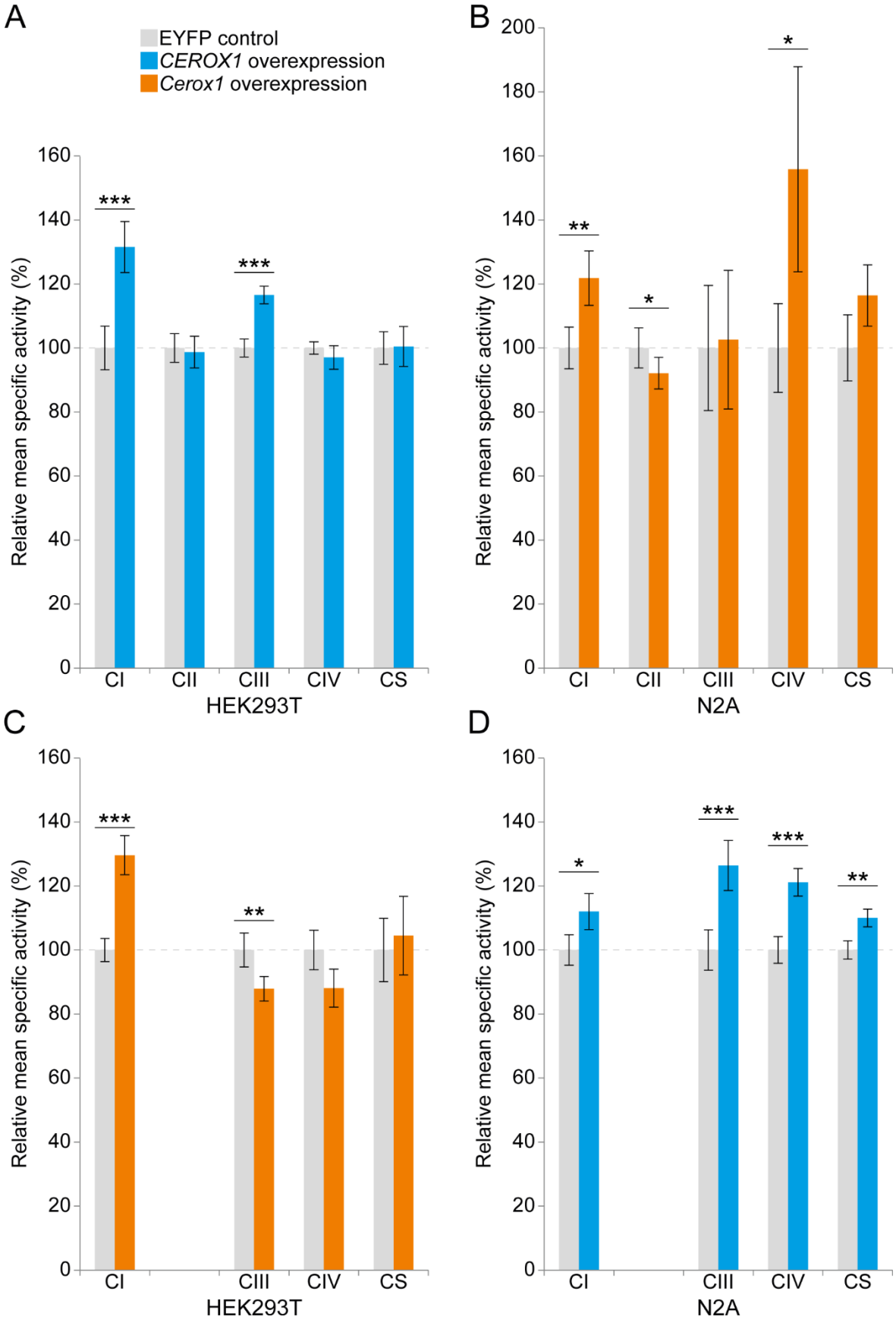


Figure 5.4: ETC complex-regulating activity of human *CEROX1* and mouse *Cerox1* in one another's respective cellular backgrounds suggests putative conservation of molecular mechanism. Shown for comparison, **(A)** and **(B)** illustrate the previous data obtained from overexpressing human *CEROX1* (Figure 5.2 above) and mouse *Cerox1* (Sirey et al, 2019[64]) in cells of their own respective species. Following overexpression of mouse *Cerox1* in human HEK293T cells **(C)** or of human *CEROX1* in mouse N2A cells **(D)**, specific enzyme activities of complexes I to IV of the electron transport chain (CI to CIV) as well as citrate synthase (CS) were assayed. Activities were assayed spectrophotometrically and normalised to protein content, measured by BCA assay. The data represent eight biological replicates and two technical replicates. Complex II (CII) activities have not yet been assayed.

5.3.4 *CEROX1* was hypothesised to be a putative conserved ceRNA

Given the evidence supporting the ceRNA activity of *Cerox1* (Sirey et al, 2019[64]), and the conserved biological activity of mouse *Cerox1* and human *CEROX1* in regulating complex I *in vitro* (Figure 5.4), it was hypothesised that human *CEROX1* also mediates its effects via a ceRNA mechanism. To this end two preliminary investigations were undertaken, into: a) the subcellular localisation of *CEROX1*; and b) the miRNA binding sites present in *CEROX1*.

5.3.4.1 Like mouse *Cerox1*, human *CEROX1* is enriched in the cytoplasm

Characterised ceRNA networks have thus far been predominantly cytoplasmic in localisation[76,77]. Therefore, as a putative ceRNA, *CEROX1* must be able to interact with the cytoplasmic miRNA pool. HEK293T cells were fractionated and the levels of *CEROX1* were assayed in the nuclear and cytoplasmic fractions by RT-qPCR, alongside the protein-coding gene *TUBA1A* and the known nuclear-retained lncRNA *MALAT1* (metastasis-associated lung adenocarcinoma transcript 1)[78], representing cytoplasmic and nuclear controls, respectively. *CEROX1* was observed to be highly enriched in the cytoplasm (5.6-fold), where 85% of the total transcript was found (Figure 5.5). This is consistent with the putative ability of *CEROX1* to interact with miRNAs in the cytoplasm, and act as a competitive endogenous RNA.

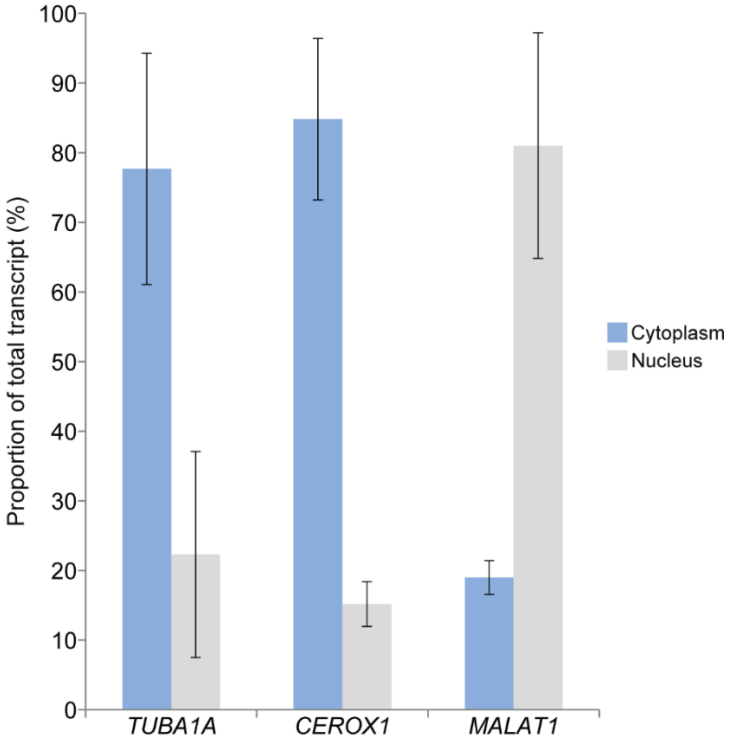


Figure 5.5: *CEROX1* is enriched in the cytoplasm of HEK293T cells. Following fractionation of HEK293T cells, the relative segregation of the transcripts of the protein coding gene *TUBA1A*, *CEROX1*, and the nuclear lncRNA *MALAT1* was assayed by RT-qPCR. As expected, *TUBA1A* was largely localised to the cytoplasm, while *MALAT1* was concentrated in the nucleus. *CEROX1* was found to be enriched in the cytoplasm.

5.3.4.2 Preliminary analysis suggested potential miRNA crosstalk between *CEROX1* and complex I

Preliminary prediction of MREs within human *CEROX1* and *Cerox1* (Prof. Ana Marques, Dr Jennifer Tan) identified 12 miRNAs or families with MREs shared by the two lncRNAs. Four of these were expressed in both HEK293T and N2A cells – miR-138-5p, miR-28/28-5p/708-5p, miR-370-3p, and miR-488-3p – and these were considered the best candidate mediators. Indeed, miR-488-3p, as mentioned above, was thereon demonstrated by Dr Tamara Sirey to mediate the *Cerox1* overexpression phenotype in N2A cells.

5.3.5 Difficulty in *CEROX1* knock-down impeded further phenotypic studies

Due to the powerful overexpression of *CEROX1* by the pCAGGS vector, which produces supra-physiological levels of RNA expression, it was desirable to assess the effects of *CEROX1* knock-down upon the ETC. A panel of five shRNAs were expressed in HEK293T cells in order to target the *CEROX1* transcript along its length but the degree of knock-down was highly variable and poorly reproducible (Figure 5.6A). Activity assays for complexes I and III were carried out following the expression of one of the shRNA constructs, but no significant changes in activity were obtained (Figure 5.6B). However, the poor knock-down efficacy and reproducibility casted doubt that this accurately reflected modulation of the *CEROX1* transcript.

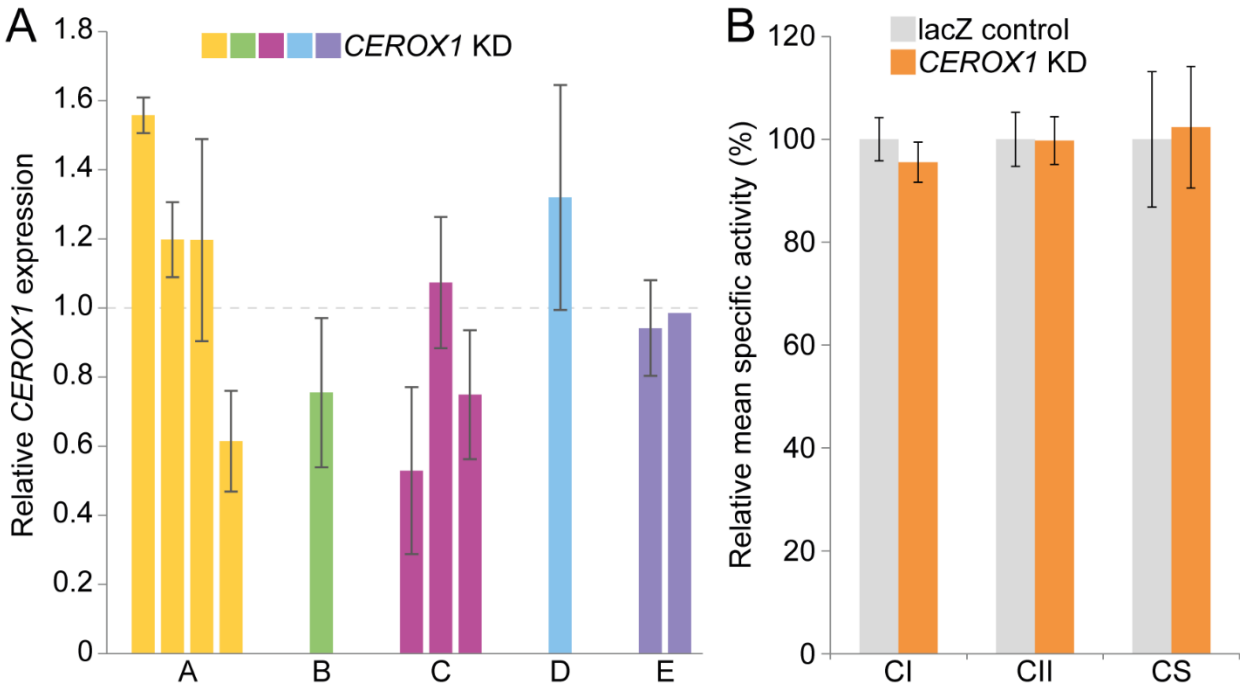


Figure 5.6: Inefficacious *CEROX1* knockdown produced no complex I or complex III phenotype. (A) Five U6 RNA polymerase III promoter-driven shRNAs (A-E) were expressed in HEK293T cells alongside a lacZ-targeting control. Levels of *CEROX1* knockdown, assayed by RT-qPCR and shown relative to the lacZ control, were largely variable or of poor efficacy. **(B)** Assay of the activities of complexes I and III (CI and CIII), and citrate synthase (CS) indicated no significant change in enzymatic activities following *CEROX1* knock-down.

5.3.6 Further investigation of the role of *CEROX1* and the underlying mechanism

Based upon these preliminary data, two routes of investigation were designed to further probe the role of *CEROX1* from mechanistic and metabolic perspectives:

A) A study of its miRNA-binding capacity, and the dependence of the overexpression phenotype upon the presence of: general miRNA-processing machinery such as Dicer, and specific miRNAs as predicted by the above networks. Tan et al.[79] and Karreth et al.[80] have demonstrated that manipulation of a ceRNA in a Dicer-deficient background removes its phenotype due to miRNA deficiency. TALEN-derived DICER-deficient HEK293T clones obtained from the Cullen lab (Duke University, US)[81] may therefore be used as a model in which to assess phenotype miRNA-dependency. Chemical inhibitors of Dicer[82] and Argonaute/RISC-loading[83] may be used to challenge the *CEROX1* phenotype in a non-miRNA-identity-specific manner. Owing to the poor false positive and false negative rates of *in silico* MRE prediction[84-86], physical approaches such as CLASH (cross-linking ligation and sequencing of hybrids)[87] are considered more reliable. CLASH has already been applied to a HEK293 cell line[88], but did not identify *CEROX1*-containing chimeric reads (David Tollervey, Aleksandra Helwak, personal communication). However, CLASH detects interactions with low efficiency[89], and HEK293 and HEK293T lines demonstrate differential miRNA expression[65,90]. Profiling miRNAs binding to not only *CEROX1*, but also ETC mRNAs, in the cell lines will aid in analysing interactions with the putative complex I target transcripts. Given that HEK293T cells are amongst the cells in which mitomiRs have been demonstrated[56], it would be prudent to identify any contribution to the putative mechanism; likewise, the possibility that MREs are shared with mitochondrially-derived lncRNAs such as *lnc-ND6*[54,91]. Noted also was the importance of obtaining an efficient and reproducible knock-down of *CEROX1* to complement overexpression studies.

B) An investigation of the downstream effects of complex I modulation. Overexpression of mouse *Cerox1* leads not only to increased complex I activity, but also increased oxygen consumption (indicative of overall increased electron transport) and yet reduced ROS production and oxidative protein damage. The central role of complex I within the cell suggested that *CEROX1* modulation might impact gross cellular metabolism. NADH and ATP, which feed into complex I and out of complex V, respectively, so encompassing the entire activity of the ETC pathway, connect the ETC to other major metabolic pathways including glycolysis, the TCA cycle, and β-oxidation. Indeed, the rate of glycolysis, pyruvate dehydrogenase, and the TCA cycle are highly dependent upon NADH/NAD^+ and ATP/ADP/AMP ratios[92,93] as both reaction participants and

allosteric regulators. Since *CEROX1* and *Cerox1* are the first described positive modulators of complex I, this study represents a novel route by which to investigate the potential of lncRNAs in regulating cellular metabolism. The experimental aims of this chapter are summarised in Figure 5.7.

Figure 5.7: The hypothetical *CEROX1* regulatory model at present, and further routes of investigation (overleaf). In its putative role as a ceRNA, *CEROX1* regulates complex I of the mitochondrial ETC. At **(A)** low levels and **(B)** higher levels, *CEROX1* differentially titrates miRNAs from the pool that regulates transcripts encoding complex I subunits. This alters their abundance at the RNA level and subsequent translation, furthermore modulating the enzymatic activity of complex I. **(C)** Identifying the miRNA interactions at the heart of this putative mechanism is of tantamount importance, and experiments proposed to address this aim during this chapter include the use of **(C')** chemical inhibitors of Dicer[82] and Argonaute/RISC-loading[83], **(C'')** mutation of predicted MREs, and **(C''')** investigation of physical miRNA-*CEROX1* interactions by CLASH. **(D)** On the other hand, the other proposed experiments in this chapter would address the downstream effects of regulation of complex I by *CEROX1*, including its role in regulating ROS production and damage, other metabolic pathways apoptosis, cell growth, and differentiation. miRISC – microRNA-induced silencing complex; ROS – reactive oxygen species; CLASH – cross-linking ligation and sequencing of hybrids.

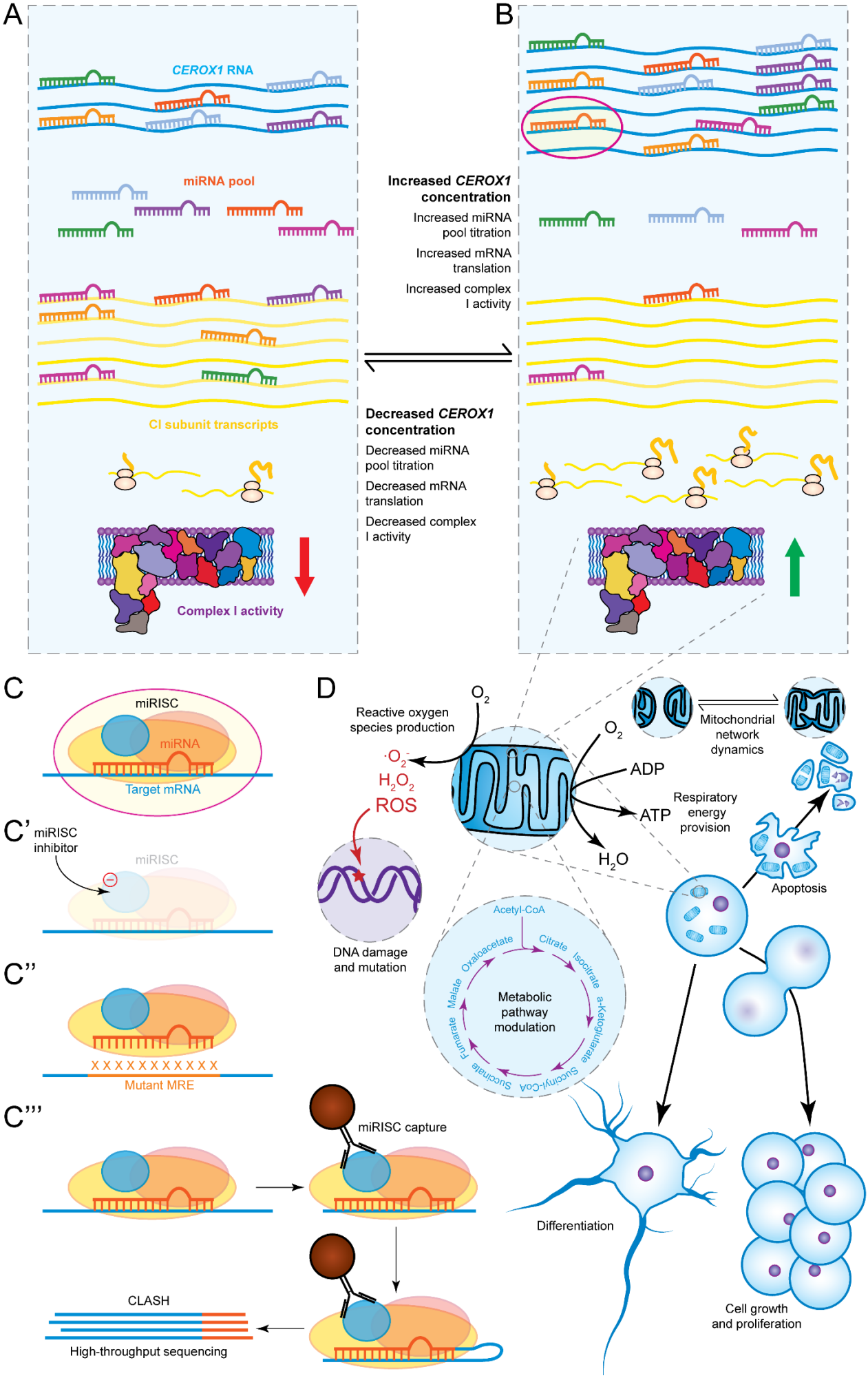


Figure 5.7: The hypothetical *CEROX1* regulatory model at present, and further routes of investigation.

5.4 *CEROX1* does not act via modulation of *SOX8* in this system

Owing to the high level of correlation of *CEROX1* and *SOX8* expression levels (Chapters 3 and 6), it was important to establish whether *CEROX1* may exert its effects by modulation of SOX8 levels. SOX8 protein was measured by western blotting following overexpression of *CEROX1*, and no significant change was observed (Figure 5.8).

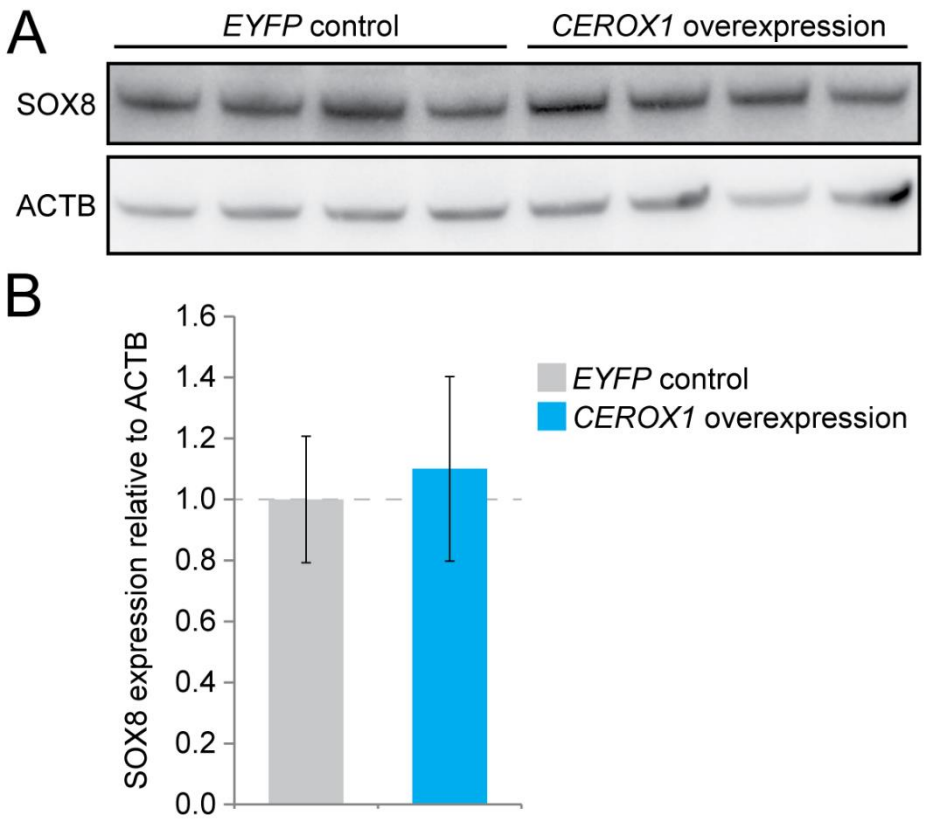


Figure 5.8: *CEROX1* overexpression does not affect SOX8 expression in the HEK293T system used here. Following overexpression of *CEROX1* RNA in HEK293T cells, protein levels of the SOX8 transcription factor were **(A)** measured by western blotting and **(B)** quantified by densitometry, normalised to β-actin (ACTB). No significant changes were observed.

In addition to the high levels of evolutionary conservation discussed in Chapter 3 as well as the apparent evolutionary conservation of molecular mechanism demonstrated above (Figure 5.4), this result strongly supported the hypothesis that human *CEROX1* may act as a ceRNA akin to mouse *Cerox1*, which formed the basis for the subsequent experiments hereon.

5.5 Human *CEROX1* contains many predicted miRNA binding sites

In order to assess whether *CEROX1* would likely interact with miRNAs that might elicit a repressive effect upon subunits of the ETC, miRNA response elements (MREs) within its sequence were predicted by Dr Oscar Bedoya-Reina using TargetScan Human v7.0[94]. Interestingly, this showed that human *CEROX1* does have a predicted binding site for hsa-miR-488-3p, the orthologue of the highly conserved mmu-miR-488-3p via which *Cerox1* mediates its activity (Sirey et al, 2019[64]). However, the site itself is not directly conserved between human and mouse, leading to speculation that a different miRNA might mediate a ceRNA functionality of human *CEROX1*. This would be consistent with the observations of Chen and Rajewsky[95], who demonstrated that MREs evolve rapidly and that functional miRNA-RNA relationships are less well conserved than the miRNAs themselves. In humans, *CEROX1* may utilise one of many primate-specific miRNAs[96].

Indeed, the *CEROX1* sequence contains 3674 TargetScan-predicted binding sites for 1509 miRNAs or miRNA families. While miRNA binding site prediction is notoriously difficult, complicated by high false positive and false negative rates[84-86,97], these predictions form a useful preliminary dataset from which more informed predictions may be refined. Indeed, they in fact include predicted binding sites for most of the miRNAs mentioned previously as characterised ETC transcript regulators: miR-210-3p/miR-147b, mir-338-3p, and miR-181c-3p. In order to identify possible candidates for the mediator(s) of a *CEROX1*-regulated ceRNA network, MREs were also predicted within the transcripts encoding the subunits of complex I, and these were intersected with those identified within the sequence of human *CEROX1*. As shown in Figure 5.9A, this led to an extremely complex hypothetical network of interactions: the complex I transcripts contain a median number of 1988 MREs per transcript, of which a median number of 1133 are for miRNAs for which *CEROX1* also has a predicted MRE (Figure 5.9B). Given the immense scale of this resulting possible *CEROX1*-miRNA-complex I network, further approaches were required to hone the predictions to those most consistent with the data already garnered and therefore most likely to be functionally relevant. The fact that human *CEROX1* and mouse *Cerox1* both demonstrated complex I-regulating activity in one another's respective miRNA backgrounds suggested that the causative miRNA(s) has a binding site in both transcripts.

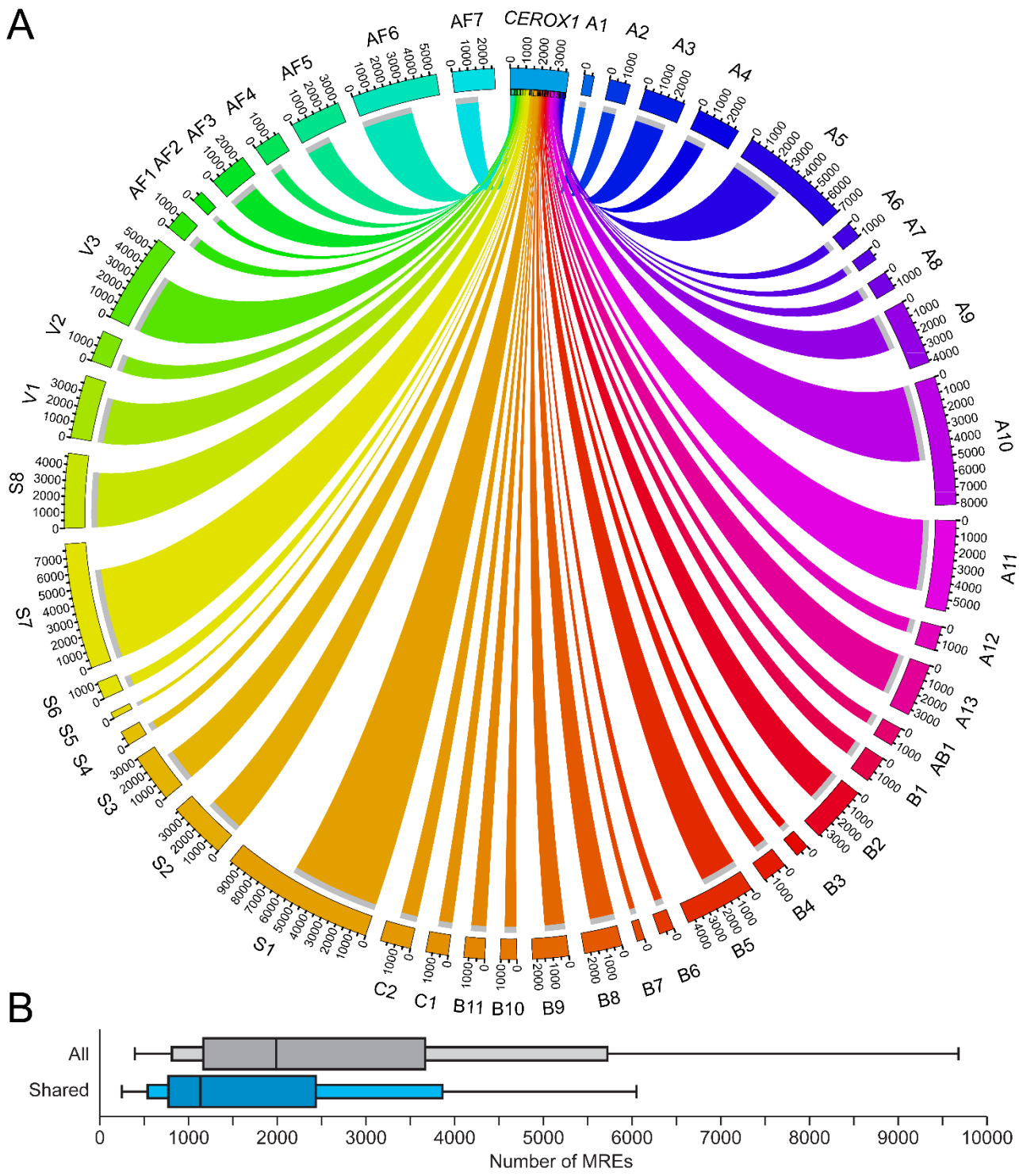


Figure 5.9: A dense network of putative interactions connects *CEROX1* to the transcripts encoding complex I subunits. (A) In this Circos plot (circos.ca)[98] each node represents a long RNA transcript, either *CEROX1* or the *NDUF* complex I subunit mRNAs encoding the 38 nuclear-encoded complex subunits and seven assembly factors. Ribbons represent shared MREs between *CEROX1* and a complex I mRNA, based upon TargetScanHuman v7.0 predictions[94]. The length of each node's bar is proportional to the total number of predicted MREs within a transcript; the width of each ribbon is proportional to the number of MREs shared with *CEROX1*. **(B)** The box plot shows the distribution of MRE count within the 45 complex I-related mRNAs illustrated in the Circos plot, for all MREs, and for MREs shared with *CEROX1*.

The list of miRNAs was therefore further refined by identification of 84 miRNAs for which MREs are found in both human *CEROX1* and mouse *Cerox1*. For these miRNAs, a median number of 65 MREs were shared between human *CEROX1* and each complex I transcript (Figure 5.10).

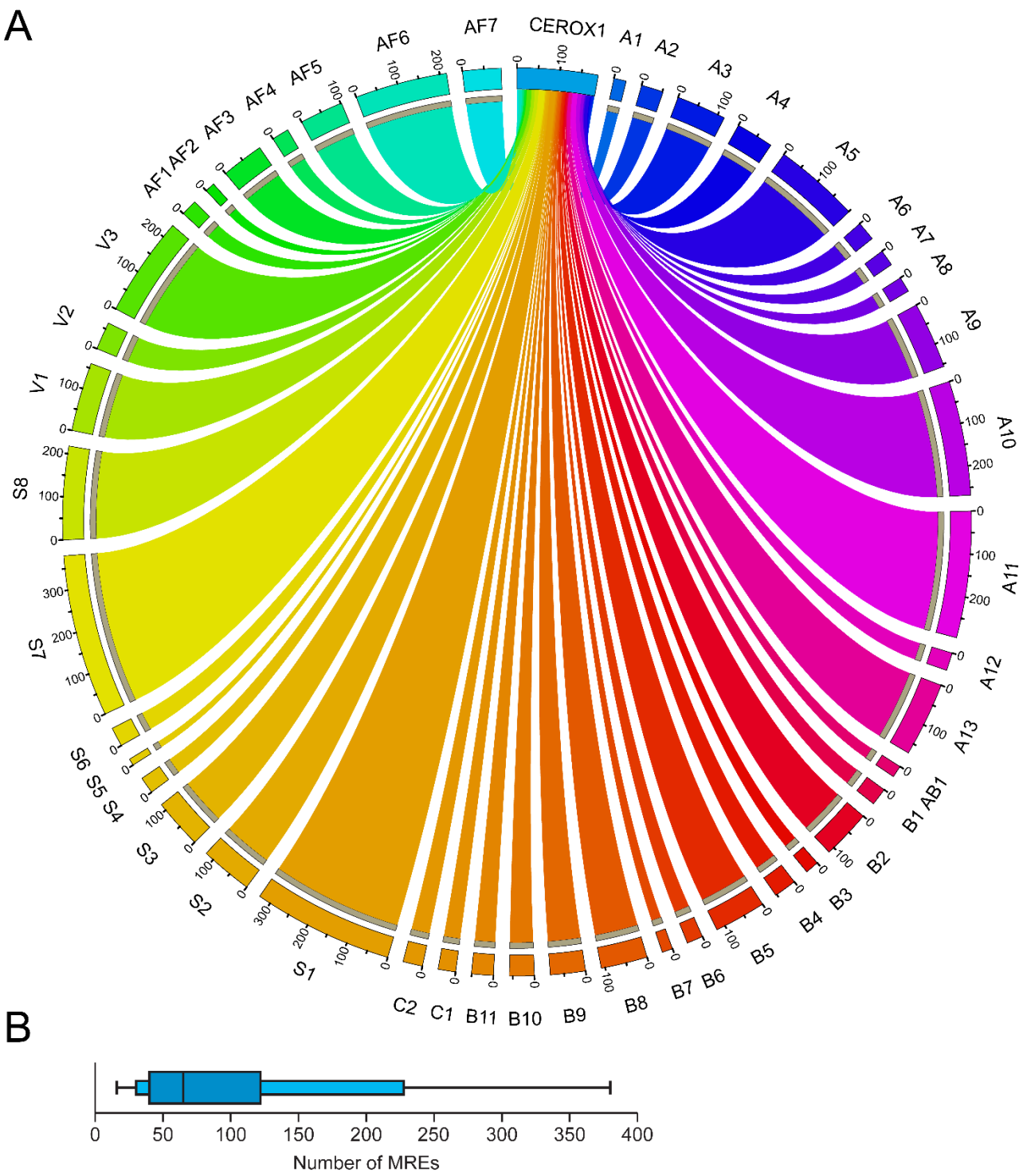


Figure 5.10: Intersection of *CEROX1-Cerox1* and *CEROX1*-complex I shared MREs provides an approach to narrow down the investigation into causative miRNAs. (A) In this Circos plot (circos.ca)[98], each node represents a long RNA transcript, either *CEROX1* or the *NDUF* complex I subunit mRNAs encoding the 38 nuclear-encoded complex subunits and seven assembly factors. Ribbons represent shared MREs between *CEROX1* and a complex I mRNA, based upon TargetScanHuman v7.0[94] predictions, for only the 84 miRNAs that have a predicted binding site in *CEROX1* and whose mouse orthologues have a predicted binding site in *Cerox1*, predicted using TargetScanMouse v7.1. **(B)** The box plot shows the distribution of MRE count within the 45 complex I-related mRNAs illustrated in the Circos plot, for the 84 miRNAs whose human-mouse orthologue pairs have predicted binding sites in human *CEROX1* and mouse *Cerox1*, respectively.

5.6 Experiments designed to further interrogate the *CEROX1* overexpression phenotype have shown it to be **poorly reproducible**

5.6.1 A *CEROX1* mutant transcript was produced to identify functionally-relevant elements

As illustrated in Chapter 3, the mouse *Cerox1* locus has been subject to a large deletion that has removed much of the sequence comprising the orthologous 5' exon in human *CEROX1*. Consequently, it was hypothesised that this region, which largely comprises a Herpes viral remnant, might be non-essential for the putative complex I-regulating function of *CEROX1*. In order to test this, and thereby eliminate a subset of possible MRE mediators, this region was deleted from the *CEROX1* overexpression construct by mutagenesis, and the complex I assay repeated in HEK293T cells. Initial data indicated no significant difference between sample means (1-way ANOVA $F(2,21) = 2.165$, $p = 0.14$) unexpectedly meaning that even the positive control, the overexpression of the non-mutated *CEROX1* transcript, did not elicit a change in complex I activity above the EYFP control (Figure 5.11A).

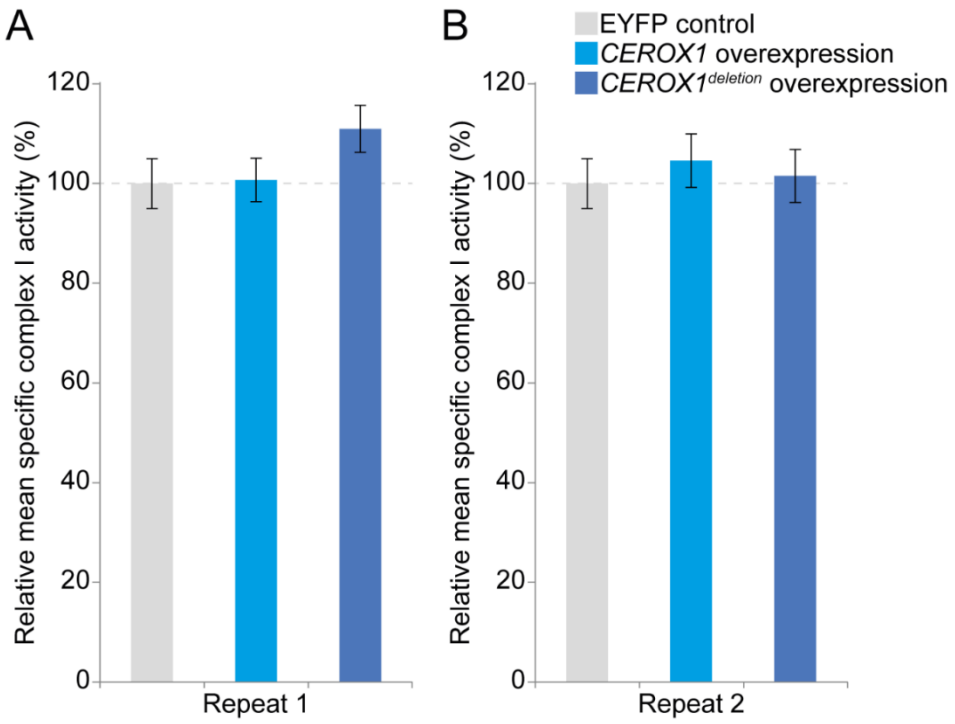


Figure 5.11: Complex I assays designed to ascertain the functional importance of a deleted region revealed non-reproducibility of the *CEROX1* overexpression phenotype. In both replicates of the experiment, overexpression of the positive control non-mutated *CEROX1* transcript (#) did not significantly upregulate complex I activity above the EYFP control. In **(A)** eight biological replicate samples were generated for each construct, and two technical replicates assays carried out for each. In **(B)**, nine biological replicate samples were generated for each construct.

In light of the apparent robustness of the evolutionarily conserved complex I phenotype in previous experiments, it was assumed that this was likely due to an anomaly in sample preparation, and the experiment was repeated. However, in a second experiment, overexpression of the wildtype *CEROX1* and the viral deletion again resulted in no significant complex I phenotype (Figure 5.11B) (1-way ANOVA $F(2,24) = 0.309$, $p = 0.687$).

5.6.2 Thorough examination of experimental procedures did not identify the reason for the loss of phenotype

Although there were no grounds upon which to suspect a specific problem in the experimental protocol, each step of the protocol was assessed for possible sources of the non-reproducible results.

No issue was identified in the growth of cells during sample preparation or otherwise. Conditions for cell growth remain unchanged, cell growth and viability are not overtly different, and testing indicates no contamination of Mycoplasma. Cell passage is routinely limited to ~25, after which a new stock is revived according to standard procedure and allowed to grow for several passages prior to experimentation.

HEK293T cells were originally chosen for this investigation in part due to their excellent transfection efficiency, and several readouts were employed in verifying that this remains unchanged. First, the control vector encodes EYFP, and the bright fluorescence of the resulting samples is testament to their continued uptake of the transfected constructs. Integrity of the *CEROX1* overexpression construct was verified using Sanger sequencing throughout the length of the insert. Furthermore, RT-qPCR indicates that the *CEROX1* overexpression construct still results in increased *CEROX1* transcript expression. Figure 5.12 shows the highly consistent magnitude of *CEROX1* overexpression in five experiments, one from the period in which a complex I phenotype was observed, and the other four since the phenotype was lost.

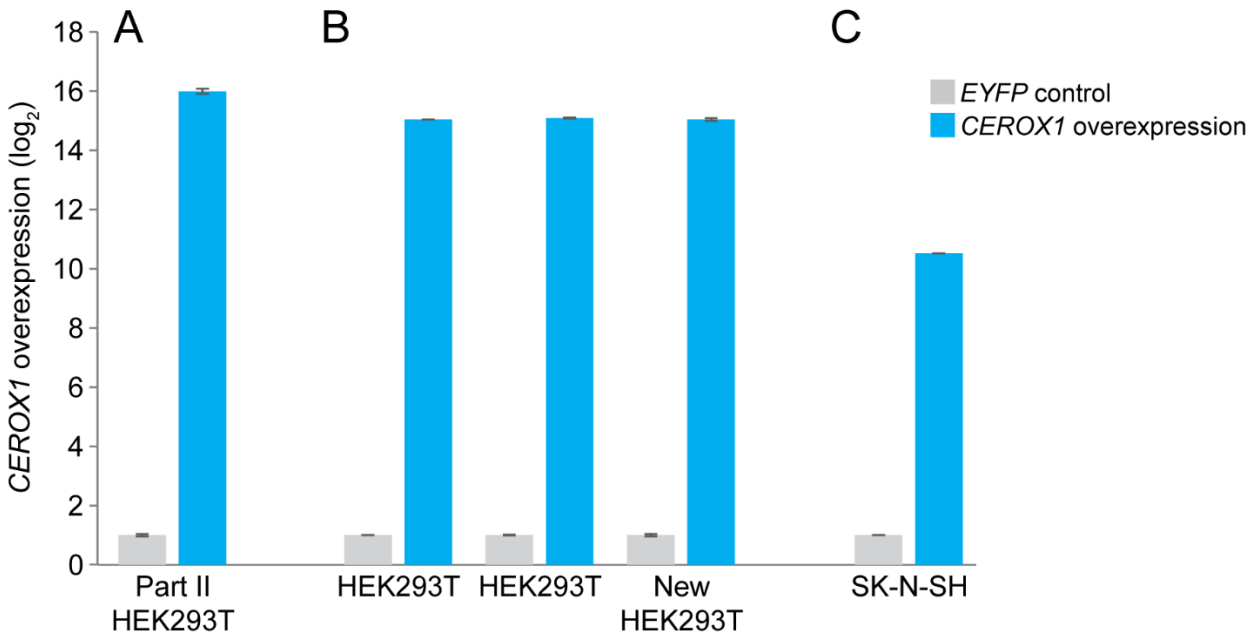


Figure 5.12: *CEROX1* overexpression is potent and consistent. Akin to the level of *CEROX1* overexpression observed in HEK293T cells when the original enzyme assay data were collected **(A)** (see Figures 5.2-5.4), continued overexpression experiments (Figure 5.11) yielded a very high level of overexpression (around 16,000-fold) in HEK293T cells **(B)**. **(C)** Mean overexpression levels observed in SK-N-SH cells were lower, though still substantial at over 1,000-fold, owing to their relatively poor transfection efficiency. All *CEROX1* levels were assayed by RT-qPCR, and quantified by normalisation to the geometric mean of the expression levels of *TUBA1A* and *POLR2A*.

No overt issue was identified during data collection by the complex I assay, with spectrophotometric absorbance changing as expected at rates within a range consistent with previous experiments. The complex I assay was additionally validated using a sample of isolated heart mitochondria (Matthew Kerr). Sensitivity of the assay to the characterised complex I inhibitor rotenone was experimentally determined to be 60-80%, consistent with previous experimental data.

Owing to the integrity of the assay and the preparation of samples, it was suggested that an unidentified issue with the cell line was the source of the altered phenotype. In order to establish any role of this parameter, a new stock of HEK293T cells was obtained from Public Health England (ECACC 12022001). *CEROX1* overexpression followed by complex I assay revealed no complex I phenotype in these cells (Figure 5.13A). Additionally, the assay was repeated in a different laboratory, in the IGMM, University of Edinburgh, using another line of HEK293 cells and entirely separate batches of reagents. This time, a decrease in complex I activity was observed in *CEROX1*-overexpressing cells (11.5% decrease, $p = 0.037$) (Figure 5.13B). In order to absolutely rule out a cell type-specific anomaly, overexpression of human *CEROX1* and mouse *Cerox1* was repeated in N2A mouse cells. No change in complex I activity was observed following overexpression of human *CEROX1*, while surprisingly mouse *Cerox1* overexpression elicited a decrease in complex I activity of 19.2% (1-way ANOVA $F(2,12) = 3.936$, $p = 0.0275$; Dunnett's post-hoc test for *Cerox1* overexpression $p < 0.05$) (Figure 5.13C). This suggested a fundamental experimental or biological issue that is not limited to a single construct or cell type.

Finally, *CEROX1* was overexpressed in two other cell lines: the neuroblastoma line SK-N-SH and their sub-clonal line SH-SY5Y. Activity assays of complexes I, III, and IV were carried out upon *CEROX1*-overexpressing SK-N-SH cells, and *CEROX1*-overexpressing SH-SY5Y cells were assayed for complex I activity. Despite a moderately increased complex IV activity level in *CEROX1*-transfected SK-N-SH cells, no statistically significant changes were observed (Figure 5.14). This may, however, be due to the lower transfection efficiency of these cells (~30% compared to >80% for HEK293T cells), which yields a lower level of overexpression (Figure 5.12).

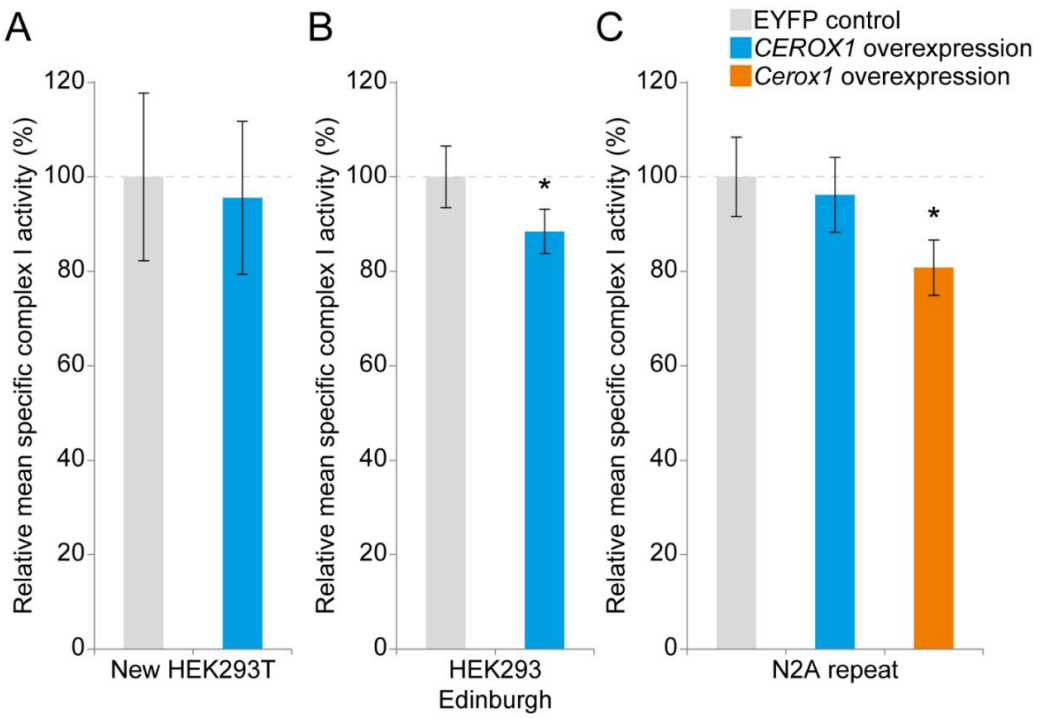


Figure 5.13: Repeated overexpression of human *CEROX1* and mouse *Cerox1* in HEK293T and N2A cells did not recapitulate the original results of these experiments. (A) Overexpression of human *CEROX1* in new HEK293T cells from the European Collection of Authenticated Cell Cultures did not result in a complex I phenotype. **(B)** Repeat of the experiment in a different facility at the University of Edinburgh indicated another different result, with a significant decrease in complex I activity. **(C)** Expression of human *CEROX1* and mouse *Cerox1* in mouse N2A cells resulted in no effect, and a significant decrease, in complex I activity, respectively, in contrast to previous results (Figure 5.4). All datasets represent averages of eight biological samples and two technical replicates.

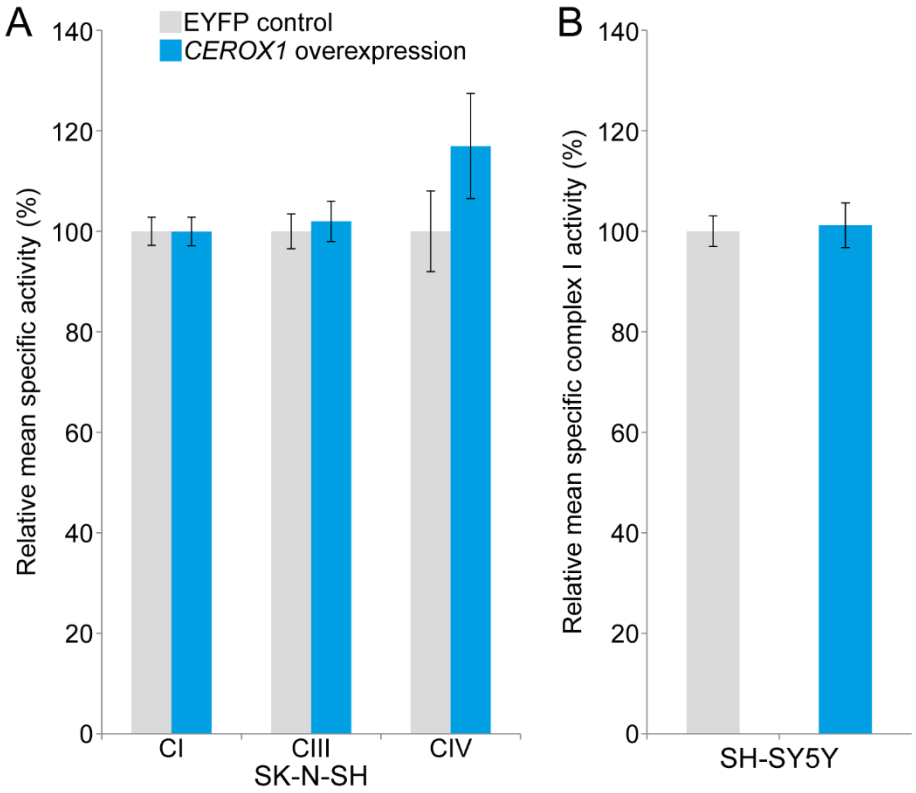


Figure 5.14: *CEROX1* overexpression resulted in no ETC phenotype in two other neuroblastoma human cell lines. Overexpression of *CEROX1* in **(A)** SK-N-SH and **(B)** SH-SY5Y neuroblastoma cell lines produced no statistically significant change in complex I, III, or IV activities. All assays were carried out using eight biological replicate samples. SK-N-SH assays incorporated two technical repeat measurements.

5.6.3 Native page was used as a secondary assay to confirm the alteration of phenotype

The data obtained from the spectrophotometric complex I assay were, between-group phenotype excepting, consistent with those obtained previously. However, a secondary assay was devised in order to confirm that the apparent absence of complex I phenotype was not a product of the assay or its preparation. Mitochondria were isolated from control and *CEROX1*-overexpressing HEK293T cells and protein was extracted for blue native PAGE (BN-PAGE). Following electrophoresis, complex I and three other complexes (II, IV, V) of the ETC were subjected to in-gel activity assay staining. Complex III staining was trialled but not achieved. The same pattern of bands, including six ETC complex forms and seven super-complexes, was observed in each group (representative lanes shown in Figure 5.15A). The smallest non-artefact complex I-stained band likely represents SCI_1III_2; the other five bands include CIV. Therefore, all or almost all complex I in this system is incorporated into super-complexes, while substantial levels of complex IV remain isolated in the monomeric form, in accordance with published data[28,30].

Following BN-PAGE and ETC complex staining of control and *CEROX1*-overexpressing HEK293T cells (Figure 5.15B), activities were assessed by densitometry. There were no statistically significant changes in ETC activity in the *CEROX1*-overepressing cells relative to the EYFP control following application of a Benjamini-Hochberg multiple testing correction (Figure 5.15C). However, all measured activities exhibited non-statistically significant decreases: complex I including all super-complexes, 13.1% decrease; complex II, 17.0% decrease; complex IV monomer, 11.1% decrease; complex IV super-complexes, 14.3% decrease; complex V, 10.1% decrease. No statistically significant change was seen in the proportion of staining exhibited for each band within a set of super-complexes, suggesting no altered stoichiometric equilibria. Contrary to all preliminary data, but consistently with some more recent data, these results suggest a concerted decrease in ETC activity following *CEROX1* overexpression.

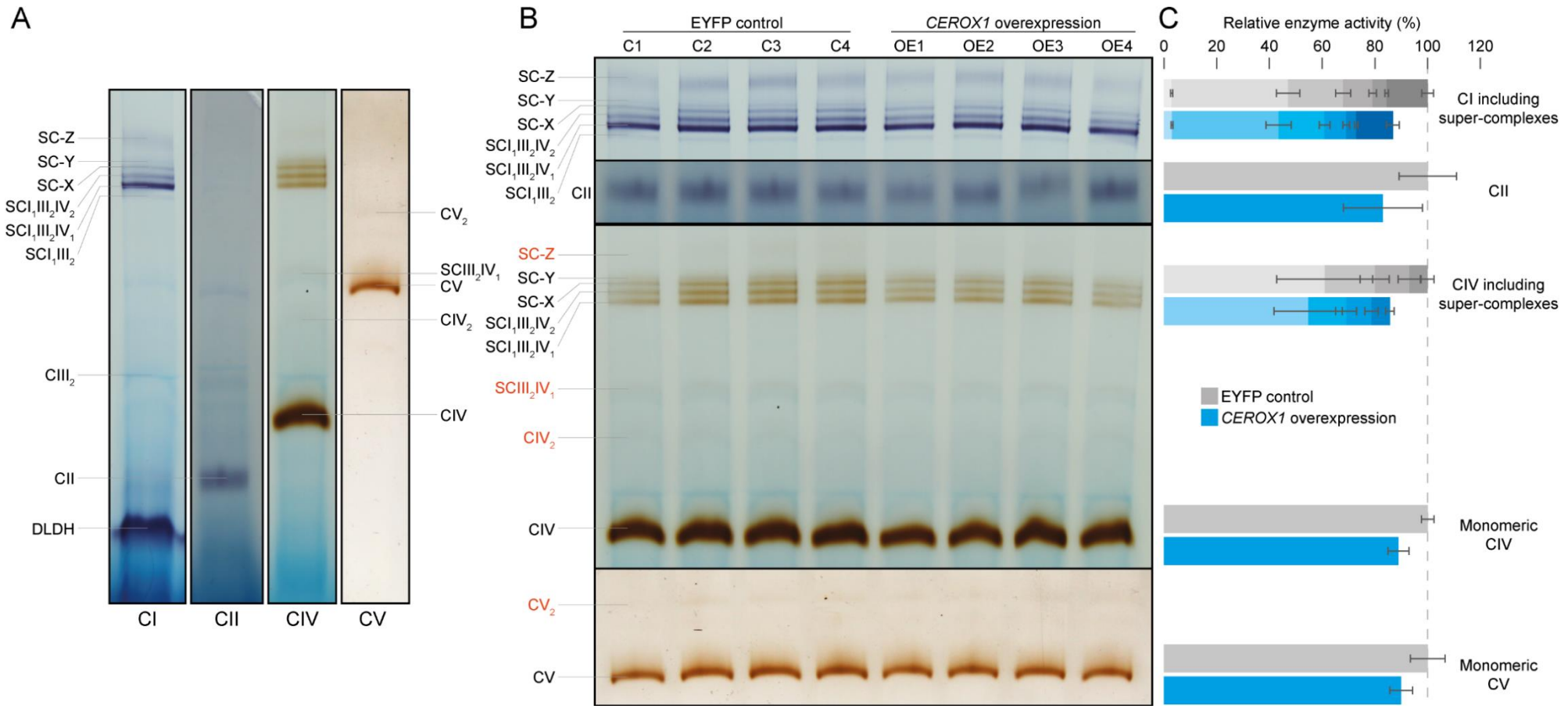


Figure 5.15: The activities of four ETC complexes were assessed by BN-PAGE following *CEROX1* overexpression in HEK293T cells. (A) Following isolation of mitochondria from HEK293T cells, protein was extracted using digitonin (8 g g^{-1}), and electrophoresed on four 3-8% NativePAGE gels (Invitrogen), each of which was then stained for a different complex I, II, IV, or V. As well as the complex I-containing super-complexes, complex I staining also stains another enzyme with diaphorase activity, dihydrolipoamide dehydrogenase (DLDH)[99], and appears to very weakly mark complex II. Complex II is visible by enzymatic staining. Complex III activity staining was unsuccessful, but the protein of the complex III dimer is visible by residual Coomassie staining. Complex IV staining highlights the dominant monomeric form, as well as the complex IV dimer and a number of super-complexes. Complex V staining stains both the monomeric and dimeric forms of complex V. At least seven super-complex forms are present: $SCIII_2IV_1$ is stained by complex IV; SCI_1III_2 is stained by complex I but not complex IV; $SCI_1III_2IV_1$ and $SCI_1III_2IV_2$ (respirasomes) are stained by both complex I and complex IV. Three further super-complexes (X, Y, and Z) cannot be unambiguously identified, but are all stained for complex I and complex IV; accordingly, one likely represents the mega-complex $MCI_2III_2IV_2$. **(B)** Four control and four *CEROX1*-overexpressing samples were generated and assayed simultaneously, then activities were quantified by densitometry **(C)**. Due to faint and diffuse staining, several activities were not assessed by densitometry (shown in red): CIV activity in dimeric CIV_2, $SCIII_2IV$, and SC-Z; and CV activity in dimeric CV_2.

5.6.4 The results of a third assay of respiratory activity are more consistent with the original *CEROX1* overexpression phenotype

All assays of ETC function and quantification thus far have utilised cell lysates, removing components of the ETC from their physiological environment. Although BN-PAGE seeks to retain higher order organisation of the ETC, the complexes are removed from the mitochondrion, and are no longer coupled to endogenous metabolic pathways or ATP synthesis. In order to assess the activity of the ETC as a complete system, the respiratory activity of HEK293T cells overexpressing *CEROX1* were assayed using Agilent Seahorse Extracellular Flux Analyzer technology (Dr Tamara Sirey) (Figure 5.16). Oxygen consumption was significantly altered in *CEROX1*-overexpressing cells under several respiratory states, with basal respiration, ATP-linked respiration, and maximal respiration all increased by 30-35% (p = 0.057, p = 0.094, p = 0.031, respectively). Lack of statistical significance at the p < 0.05 threshold was deemed likely due to the small sample size. These results support the original presence of an ETC phenotype.

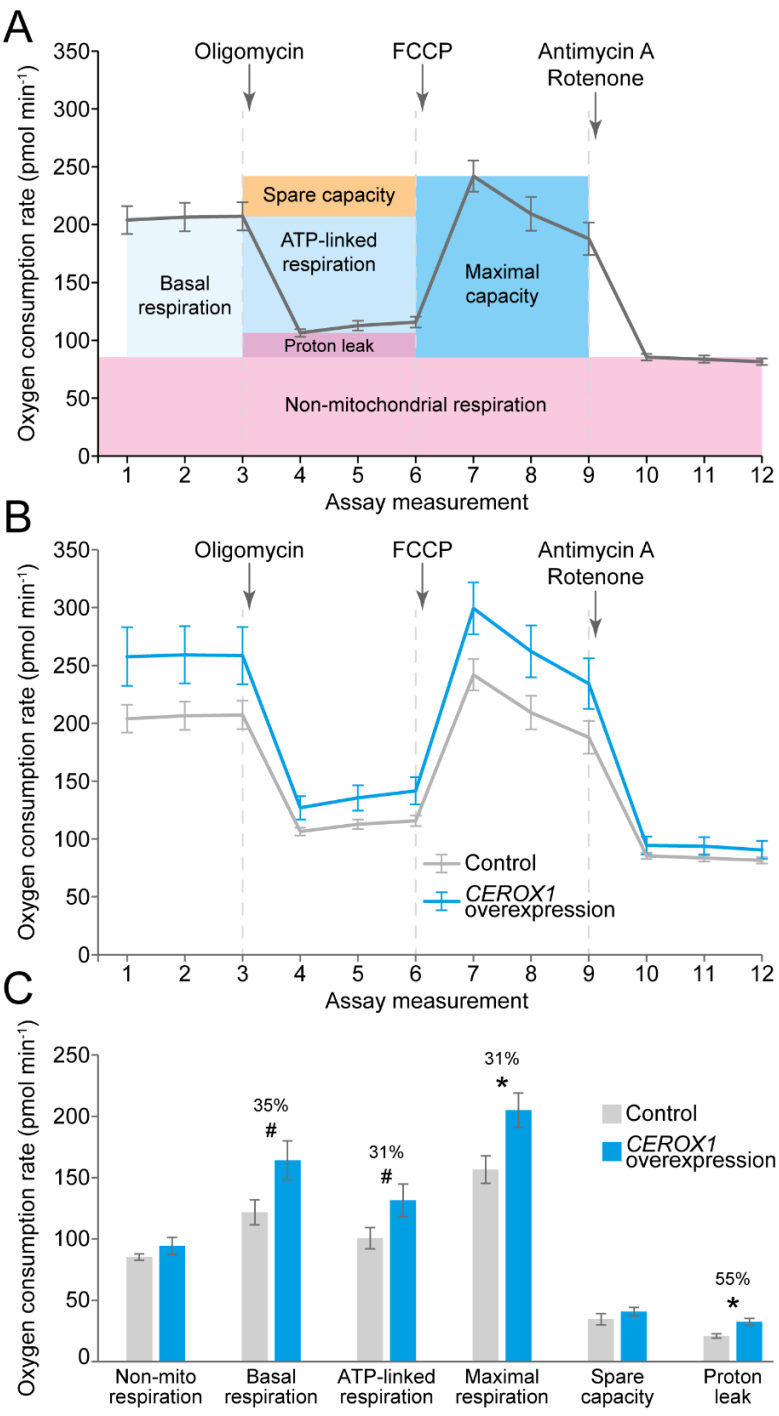


Figure 5.16: Analysis of live HEK293 cells supports upregulation of ETC activity following overexpression of *CEROX1*. (A) Using a sequential combination of respiratory inhibitors (oligomycin inhibits complex V; FCCP uncouples the proton gradient; antimycin A and rotenone inhibit complexes III and I, respectively), the Agilent Seahorse Extracellular Flux Analyzer quantitatively assesses the activity of the ETC, permitting the calculation of basal, ATP-linked, and maximal respiration, as well as spare respiratory capacity and proton leakage. **(B)** Respiratory traces for control and *CEROX1*-overexpressing HEK293 cells, and quantification **(C)**. (# p < 0.1; * p < 0.05). The data represent the results of one experiment, with n = 7 for control cells, and n = 5 for *CEROX1* overexpression.

5.6.5 RNA and protein expression of complex I subunits also supports an altered phenotype

With the original enzymatic complex I phenotype apparently no longer present in *CEROX1*/*Cerox1*-overexpressing cells, it is important to assess whether other previously identified phenotypes are also absent. To this end, levels of gene expression and protein expression of complex I subunits were also re-assayed. The RNA levels of seven complex I-associated genes, four subunits and three assembly factors, which previously exhibited non-statistically significant but reproducible changes were re-assessed by RT-qPCR. No changes in expression were seen between the EYFP control- and *CEROX1* overexpression-transfected cells (Figure 5.17).

Changes in complex I subunit protein levels have previously been observed in N2A cells overexpressing mouse *Cerox1* (Sirey et al, 2019[64]). Protein samples prepared for the N2A complex I assay above (Figure 5.13C) were analysed by western blotting and densitometry and compared to previous data (Figure 5.17B-C). While a change in average expression of Ndufs3 was similar in magnitude to previous data (97% increase) cellular expression levels were variable and the change was not statistically significant, unlike in previous data (Figure 5.17C). No change reminiscent of previous data was observed for Ndufs1.

Protein levels of complex I subunits had not yet been carried out for *CEROX1*-overexpressing cells during a time when the enzymatic activity phenotype was observed. However, the data of Sirey et al. support a model whereby *Cerox1*-mediates regulation at the RNA level that is subsequently transduced into changes in enzymatic activity via altered protein levels[64]. Consequently, it was hypothesised that changes in levels of subunit proteins in this human cell system should also correlate with the levels of RNA transcripts, and so the levels of several complex I subunit proteins were assayed by western blotting in HEK293T cells overexpressing *CEROX1*. Small changes were observed in protein expression of some of these subunits, but not all (Figure 5.17D-E). Only the changed expression of the NDUFA10 protein subunit was statistically significant (25.4% increase, $p = 0.0067$).

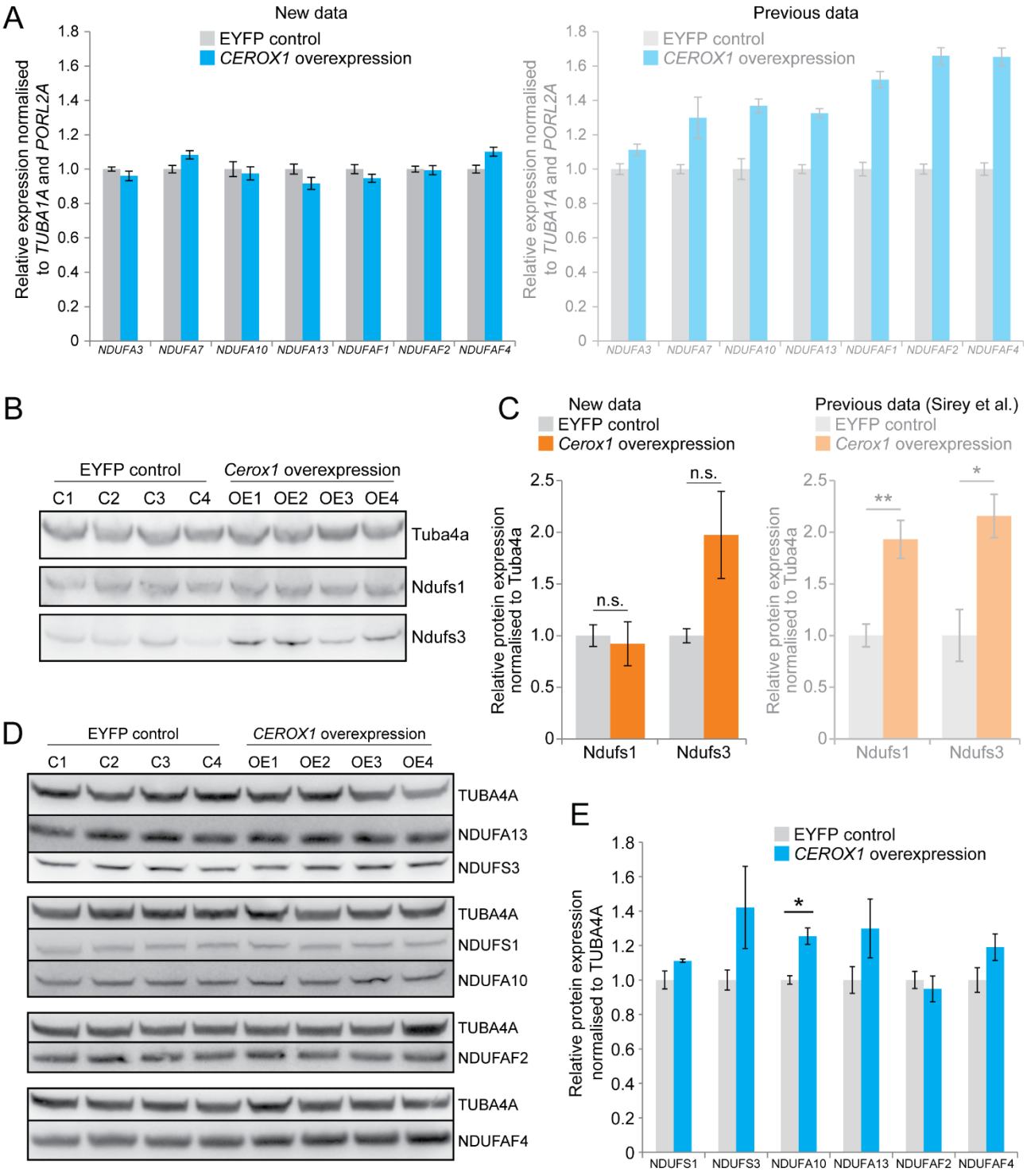


Figure 5.17: Altered RNA and protein changes in response to *CEROX1* overexpression. (A) Unlike original gene expression data (right), which indicated consistent if statistically non-significant changes in RNA levels of 10-60%, more recent RT-qPCR (left) shows little to no difference in subunit gene expression. **(B)** Following overexpression of mouse *Cerox1* in N2A cells, complex I subunit protein levels were analysed by western blotting and densitometry **(C)**, which indicated no statistically significant recapitulation of previous data (right; Sirey et al, 2019[64]). **(D)** Protein subunit levels in HEK293T cells following human *CEROX1* overexpression were also assayed by western blotting and then quantified by densitometry relative to TUBA4A **(E)**.

DISCUSSION

The respiratory analysis that supports the original putative hypothesis was carried out late during this project. In the absence of a stable phenotype upon modulation of *CEROX1*, it was no longer tenable to continue experiments to investigate its putative miRNA-regulating role. The expense and low efficiency of high-throughput approaches such as CLASH, and the time-consuming nature of sequential mutagenesis of MREs within the transcript place these approaches outside the scope of this project, given the now-contentious evidence for the biological activity of *CEROX1* in the model systems trialled here. Likewise, enquiries into any downstream effects of ETC modulation by *CEROX1* are of limited, at most, meaning in the context of a poorly reproducible phenotype.

Gene ontology analysis of the original list of genes differentially expressed upon *Cerox1* overexpression indicated several other enriched categories, though none so significantly over-represented as the mitochondrial ETC. These included mitochondrial ribosome components, and several regulators of the cell cycle. Flow cytometry indicated that *Cerox1*-pertrubation effected no change upon overall N2A cell division, though cell growth is slightly reduced (Sirey et al, 2019[64]); no reproducible effect of *CEROX1* overexpression upon cell growth was observed in HEK293T cells here. It was hypothesised that the upregulation of mitochondrial translation might follow perturbation of ETC-regulating miRNAs but an investigation into the mitochondrial ribosome was deemed a) beyond the scope of this project, and b) of unknown relevance until the original mitochondrial phenotype was clarified.

While the failure of *CEROX1* overexpression to modulate SOX8 expression in the *in vitro* system of this Chapter does not preclude the possibility that expression levels of these genes may regulate one another *in vivo*, it does indicate that the *CEROX1* overexpression phenotype seen here is not mediated by a regulatory effect in *cis*.

Why has the *CEROX1* overexpression complex I assay phenotype, which was previously robust, disappeared? While some still consider the biological significance of the ceRNA mechanism contentious[100], and it has seldom yet been thoroughly demonstrated *in vivo*[77], doubt into the mechanism by which *CEROX1* acts does not address the conundrum posed by the data of this chapter: given that HEK293T cells and N2A cells previously hosted a robust phenotype, be it ceRNA-mediated or not, why do the same cells no longer demonstrate the same phenomenon? Indeed, even after the original 'disappearance' of the complex I

enzymatic phenotype, the responses to *CEROX1* overexpression are still inconsistent. Likewise, while aspects of HEK293T cell models are sub-optimal for modelling normal physiology, this again does not address the reproducibility problems encountered here.

Several hypotheses have been considered: a) the original phenotype observed in HEK293T cells represented biological noise; b) the alteration of phenotype is due to clonal differences in the cell lines used; c) the phenotype is dependent upon a previously unexamined parameter that remains uncontrolled.

a) Given the crucial nature of the ETC, it seems highly unlikely that a change in enzymatic activity of 10-35% would simply be permissible due to stochasticity. Indeed, alterations in ETC complex activity of this magnitude range in the opposite direction have been associated with disease states: Schapira et al.[101] and Keeney et al.[102] both showed a 30% decrease in complex I activity in the brains of Parkinson's disease patients, while a 36% decrease in complex IV activity was observed in a β-amyloid-treated mitochondrial model of Alzheimer's disease[103]. In a myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis (EAE) model of multiple sclerosis, a selective 10% decrease in complex I activity was observed in the spinal cord in association with the onset of neurological symptoms[104], alongside mitochondrial fragmentation and loss of mitochondrial membrane potential. In the striatum of Huntington's disease-affected brains, selective 30-70% decreases have been observed in complexes II, III, and IV[21,23,105]. Crucially, as described above, Alvarez-Fischer et al. showed that a 20% increase in complex I activity mediated by exogenous Engrailed was sufficient to protect midbrain dopaminergic neurons from the potent Parkinson's disease model agent MPP^{+}[36]. So, while the magnitude of effects seen may seem minor compared to the many-fold gene or activity differences present upon perturbation of other biological systems, the importance and high basal activity of the ETC grants even small changes substantial importance. Alongside the initial reproducibility of the phenotype, this strongly suggests that the original data obtained following *CEROX1* overexpression were not due to biological noise. Additionally, preliminary analysis of respiratory function using the Agilent Seahorse system supports the original phenotype identified. Furthermore, the evolutionary conservation demonstrated in the *CEROX1* locus and transcript strongly supports the notion of molecular conservation of the activities and mechanism of the more thoroughly characterised mouse *Cerox1*. The unique nature of the role to which we have putatively ascribed *Cerox1* and the magnitude of responses observed at the RNA, protein, and enzymatic activity levels make assay validation difficult. No other monogenic perturbations that result in 10-35% changes in ETC complex

activity that might have been used as controls in verifying small responses of the assays have been described. As evidenced by studies of complex assembly pathways as well as the plethora of mitochondrial diseases, inhibition or loss of even single subunits profoundly impairs the activity of the entire complex[6], and chemical inhibitors of the ETC such as rotenone are typically extremely potent.

b) If indeed *CEROX1* is a ceRNA, a hypothesis that was un-contested until the observation of the later data presented in this chapter, then its activity is dependent upon specific miRNAs. If the miRNA pool were to be altered, then the phenotype would also. As an immortalised cancerous cell line that has been under dis-contiguous culture for almost half a century, HEK293 cells, including HEK293T[106] and other derivatives, have been subject to extensive mutation since they were last in a physiological environment. The inhibition of p53 by the SV40 T antigen[107] present in the HEK293T lineage[106,108] has resulted in an accelerated level of genome rearrangement relative to the parental line[65]. This includes, for example, 8-fold amplification of the *MIR17HG* locus, which encodes three miRNAs miR-17, miR-20a, and miR-92a[65]. The failure to recapitulate the original results in a different HEK293T cell line is therefore perhaps unsurprising, since different sources might in fact demonstrate profound differences: Bylund et al.[109] showed that batches of HEK293 cells from the European Collection of Cell Cultures had between 62 and 70 chromosomes. This is further different to other sources of the 'same' cell line[110,111], and different chromosomal aberrations were identified[112]. In light of this hypothesis, it may prove very informative to thoroughly assess by NanoString nCounter® Technology, for example, the miRNA species present, and their relative abundances, within the lines of HEK293T cells in which a *CEROX1* overexpression phenotype was observed, and in which it was not. However, this strategy relies upon the rediscovery of a clonal line in which the phenotype is robustly present: while new HEK293T cells were obtained during the troubleshooting phase of this chapter, the initial disappearance of phenotype occurred while using the same stocks of HEK293T cells as were used in the collection of the previous data. Phenotypic drift has been characterised within a number of cell culture models, but usually after extensive continuous culture, or in specific cell types[113-115]. Indeed, while the diversity of HEK293T cell lines is well established, studies have also identified relative stability during routine passaging and cryostorage[65,109], suggesting that the alteration or loss of the phenotype within the original batch of HEK293T cells is less likely due to a genomic alteration. However, the alteration of the mouse *Cerox1* overexpression phenotype in N2A cells also observed here, as well as sporadically in the Ponting laboratory in Edinburgh (Dr Tamara Sirey, personal communication) suggests a fundamentally non-cell type-specific issue. No phenotype was

observed in SH-SY5Y and SK-N-SH cells following *CEROX1* overexpression, but owing to the aforementioned proposed miRNA-dependence of the complex I phenotype, it may be the case that these cell lines do not express the requisite miRNA species: the slight differences in the original phenotypes observed between HEK293T and N2A cells upon human or mouse *CEROX1* overexpression is likely testament to their different, cell type-specific miRNA pools. Additionally, the transfection efficiency of these neuroblastoma cell lines is lower than that of HEK293T cells. Unfortunately, without knowing which miRNA(s) mediate the *CEROX1* phenotype, it is difficult to choose an alternative cell line in which a complex I phenotype would be expected to be apparent.

As well as differences in miRNA levels, it is possible that differences in the endogenous expression levels of complex I transcripts between different HEK293T sub-clonal lines results in differential stoichiometry of miRNA-mRNA binding and resulting ceRNA-mediated regulation. As with miRNA levels, it would be desirable to assess comparatively and concurrently complex I transcript expression between cells that host the original complex I phenotype and that do not.

c) Is there any possibility that the phenotype is dependent upon a parameter not identified, other than genomic diversity as discussed above? If so, it is hard to identify, since all efforts were made to replicate the original experiments to as great an extent as possible, including growth conditions. No differences in media or the foetal bovine serum (which is well known as a major source of cell culture variation[116]) routinely used as a growth additive were observed to correlate with the presence or absence of the *CEROX1* overexpression phenotype. A different commercial source of cytochrome *c* was once used in complex III and complex IV assays due to discontinuation of the original reagent by the supplier, and, interestingly, the complex IV activity assay behaved unusually. The reagent was replaced, to the returned success of the complex IV assay, but this demonstrates that some reagents may inexplicably alter the behaviour of phenotyping assays. There is, however, no specific evidence to suggest an issue with other reagents, and multiple different stocks have been used during the course of the experiments presented here. The temperature sensitive SV40 T-antigen present in HEK293T cells[106] is active at 33°C but not 40°C, and 37°C is considered semi-permissive[108]. It may be that temperature fluctuations during the growth of HEK293T cells influence gene expression networks via the interaction of differential T-antigen activity levels with p53[107,117,118] and retinoblastoma protein[119]. However, this hypothesis would depend upon a consistent difference in growth conditions between the two phases of experimental work presented here, and seems

highly unlikely. Wright Muelas et al.[120] have demonstrated the profound implications that a lack of optimal conditions in cell culture present for experimental reproducibility. While examining the effects of a glutaminase-1 inhibitor upon two cancer cell lines, it was shown that cell density and culture volume strongly influenced the observations, with fluctuation of growth rate, metabolite levels, media pH, and pathway activities convoluting the results[120]. While no overt changes were observed here in the growth of cells, and the same growth density and end confluency were obtained when assessed by eye, it may be that differences between the growth rate and metabolic behaviour of batches and sub-clones of cells, due to the genomic diversity explained above, contribute to an altered metabolic state that potentiates or de-potentiates a *CEROX1* modulation phenotype. Abrahamsen et al.[121] demonstrated that almost 60% of expressed genes were differentially expressed in SPAST mutant cells, despite normal growth.

Ultimately, the reason for the changed phenotype in the cell culture models, following *CEROX1*/*Cerox1* overexpression is not clear, and complicates the process of moving forward with the functional investigation of *CEROX1*. However, the positive trends garnered through preliminary Seahorse respiratory data suggest that an ETC phenotype may still be observed, and living growing cells are, for many obvious reasons, a better reflection of physiological conditions. Repeats of this experiment will be required to establish robustness of the preliminary data, but the data are consistent with previous results, including the magnitude of increase in activity (~30%). Given that mouse *Cerox1* has been shown to protect N2A cells against rotenone (Sirey et al, 2019[64]), the small increase in rotenone- and antimycin A-inhibited respiration, usually interpreted as "non-mitochondrial respiration" might instead represent a small independent effect upon the sensitivity of complexes I and III to their canonical inhibitors under *CEROX1* overexpression. The smaller scale of sample preparation for this respiratory assay (96-well plate-style) compared to the spectrophotometric assay (>25 cm^2 flasks) would also allow more economical use of LNA™ GapmeR inhibitors in order to investigate the respiratory phenotype of *CEROX1* knockdown (see Appendix A11). Based upon the characterisation undertaken in Chapter 3, it might be logical to transfer the investigation into a more physiologically relevant system, such the MO3.13 oligodendrocyte cell line, which may be differentiated *in vitro* as a model of oligodendrocyte development[122]. The use of a more physiologically relevant system was already planned for this investigation, and halted in light of the unpredictable behaviour of the previously positive control. Interestingly, one of the miRNAs for which an MRE is conserved in mouse and human *CEROX1*, miR-138, has been implicated in the differentiation of oligodendrocytes[123] and Schwann cells[124]. MiR-138 specifically

promotes, early stages of differentiation, while repressing maturation[123], and thus its levels are negatively correlated with those of *Cerox1* during late oligodendrocyte development. Other miRNAs whose levels are elevated throughout oligodendrocyte differentiation[125] include miR-338 and miR-181c the ETC-repressing activities of which were introduced above. MiR-338 and miR-219 both contribute to repression of precursory and immature gene expression programs in differentiating oligodendrocytes[126,127].

While it was out of the scope of this investigation due to cost and time constraints, in further disentangling the results presented here it would be prudent to follow the suggestion of Wright Muelas et al.[120] and assay growth parameters during the preparation of samples, including nutrient concentrations and cell density, in order to identify any blocks that would suggest a shift in metabolic control. Indeed, this may identify other metabolic effects of *CEROX1* perturbation. Additionally, challenging cells with respiratory toxins, such as sub-lethal doses of rotenone, or different media, including a low/no glucose medium or galactose-containing medium, which stimulates oxidative metabolism[128-130], might clarify or potentiate a *CEROX1* overexpression phenotype. With a *Cerox1* knock-out mouse in development (Dr Tamara Sirey, personal communication), a more physiologically relevant model will be available to examine the still-promising role of the lncRNA in regulating energy metabolism.

REFERENCES

1. Wallace, D.C. Mitochondrial DNA in aging and disease. *Sci Am* **277**, 40-7 (1997).
2. Vinothkumar, K.R., Zhu, J. & Hirst, J. Architecture of mammalian respiratory complex I. *Nature* **515**, 80 (2014).
3. Guo, R., Zong, S., Wu, M., Gu, J. & Yang, M. Architecture of Human Mitochondrial Respiratory Megacomplex I2III2IV2. *Cell* **170**, 1247-1257.e12 (2017).
4. Sazanov, L.A. A giant molecular proton pump: structure and mechanism of respiratory complex I. *Nature Reviews Molecular Cell Biology* **16**, 375 (2015).
5. Zhu, J., Vinothkumar, K.R. & Hirst, J. Structure of mammalian respiratory complex I. *Nature* **536**, 354-358 (2016).
6. Stroud, D.A., Surgenor, E.E., Formosa, L.E., Reljic, B. *et al.* Accessory subunits are integral for assembly and function of human mitochondrial complex I. *Nature* **538**, 123 (2016).
7. Guerrero-Castillo, S., Baertling, F., Kownatzki, D., Wessels, H.J. *et al.* The Assembly Pathway of Mitochondrial Respiratory Chain Complex I. *Cell Metab* **25**, 128-139 (2017).
8. Balsa, E., Marco, R., Perales-Clemente, E., Szklarczyk, R. *et al.* NDUFA4 Is a Subunit of Complex IV of the Mammalian Electron Transport Chain. *Cell Metabolism* **16**, 378-386 (2012).
9. DiMauro, S. & Emmanuele, V. The Clinical Spectrum of Nuclear DNA-Related Mitochondrial Disorders. in *Mitochondrial Disorders Caused by Nuclear Genes* (ed. Wong, L.-J.C.) 3-25 (Springer New York, 2013).
10. Distelmaier, F., Koopman, W.J.H., van den Heuvel, L.P., Rodenburg, R.J. *et al.* Mitochondrial complex I deficiency: from organelle dysfunction to clinical disease. *Brain* **132**, 833-842 (2009).
11. Keane, P.C., Kurzawa, M., Blain, P.G. & Morris, C.M. Mitochondrial Dysfunction in Parkinson's Disease. *Parkinsons Disease* (2011).
12. Ryan, B.J., Hoek, S., Fon, E.A. & Wade-Martins, R. Mitochondrial dysfunction and mitophagy in Parkinson's: from familial to sporadic disease. *Trends Biochem Sci* **40**, 200-10 (2015).
13. Zhou, L., Wang, W., Hoppel, C., Liu, J. & Zhu, X. Parkinson's disease-associated pathogenic VPS35 mutation causes complex I deficits. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1863**, 2791-2795 (2017).
14. Nunomura, A., Perry, G., Aliev, G., Hirai, K. *et al.* Oxidative damage is the earliest event in Alzheimer disease. *Journal of Neuropathology and Experimental Neurology* **60**, 759-767 (2001).
15. Moreira, P.I., Cardoso, S.M., Santos, M.S. & Oliveira, C.R. The key role of mitochondria in Alzheimer's disease. *J Alzheimers Dis* **9**, 101-10 (2006).
16. Moreira, P.I., Santos, M.S. & Oliveira, C.R. Alzheimer's disease: a lesson from mitochondrial dysfunction. *Antioxid Redox Signal* **9**, 1621-30 (2007).
17. Moreira, P.I., Duarte, A.I., Santos, M.S., Rego, A.C. & Oliveira, C.R. An integrative view of the role of oxidative stress, mitochondria and insulin in Alzheimer's disease. *J Alzheimers Dis* **16**, 741-61 (2009).
18. Moreira, P.I., Carvalho, C., Zhu, X., Smith, M.A. & Perry, G. Mitochondrial dysfunction is a trigger of Alzheimer's disease pathophysiology. *Biochim Biophys Acta* **1802**, 2-10 (2010).
19. Zeitlow, K., Charlambous, L., Ng, I., Gagrani, S. *et al.* The biological foundation of the genetic association of TOMM40 with late-onset Alzheimer's disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1863**, 2973-2986 (2017).
20. Benchoua, A., Trioulier, Y., Zala, D., Gaillard, M.-C. *et al.* Involvement of Mitochondrial Complex II Defects in Neuronal Death Produced by N-Terminus Fragment of Mutated Huntingtin. *Molecular Biology of the Cell* **17**, 1652-1663 (2006).
21. Gu, M., Gash, M.T., Mann, V.M., Javoy-Agid, F. *et al.* Mitochondrial defect in Huntington's disease caudate nucleus. *Annals of Neurology* **39**, 385-389 (1996).
22. Mochel, F. & Haller, R.G. Energy deficit in Huntington disease: why it matters. *The Journal of Clinical Investigation* **121**, 493-499 (2011).
23. Stahl, W.L. & Swanson, P.D. Biochemical abnormalities in Huntington's chorea brains. *Neurology* **24**, 813-9 (1974).
24. Leigh, D. Subacute necrotizing encephalomyelopathy in an infant. *J Neurol Neurosurg Psychiatry* **14**, 216-21 (1951).
25. Morris, S.A. & Harbord, M.G. Infant onset subacute necrotizing encephalomyelopathy (Leigh's disease). *J Paediatr Child Health* **29**, 363-7 (1993).

26. Loeffen, J., Smeitink, J., Triepels, R., Smeets, R. *et al.* The first nuclear-encoded complex I mutation in a patient with Leigh syndrome. *Am J Hum Genet* **63**, 1598-608 (1998).
27. Lake, N.J., Compton, A.G., Rahman, S. & Thorburn, D.R. Leigh Syndrome: One disorder, more than 75 monogenic causes. *Annals of Neurology* **79**, 190-203 (2015).
28. Schagger, H. & Pfeiffer, K. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *EMBO J* **19**, 1777-83 (2000).
29. Acin-Perez, R., Fernandez-Silva, P., Peleato, M.L., Perez-Martos, A. & Enriquez, J.A. Respiratory active mitochondrial supercomplexes. *Mol Cell* **32**, 529-39 (2008).
30. Milenkovic, D., Blaza, J.N., Larsson, N.-G. & Hirst, J. The Enigma of the Respiratory Chain Supercomplex. *Cell Metabolism* **25**, 765-776 (2017).
31. Buck, K.J., Walter, N.A.R. & Denmark, D.L. Genetic variability of respiratory complex abundance, organization and activity in mouse brain. *Genes Brain and Behavior* **13**, 135-143 (2014).
32. Davies, K.M., Blum, T.B. & Kühlbrandt, W. Conserved in situ arrangement of complex I and III2 in mitochondrial respiratory chain supercomplexes of mammals, yeast, and plants. *Proceedings of the National Academy of Sciences* **115**, 3024-3029 (2018).
33. Fedor, J.G. & Hirst, J. Mitochondrial Supercomplexes Do Not Enhance Catalysis by Quinone Channeling. *Cell Metabolism* **28**, 525-531.e4 (2018).
34. Matsumoto, S., Uchiumi, T., Tanamachi, H., Saito, T. *et al.* Ribonucleoprotein Y-box-binding protein-1 regulates mitochondrial oxidative phosphorylation (OXPHOS) protein expression after serum stimulation through binding to OXPHOS mRNA. *Biochem J* **443**, 573-584 (2012).
35. Evdokimova, V.M., Kovrigina, E.A., Nashchekin, D.V., Davydova, E.K. *et al.* The Major Core Protein of Messenger Ribonucleoprotein Particles (p50) Promotes Initiation of Protein Biosynthesis in Vitro. *J Biol Chem* **273**, 3574-3581 (1998).
36. Alvarez-Fischer, D., Fuchs, J., Castagner, F., Stettler, O. *et al.* Engrailed protects mouse midbrain dopaminergic neurons against mitochondrial complex I insults. *Nature Neuroscience* **14**, 1260 (2011).
37. Sanders, L.H. & Greenamyre, J.T. Regulation of complex I by Engrailed is complex too. *Nature Neuroscience* **14**, 1221 (2011).
38. Kulshreshtha, R., Ferracin, M., Wojcik, S.E., Garzon, R. *et al.* A MicroRNA Signature of Hypoxia. *Mol Cell Biol* **27**, 1859-1867 (2007).
39. Puisségur, M.P., Mazure, N.M., Bertero, T., Pradelli, L. *et al.* miR-210 is overexpressed in late stages of lung cancer and mediates mitochondrial alterations associated with modulation of HIF-1 activity. *Cell Death and Differentiation* **18**, 465-478 (2011).
40. Grosso, S., Doyen, J., Parks, S.K., Bertero, T. *et al.* MiR-210 promotes a hypoxic phenotype and increases radioresistance in human lung cancer cell lines. *Cell Death & Disease* **4**, e544 (2013).
41. Chen, Z., Li, Y., Zhang, H., Huang, P. & Luthra, R. Hypoxia-regulated microRNA-210 modulates mitochondrial function and decreases ISCU and COX10 expression. *Oncogene* **29**, 4362 (2010).
42. Chan, S.Y., Zhang, Y.-Y., Hemann, C., Mahoney, C.E. *et al.* MicroRNA-210 Controls Mitochondrial Metabolism during Hypoxia by Repressing the Iron-Sulfur Cluster Assembly Proteins ISCU1/2. *Cell Metabolism* **10**, 273-284 (2009).
43. Sirey, T.M. & Ponting, C.P. Insights into the post-transcriptional regulation of the mitochondrial electron transport chain. *Biochemical Society Transactions* **44**, 1491-1498 (2016).
44. Muralimanoharan, S., Maloyan, A., Mele, J., Guo, C. *et al.* miR-210 Modulates Mitochondrial Respiration in Placenta with Preeclampsia. *Placenta* **33**, 816-823 (2012).
45. Lee, D.-C., Romero, R., Kim, J.-S., Tarca, A.L. *et al.* miR-210 Targets Iron-Sulfur Cluster Scaffold Homologue in Human Trophoblast Cell Lines: Siderosis of Interstitial Trophoblasts as a Novel Pathology of Preterm Preeclampsia and Small-for-Gestational-Age Pregnancies. *The American Journal of Pathology* **179**, 590-602 (2011).
46. Chan, S.Y. & Loscalzo, J. MicroRNA-210: A unique and pleiotropic hypoxamir. *Cell Cycle* **9**, 1072-1083 (2010).
47. Bertero, T., Grosso, S., Robbe-Sermesant, K., Lebrigand, K. *et al.* "Seed-Milarity" Confers to hsa-miR-210 and hsa-miR-147b Similar Functional Activity. *PLOS ONE* **7**, e44919 (2012).
48. Aschrafi, A., Schwechter, A.D., Mameza, M.G., Natera-Naranjo, O. *et al.* MicroRNA-338 Regulates Local Cytochrome c Oxidase IV mRNA Levels and Oxidative Phosphorylation in the Axons of Sympathetic Neurons. *J Neurosci* **28**, 12581-12590 (2008).

49. Aschrafi, A., Kar, A.N., Natera-Naranjo, O., MacGibeny, M.A. *et al.* MicroRNA-338 regulates the axonal expression of multiple nuclear-encoded mitochondrial mRNAs encoding subunits of the oxidative phosphorylation machinery. *Cellular and Molecular Life Sciences* **69**, 4017-4027 (2012).
50. Nie, H., Yu, X., He, H., Zhou, L. *et al.* Hepatocyte miR-33a mediates mitochondrial dysfunction and hepatosteatosis by suppressing NDUFA5. *Journal of Cellular and Molecular Medicine* **0**(2018).
51. Aoi, W., Naito, Y., Mizushima, K., Takanami, Y. *et al.* The microRNA miR-696 regulates PGC-1α in mouse skeletal muscle in response to physical activity. *Am J Physiol Endocrinol Metab* **298**, E799-E806 (2010).
52. Kren, B.T., Wong, P.Y.P., Sarver, A., Zhang, X. *et al.* MicroRNAs identified in highly purified liver-derived mitochondria may play a role in apoptosis. *RNA Biology* **6**, 65-72 (2009).
53. Bian, Z., Li, L.-M., Tang, R., Hou, D.-X. *et al.* Identification of mouse liver mitochondria-associated miRNAs and their potential biological functions. *Cell Research* **20**, 1076 (2010).
54. Mercer, Tim R., Neph, S., Dinger, Marcel E., Crawford, J. *et al.* The Human Mitochondrial Transcriptome. *Cell* **146**, 645-658 (2011).
55. Barrey, E., Saint-Auret, G., Bonnamy, B., Damas, D. *et al.* Pre-microRNA and Mature microRNA in Human Mitochondria. *PLOS ONE* **6**, e20220 (2011).
56. Sripada, L., Tomar, D., Prajapati, P., Singh, R. *et al.* Systematic Analysis of Small RNAs Associated with Human Mitochondria by Deep Sequencing: Detailed Analysis of Mitochondrial Associated miRNA. *PLOS ONE* **7**, e44873 (2012).
57. Borralho, P.M., Rodrigues, C.M.P. & Steer, C.J. Mitochondrial MicroRNAs and Their Potential Role in Cell Function. *Current Pathobiology Reports* **2**, 123-132 (2014).
58. Das, S., Ferlito, M., Kent, O.A., Fox-Talbot, K. *et al.* Nuclear miRNA Regulates the Mitochondrial Genome in the Heart. *Circulation Research* **110**, 1596-1603 (2012).
59. Das, S., Bedja, D., Campbell, N., Dunkerly, B. *et al.* miR-181c Regulates the Mitochondrial Genome, Bioenergetics, and Propensity for Heart Failure In Vivo. *PLOS ONE* **9**, e96820 (2014).
60. Zhou, R., Wang, R., Qin, Y., Ji, J. *et al.* Mitochondria-related miR-151a-5p reduces cellular ATP production by targeting CYTB in asthenozoospermia. *Scientific Reports* **5**, 17743 (2015).
61. Zhang, X., Zuo, X., Yang, B., Li, Z. *et al.* MicroRNA Directly Enhances Mitochondrial Translation during Muscle Differentiation. *Cell* **158**, 607-619 (2014).
62. Reeves, M.B., Davies, A.A., McSharry, B.P., Wilkinson, G.W. & Sinclair, J.H. Complex I Binding by a Virally Encoded RNA Regulates Mitochondria-Induced Cell Death. *Science* **316**, 1345-1348 (2007).
63. Huang, G., Lu, H., Hao, A., Ng, D.C.H. *et al.* GRIM-19, a Cell Death Regulatory Protein, Is Essential for Assembly and Function of Mitochondrial Complex I. *Molecular and Cellular Biology* **24**, 8447-8456 (2004).
64. Sirey, T.M., Roberts, K., Haerty, W., Bedoya-Reina, O. *et al.* The long non-coding RNA Cerox1 is a post transcriptional regulator of mitochondrial complex I catalytic activity. *eLife* **8**, e45051 (2019).
65. Lin, Y.-C., Boone, M., Meuris, L., Lemmens, I. *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications* **5**, 4767 (2014).
66. Shaw, G., Morse, S., Ararat, M. & Graham, F.L. Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. **16**, 869-871 (2002).
67. Falkenburger, B.H., Saridaki, T. & Dinter, E. Cellular models for Parkinson's disease. **139**, 121-130 (2016).
68. Schlachetzki, J.C.M., Saliba, S.W. & Oliveira, A.C.P.d. Studying neurodegenerative diseases in culture models. *Brazilian Journal of Psychiatry* **35**, S92-S100 (2013).
69. Sinha, D., Srivastava, S., Krishna, L. & D'Silva, P. Unraveling the Intricate Organization of Mammalian Mitochondrial Presequence Translocases: Existence of Multiple Translocases for Maintenance of Mitochondrial Function. **34**, 1757-1775 (2014).
70. de Moura, M.B., Uppala, R., Zhang, Y., Van Houten, B. & Goetzman, E.S. Overexpression of mitochondrial sirtuins alters glycolysis and mitochondrial function in HEK293 cells. *PloS one* **9**, e106028-e106028 (2014).
71. Kim, C., Patel, P., Gouvin, L.M., Brown, M.L. *et al.* Comparative Analysis of the Mitochondrial Physiology of Pancreatic β Cells. *Bioenergetics : open access* **3**, 110 (2014).
72. Larsen, S., Nielsen, J., Hansen, C.N., Nielsen, L.B. *et al.* Biomarkers of mitochondrial content in skeletal muscle of healthy young human subjects. *J Physiol* **590**, 3349-60 (2012).
73. Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A. *et al.* Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science* **334**, 521 (2011).

74. dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J. *et al.* Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences* **279**, 3491 (2012).
75. dos Reis, M., Donoghue, P.C.J. & Yang, Z. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. *Biology Letters* **10**(2014).
76. Leung, A.K.L. The Whereabouts of microRNA Actions: Cytoplasm and Beyond. *Trends in Cell Biology* **25**, 601-610 (2015).
77. Smillie, C.L., Sirey, T. & Ponting, C.P. Complexities of post-transcriptional regulation and the modeling of ceRNA crosstalk. *Critical Reviews in Biochemistry and Molecular Biology*, 1-15 (2018).
78. Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**, 925-38 (2010).
79. Tan, J.Y., Vance, K.W., Varela, M.A., Sirey, T. *et al.* Cross-talking noncoding RNAs contribute to cell-specific neurodegeneration in SCA7. *Nat Struct Mol Biol* **21**, 955-61 (2014).
80. Karreth, F.A., Tay, Y., Perna, D., Ala, U. *et al.* In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147**, 382-95 (2011).
81. Bogerd, H.P., Whisnant, A.W., Kennedy, E.M., Flores, O. & Cullen, B.R. Derivation and characterization of Dicer- and microRNA-deficient human cells. *RNA* **20**, 923-37 (2014).
82. Murata, A., Fukuzumi, T., Umemoto, S. & Nakatani, K. Xanthone derivatives as potential inhibitors of miRNA processing by human Dicer: targeting secondary structures of pre-miRNA by small molecules. *Bioorg Med Chem Lett* **23**, 252-5 (2013).
83. Tan, G.S., Chiu, C.H., Garchow, B.G., Metzler, D. *et al.* Small molecule inhibition of RISC loading. *ACS Chem Biol* **7**, 403-10 (2012).
84. Yue, D., Liu, H. & Huang, Y. Survey of Computational Algorithms for MicroRNA Target Prediction. *Curr Genomics* **10**, 478-92 (2009).
85. Witkos, T.M., Koscianska, E. & Krzyzosiak, W.J. Practical Aspects of microRNA Target Prediction. *Curr Mol Med* **11**, 93-109 (2011).
86. Tabas-Madrid, D., Muniategui, A., Sanchez-Caballero, I., Martinez-Herrera, D.J. *et al.* Improving miRNA-mRNA interaction predictions. *BMC Genomics* **15 Suppl 10**, S2 (2014).
87. Helwak, A. & Tollervey, D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat Protoc* **9**, 711-28 (2014).
88. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654-65 (2013).
89. Cloonan, N. Re-thinking miRNA-mRNA interactions: Intertwining issues confound target discovery. *Bioessays* **37**, 379-388 (2015).
90. Tian, W., Dong, X., Liu, X., Wang, G. *et al.* High-throughput functional microRNAs profiling by recombinant AAV-based microRNA sensor arrays. *PLoS One* **7**, e29551 (2012).
91. Mercer, T.R., Wilhelm, D., Dinger, M.E., Solda, G. *et al.* Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res* **39**, 2393-403 (2011).
92. Berg JM, T.J., Stryer L. 17.2 Entry to the Citric Acid Cycle and Metabolism Through It Are Controlled. in *Biochemistry* (ed. Freeman, W.) (New York, 2002).
93. Berg JM, T.J., Stryer L. 16.2 The Glycolytic Pathway Is Tightly Controlled. in *Biochemistry* (ed. Freeman, W.) (New York, 2002).
94. Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**(2015).
95. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8**, 93-103 (2007).
96. Lin, S., Cheung, W.K., Chen, S., Lu, G. *et al.* Computational identification and characterization of primate-specific microRNAs in human genome. *Comput Biol Chem* **34**, 232-41 (2010).
97. Pinzón, N., Li, B., Martinez, L., Sergeeva, A. *et al.* microRNA target prediction programs predict many false positives. *Genome Res* **27**, 234-245 (2017).
98. Krzywinski, M., Schein, J., Birol, I., Connors, J. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).
99. Yan, L.-J. & Forster, M.J. Resolving mitochondrial protein complexes using non-gradient blue native polyacrylamide gel electrophoresis. *Analytical biochemistry* **389**, 143-149 (2009).
100. Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P. & Stoffel, M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell* **54**, 766-76 (2014).

101. Schapira, A.H., Cooper, J.M., Dexter, D., Clark, J.B. *et al.* Mitochondrial complex I deficiency in Parkinson's disease. *J Neurochem* **54**, 823-7 (1990).
102. Keeney, P.M., Xie, J., Capaldi, R.A. & Bennett, J.P. Parkinson's Disease Brain Mitochondrial Complex I Has Oxidatively Damaged Subunits and Is Functionally Impaired and Misassembled. *J Neurosci* **26**, 5256-5264 (2006).
103. Canevari, L., Clark, J.B. & Bates, T.E. beta-Amyloid fragment 25-35 selectively decreases complex IV activity in isolated mitochondria. *FEBS Lett* **457**, 131-4 (1999).
104. Sadeghian, M., Mastrolia, V., Rezaei Haddad, A., Mosley, A. *et al.* Mitochondrial dysfunction is an important cause of neurological deficits in an inflammatory model of multiple sclerosis. *Scientific Reports* **6**, 33249 (2016).
105. Browne, S.E., Bowling, A.C., Macgarvey, U., Baik, M.J. *et al.* Oxidative damage and metabolic dysfunction in Huntington's disease: Selective vulnerability of the basal ganglia. *Annals of Neurology* **41**, 646-653 (2004).
106. DuBridge, R.B., Tang, P., Hsia, H.C., Leong, P.M. *et al.* Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol Cell Biol* **7**, 379-387 (1987).
107. Lilyestrom, W., Klein, M.G., Zhang, R., Joachimiak, A. & Chen, X.S. Crystal structure of SV40 large T-antigen bound to p53: interplay between a viral oncoprotein and a cellular tumor suppressor. *Genes Dev* **20**, 2373-2382 (2006).
108. Rio, D., Clark, S. & Tjian, R. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science* **227**, 23-28 (1985).
109. Bylund, L., Kytölä, S., Lui, W.O., Larsson, C. & Weber, G. Analysis of the cytogenetic stability of the human embryonal kidney cell line 293 by cytogenetic and STR profiling approaches. *Cytogenetic and Genome Research* **106**, 28-32 (2004).
110. Stavropoulou, V., Xie, J., Henriksson, M., Tomkinson, B. *et al.* Mitotic Infidelity and Centrosome Duplication Errors in Cells Overexpressing Tripeptidyl-Peptidase II. *Cancer Res* **65**, 1361-1368 (2005).
111. Stepanenko, A., Andreieva, S., Korets, K., Mykytenko, D. *et al.* Step-wise and punctuated genome evolution drive phenotype changes of tumor cells. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **771**, 56-69 (2015).
112. Stepanenko, A.A. & Dmitrenko, V.V. HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene* **569**, 182-190 (2015).
113. Babaev, V.R., Antonov, A.S., Domogatsky, S.P. & Kazantseva, I.A. Phenotype related changes of intimal smooth muscle cells from human aorta in primary culture. *Atherosclerosis* **96**, 189-202 (1992).
114. Yao, L., Bestwick, C.S., Bestwick, L.A., Maffulli, N. & Aspden, R.M. Phenotypic drift in human tenocyte culture. *Tissue Eng* **12**, 1843-9 (2006).
115. Zavyalova, M.V., Denisov, E.V., Tashireva, L.A., Gerashchenko, T.S. *et al.* Phenotypic Drift as a Cause for Intratumoral Morphological Heterogeneity of Invasive Ductal Breast Carcinoma Not Otherwise Specified. *BioResearch Open Access* **2**, 148-154 (2013).
116. Baker, M. Reproducibility: Respect your cells! *Nature* **537**, 433 (2016).
117. Linzer, D.I.H. & Levine, A.J. Characterization of a 54K Dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. *Cell* **17**, 43-52 (1979).
118. Lane, D.P. & Crawford, L.V. T antigen is bound to a host protein in SY40-transformed cells. *Nature* **278**, 261 (1979).
119. DeCaprio, J.A., Ludlow, J.W., Figge, J., Shew, J.-Y. *et al.* SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. *Cell* **54**, 275-283 (1988).
120. Wright Muelas, M., Ortega, F., Breitling, R., Bendtsen, C. & Westerhoff, H.V. Rational cell culture optimization enhances experimental reproducibility in cancer cells. *Scientific Reports* **8**, 3029 (2018).
121. Abrahamsen, G., Fan, Y., Matigian, N., Wali, G. *et al.* A patient-derived stem cell model of hereditary spastic paraplegia with SPAST mutations. *Dis Model Mech* **6**, 489-502 (2013).
122. Buntinx, M., Vanderlocht, J., Hellings, N., Vandenabeele, F. *et al.* Characterization of three human oligodendroglial cell lines as a model to study oligodendrocyte injury: Morphology and oligodendrocyte-specific gene expression. *Journal of Neurocytology* **32**, 25-38 (2003).
123. Dugas, J.C., Cuellar, T.L., Scholze, A., Ason, B. *et al.* Dicer1 and miR-219 Are Required for Normal Oligodendrocyte Differentiation and Myelination. *Neuron* **65**, 597-611 (2010).

124. Yun, B., Anderegg, A., Menichella, D., Wrabetz, L. *et al.* MicroRNA-Deficient Schwann Cells Display Congenital Hypomyelination. *J Neurosci* **30**, 7722-7728 (2010).
125. Lau, P., Verrier, J.D., Nielsen, J.A., Johnson, K.R. *et al.* Identification of Dynamically Regulated MicroRNA and mRNA Networks in Developing Oligodendrocytes. *J Neurosci* **28**, 11720-11730 (2008).
126. Shin, D., Shin, J.-Y., McManus, M.T., Ptáček, L.J. & Fu, Y.-H. Dicer ablation in oligodendrocytes provokes neuronal impairment in mice. *Ann Neurol* **66**, 843-857 (2009).
127. Zhao, X., He, X., Han, X., Yu, Y. *et al.* MicroRNA-Mediated Control of Oligodendrocyte Differentiation. *Neuron* **65**, 612-626 (2010).
128. Rossignol, R., Gilkerson, R., Aggeler, R., Yamagata, K. *et al.* Energy Substrate Modulates Mitochondrial Structure and Oxidative Capacity in Cancer Cells. *Cancer Research* **64**, 985-993 (2004).
129. Marroquin, L.D., Hynes, J., Dykens, J.A., Jamieson, J.D. & Will, Y. Circumventing the Crabtree Effect: Replacing Media Glucose with Galactose Increases Susceptibility of HepG2 Cells to Mitochondrial Toxicants. *Toxicological Sciences* **97**, 539-547 (2007).
130. Aguer, C., Gambarotta, D., Mailloux, R.J., Moffat, C. *et al.* Galactose Enhances Oxidative Metabolism and Reveals Mitochondrial Dysfunction in Human Primary Muscle Cells. *PLoS ONE* **6**, e28536 (2011).

CHAPTER 6: THE *CEROX1-SOX8* PROMOTER

INTRODUCTION

6.1 *CEROX1* is surely regulated, as well as being a putative regulator

The body of this thesis thus far has focused on the potential regulatory function of *CEROX1*, and contextualising its expression and conservation in light of this. However, it is a gene like any other, and is presumably subject to regulatory mechanisms in its own right, thereby integrating its function(s) into the complex demands of the mammalian cell. Regulation begins at the point of creation, in the form of transcriptional modulation at the *CEROX1* promoter, which is shared with the divergent gene *SOX8*.

6.1.1 Regulation would likely reflect *CEROX1*'s expression pattern and putative function

In keeping with the proposed role of *CEROX1* in regulating mitochondrial energy metabolism, it seems likely that levels of *CEROX1* might be modulated in response to stimuli whose activation or withdrawal indicate changes in energy level or metabolic activity on either a cellular, tissue, or organismal level. Examples of these include the level of metabolites such as glucose, presence of key metabolic hormones such as insulin, and activity of enzymes integral to energy metabolism such as complex I, perhaps via its influence on the redox state of the cell[1]. Due to the contribution of mitochondria in producing reactive oxygen species that lead to oxidative stress, it is possible that *CEROX1* may also be found to respond to general stress stimuli, in doing so relieving the burden of combinatorial cell stress.

Trans-acting regulators such as transcription factors are important in mediating cell type- and tissue-specific expression[2], and so the identification of binding sites for known central nervous system (CNS)-enriched regulators, such as the essential brain development transcription factor Nfix[3], would be consistent with the enrichment of *CEROX1* in the CNS. ChIP peaks representing binding of the oligodendrocyte differentiation regulator Olig2[4] to the *Cerox1* locus in mouse and rat have already been illustrated in Chapter 3 (Figure 3.12).

6.1.2 Combined approaches facilitate the identification of functionally relevant regulatory transcriptional mechanisms

Synergistic approaches have been most fruitful in almost all avenues of biomedical research, and the investigation of transcriptional regulation is no exception. In line with other chapters of this thesis in which multidisciplinary studies have been applied to aspects of *CEROX1* biology, this chapter will incorporate both bioinformatic and experimental methodologies. Care must be taken in interpreting both bioinformatic data and cellular models, which both incur limitations. A review of approaches, as well as their foundations and their constraints, will be discussed below.

6.1.2.1 Finding starting points: from transcription factors upwards…

In the investigation of gene control, as with many genetic studies, it is possible to take a 'bottom-up' or a 'top-down' approach[5]. In the former, also known as reverse chemical genetics, definition of putative transcription factor binding sites, based upon consensus binding sequences or motifs, is used to identify transcriptional regulator candidates. The binding and subsequent effects of which may then be assessed by techniques such as the electrophoretic mobility shift assay (EMSA), combined with binding site mutagenesis and transcription factor inhibitors. The effect of experimentally-confirmed binding factors may then be established using reporter systems and secondary assays. This approach has the advantage of establishing biochemical knowledge of the promoter elements and their interaction partners. However, it works under the treacherous assumption that the selected target protein and its pathway have significant control over the process or gene of interest, and that modulation will therefore contribute to remediation of a disease state[6]. It therefore might not quickly establish gross modulators of gene expression, such as drug candidates, though the effects and possible cross-interactions of any eventual modulators are theoretically well understood. For example, via interaction with Brca1 *lncRNA-JADE* was found to regulate transcription of the neighbouring protein-coding gene JADE1, which is a component of the histone acetylase complex and contributes to DNA damage response-promoting H4 acetylation[7]. *LncRNA-JADE* was found to be transcriptionally upregulated by NF-κB, downstream of the DNA damage protein ATM (ataxia-telangiectasia mutated)[7]. *LncRNA-JADE* was upregulated in a majority of human breast cancer tumours examined, and it was found that knockdown reduced growth of breast cancer cell lines *in vitro* and the rate of xenograft tumourigenesis in injected mice[7]. From this it might be suggested that therapies targeting components of

the NF-κB pathway may prove beneficial in breast cancer. However, this notion is far from novel[8,9] and despite over 30 years of research into their clinical application NF-κB inhibitors have shown "little efficacy, except for certain types of lymphoma and leukemia"[10]. The biology of a pleiotropic factor such as NF-κB is highly complex and not simply targeted.

6.1.2.2 … or from pathways downwards?

Alternatively, in the top-down or forward chemical genetics approach, a stimulus, drug, or phenotype associated with altered gene expression or transcriptional activity is sought, again using a reporter system. This approach initially carries the advantage of swift and efficient high-throughput screening, and may relatively quickly identify an exogenous or endogenous modulator, since the multitude of potential targets are situated within a more physiologically relevant environment, unlike the purified proteins typically used in reverse chemical genetics. Indeed, it is also possible to discover new targets that would not have been addressed otherwise. Subsequently, the mechanism via which the stimulus elicits its effect may be traced through a known pathway(s), using combinatorial inhibitor studies and next-generation 'omics' experiments including RNA-seq, quantitative (phospho-)proteomics, chromatin immunoprecipitation (ChIP-seq)[11], chromosome conformation capture[12] (Hi-C[13], Capture-C[14,15], ChIA-PET[16]), and the assay for transposase-accessible chromatin (ATAC-seq)[17]. Therefore, while the effect of drug candidates upon overall gene expression may be quickly established, the biological promiscuity of many drugs and the substantial cross-talk of pathways may be reflected in the relative difficulty of identification of their mechanism of action upon a given promoter[18].

For example, a GFP-reporter containing the promoter of foetal γ-globin alongside a known enhancer from the globin locus control region was utilised in a high-throughput screen 10,000 compounds in search of γ-globin modulators to treat β-haemoglobinopathies such as β-thalassemia by a replacement-upregulation therapy[19]. The existing FDA-approved drug benserazide, an aromatic L-amino acid decarboxylase inhibitor used in combination with L-dopa for the treatment of Parkinson's disease, was found to upregulate transcriptional activity of the reporter, and its efficacy in upregulating γ-globin mRNA and protein, as well as foetal haemoglobin and foetal haemoglobin-containing reticulocytes, was demonstrated in human erythroid cell cultures, a non-human primate model, and a transgenic mouse model containing the human globin locus[19]. Subsequently, the mechanism of action was sought, making use of existing knowledge and

hypotheses about regulation of the γ-globin gene during foetal to adult stage switching. It was found that benserazide treatment decreases expression of a known γ-globin silencer, the lysine-specific histone demethylase LSD1, as well as its co-repressor transcription factor BCL11A. Reduced occupancy at the γ-globin locus by LSD1 and the histone deacetylase HDAC3 were demonstrated by ChIP, and this was corroborated by increased density of H3K4 dimethylation and H3K9 acetylation[20]. While the exact molecular target of this function of benserazide remains unidentified, this forward genetics approach has identified a promising treatment for β-haemoglobinopathies, and benserazide has been designated orphan drug status for the treatment of β-thalassemia (EMA/COMP/736552/2014).

6.1.3 Previous investigation of the regulation of *Sox8* produced an incomplete model

Clearly, the distal *cis*- or *trans*-regulatory elements of a gene may not be recapitulated using a simple proximal promoter reporter construct, especially since regulatory sites may be situated tens or hundreds of kb from the TSS[21-24]. A previous study by Guth et al.[25] attempted to identify the importance of several putative distal enhancer elements in the regulation of *Sox8*, with which *Cerox1* shares a bidirectional promoter. Seven putative enhancers were identified upon the basis of evolutionary conservation within a 220 kb region surrounding the *Sox8* TSS, and the *Cerox1* locus was included within the annotation of the 'minimal promoter', since it was not yet annotated. This also therefore included the putative enhancer element identified using ChIP-seq data in Chapter 3. Of the seven putative sequences, three were found to induce reproducible expression of an *Hsp68* minimal promoter-conjugated *lacZ* gene in mouse embryos; the other four showed no enhancing activity in that assay, though might be associated with regulation at a different developmental stage[25]. Importantly, the minimal *Sox8* promoter alone stimulated little to no expression during the developmental period (E10.5-12.5) examined *in vivo* without its enhancers. When compared to *lacZ* expression driven by the endogenous *Sox8* locus, therefore including its full complement of identified and not-yet identified regulatory elements, the three enhancer elements were each associated with specific tissue expression patterns, recapitulating parts of endogenous *Sox8* expression. However, their combined activities still do not recapitulate the full expression pattern of the gene during development, from which it must be inferred that there are additional regulatory elements more distal to *Sox8*[25] that are important at this stage.

Following the identification of its promoter as a site of binding, *Sox8* was also the subject of a study[26] investigating the mouse lncRNA *mrhl*[27,28]. It was shown *in vitro* and *in vivo* that Wnt activation of Sertoli cells and spermatogonial cells in the testes leads to upregulation of *Sox8* and that this is mediated by *mrhl*. Binding of *mrhl* to a site in the *Sox8* promoter is necessary for the recruitment of a repressive complex consisting of Max-Mad and the co-repressors Sin3A and HDAC1 (histone deacetylase 1). Under stimulation of Wnt signalling, the removal of *mrhl* is followed by recruitment of the transcriptionally activating Myc-Max complex and its histone acetyltransferase co-activator Pcaf (p300/CBP-associated factor). This leads to H3K9 acetylation and subsequent activation of *Sox8* expression[26]. The putative regulation of Sox8 was initially observed based upon microarray data that showed an increase in Sox8 expression following siRNA knockdown of *mrhl* in the cultured spermatogonial cells line Gc1-Spg[27]. However, the same experiment did not show an effect upon *Cerox1* expression, despite the fact that the *mrhl* binding region, according to chromatin oligo-affinity precipitation (ChOP) sits in the middle of their putative shared promoter region[28]. This is suggestive that regulation of *SOX8* and *CEROX1* is not completely correlated.

6.1.4 Addressing the *CEROX1* promoter

In this chapter, several approaches will be taken to begin the analysis of *CEROX1* regulation. This will begin with the establishment of a core promoter, confirming the assumption that transcription of the *CEROX1* and *SOX8* genes is initiated from the region between them. Following the above appraisal of the potential methodologies and their caveats, a forward chemical genetics approach using a luciferase reporter screen will then be applied in order to develop a tool assay and potential tool compounds for use in the further investigation of the role of *CEROX1* as well as its regulation. Together with bioinformatic analyses of putative transcriptional regulators, hypotheses of several potential *CEROX1*-modulating pathways will be formulated and discussed in light of the endogenous expression of the lncRNA.

RESULTS

6.2 *CEROX1* and *SOX8* share a bidirectional promoter

6.2.1 Characteristics of the *CEROX1-SOX8* promoter are consistent with bidirectionality

Within the human genome, divergent transcription is a common form of promoter architecture that accounts for over 11% of gene expression, with ~1350 pairs of genes being expressed from bidirectional promoters[29]. This group of bidirectional promoters exhibit several characteristic trends to which the *CEROX1-SOX8* promoter adhere: they have higher GC content (average 66%), are more likely to be situated within a CpG island, and less likely to host a consensus TATA box element[29]. The *CEROX1-SOX8* promoter, like *CEROX1* itself, is very GC rich (~75%). It is situated within one of the largest CpG islands in the genome (Chapter 3), and no strict TATA box is present. Bidirectional promoters also contain disproportionate ratios of transcription factor binding sites relative to the whole genome[30].

6.2.2 Expression of *CEROX1* and *SOX8* is highly correlated across human tissues

The similarity in gene expression patterns of *CEROX1* and *SOX8* within the CNS was already noted in Chapter 3. When examining GTEx[31] expression data for *CEROX1*, alongside that of its genomic neighbour *SOX8*, it is notable that the two share a striking similarity in their expression patterns across all tissues (Figure 6.1A). While expression of *CEROX1* is higher than that of *SOX8* in any given tissue, both genes are most highly expressed in the central nervous system. Given that the expression of bidirectionally-paired genes is typically more highly correlated than otherwise neighbouring genes[29,30], the correlation between *CEROX1* and *SOX8* was calculated. There is a highly significant level of strong positive correlation between the expression values of *CEROX1* and *SOX8* across the 53 tissues represented (Spearman's correlation coefficient $\rho = 0.943$; $p = 1.62 \times 10^{-26}$) (Figure 6.1B). There is no significant comparable relationship between expression of *CEROX1* and the neighbouring gene *LMF1* ($\rho = 0.156$; $p = 0.790$) or between expression of *SOX8* and *LMF1* ($\rho = 0.142$; $p = 0.927$) (Figure 6.1C-D). It is difficult to disentangle the extent to which the correlation in their expression levels is indicative that *CEROX1* and *SOX8* are purposefully regulated in concert through their common promoter region, or merely due to their incidental reliance upon permissivity of the same region of chromatin.

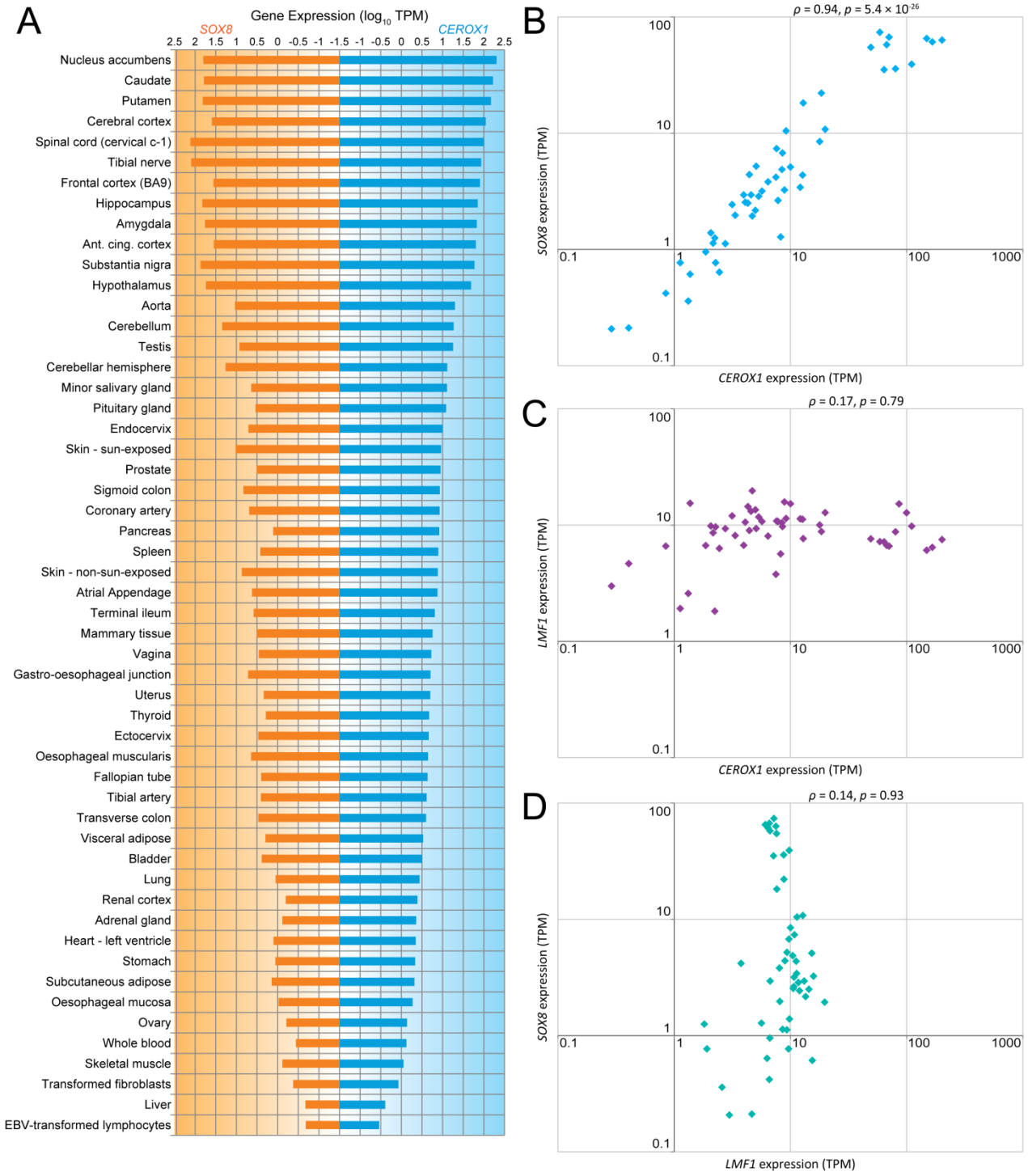


Figure 6.1: Expression of *CEROX1* and *SOX8* is highly correlated across human tissues. (A) The tissue-specific expression patterns of *CEROX1* and *SOX8*, as taken from the GTEx project[31], are strikingly similar when plotted aside one another. The data are ordered according to level of *CEROX1* expression, which is higher than that of *SOX8* in all tissues. **(B)** Spearman's rank correlation test indicated that the two datasets are highly positively correlated across the 50 tissues for which expression data is provided for both genes (ρ = 0.943; p = 1.62 × 10^{-26}). Comparisons of *CEROX1* and LMF1 expression **(C)**, and of *LMF1* and *SOX8* expression **(D)**, on the other hand, show no significant correlation (ρ = 0.156; p = 0.790 and ρ = 0.142; p = 0.927, respectively).

6.3 Identification of the *CEROX1-SOX8* promoter and analysis *in silico*

6.3.1 The core promoter was defined so as to include all major transcriptional start sites

As shown in Chapter 3, *CEROX1* encompasses a number of isoforms, resulting in part from multiple transcriptional start sites. In a thorough appraisal of isoform-specific transcriptional regulation, activity at each transcriptional start site would ideally be detectable via a separate reporter system. In this preliminary investigation, however, the promoter sequence for experimentation was putatively defined from the 5' end of *SOX8* to the 5' end of the downstream-most *CEROX1* isoform (Figure 6.2), allowing initiation of transcription at any start site in the *CEROX1*-oriented sequence to result in luciferase reporter synthesis.

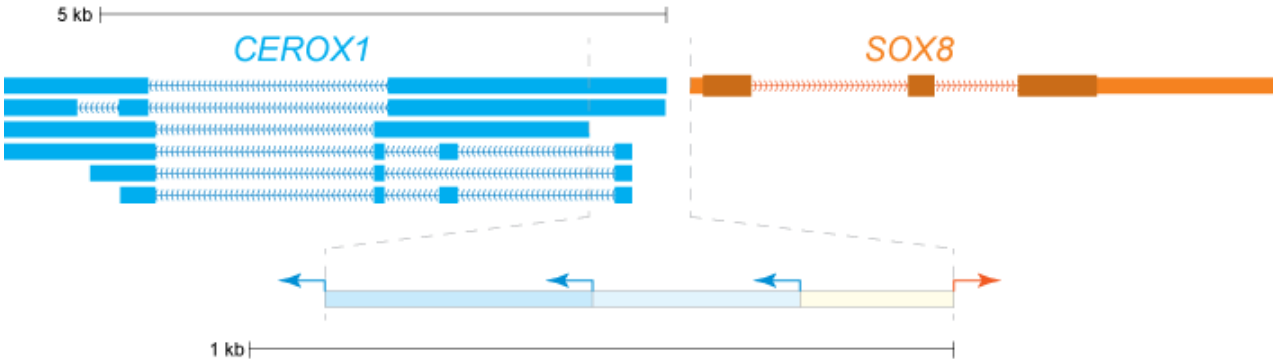


Figure 6.2: The *CEROX1-SOX8* promoter region used in this study. The sequence of interest was defined by the 5' terminus of *SOX8* (right), and that of the downstream-most isoform of *CEROX1* (left), so encapsulating the *SOX8* TSS, and three *CEROX1* TSSs (from GENCODE/Ensembl, but also including the two validated by FANTOM CAT/5' RACE in Chapter 3) in a sequence ~900 bp long.

6.3.2 Bioinformatic approaches identify conserved transcription factor binding sites

6.3.2.1 Many transcription factor binding sites are predicted in the putative *CEROX1-SOX8* promoter

Transcription factor binding sites are difficult to predict accurately, with high false positive and false negative rates. Even for those that are associated with consistent consensus binding motifs, there is often evidence for functional binding at only a fraction of the sites throughout the genome[32]. As a result, computational prediction of transcription factor binding sites has limited value in the absence of experimental contributions that identify protein-DNA interactions such as ChIP-seq and ChIA-PET. Predictions of transcription factor motifs within the *CEROX1-SOX8* promoter sequence were initially generated with several search programs: MATCH[33,34], PATCH[34] (gene-regulation.com/pub/programs.html),

and Signal Scan (www-bimas.cit.nih.gov/molbio/signal)[35,36]. Together these generated hundreds of unique and overlapping predictions.

6.3.2.2 Evolutionary conservation of core promoter elements reflects that of the *CEROX1* and *SOX8*

Alignment of the *CEROX1-SOX8* promoter region in vertebrates reveals that the *SOX8*-proximal end exhibits a greater degree of evolutionary alignment, indicating conservation back through mammals to birds and fish; this is in accordance with the conservation of the *SOX8* gene body, which is conserved in all vertebrates. The *CEROX1* end, on the other hand, shows recognisable alignment and conservation only in eutherian mammals, the clade within which sequence and expression of *CEROX1* is conserved.

As with all other genetic elements, evolutionary conservation of transcription factor binding motifs is suggestive of function. A list of all unique predicted binding factors was produced by combining the results of PATCH, MATCH, and Signal Scan searches. The conservation of the putative binding sites identified in the *CEROX1-SOX8* promoter was then assessed using ConTra v3[37-39], which visualises matches to known transcription factor sites within alignments of promoter regions from multiple species. This refinement revealed a number of conserved transcription factor binding site predictions throughout the *CEROX1-SOX8* promoter. One hundred conserved transcription factor sites, as analysed by ConTra, are shown in Figure 6.3. Not shown are additional sites for AP-1, AP-2(α), ELK1, EGR1, EGR2, SP1, WT1, ZNF740 and HOXB13, each of which had numerous conserved predictions throughout the *CEROX1-SOX8* sequence. AP-1 and SP1 are both common transcriptional promoter regulators, the latter comprising heterodimers of c-FOS, c-JUN, and ATF. Interestingly, EGR1 and EGR2 are important during development of the brain, and ELK1 is highly and ubiquitously expressed, but is enriched in the brain and testes, in correlation with *CEROX1*. The sequence contains three conserved binding predictions for HNF4 (hepatocyte nuclear factor 4), which is mostly expressed in the liver, one tissue from which *CEROX1* is conspicuously absent. This raises the question of whether factors such as HNF4 are involved in repression of transcription there, either by recruitment of a repressor or occlusion of another binding site required for transcriptional activation. CLOCK-BMAL is the master regulator of circadian rhythm, and the two predicted E-box CLOCK binding sites suggest that *CEROX1* and *SOX8* might be among the large proportion of genes that are subject to circadian modulation of expression[40].

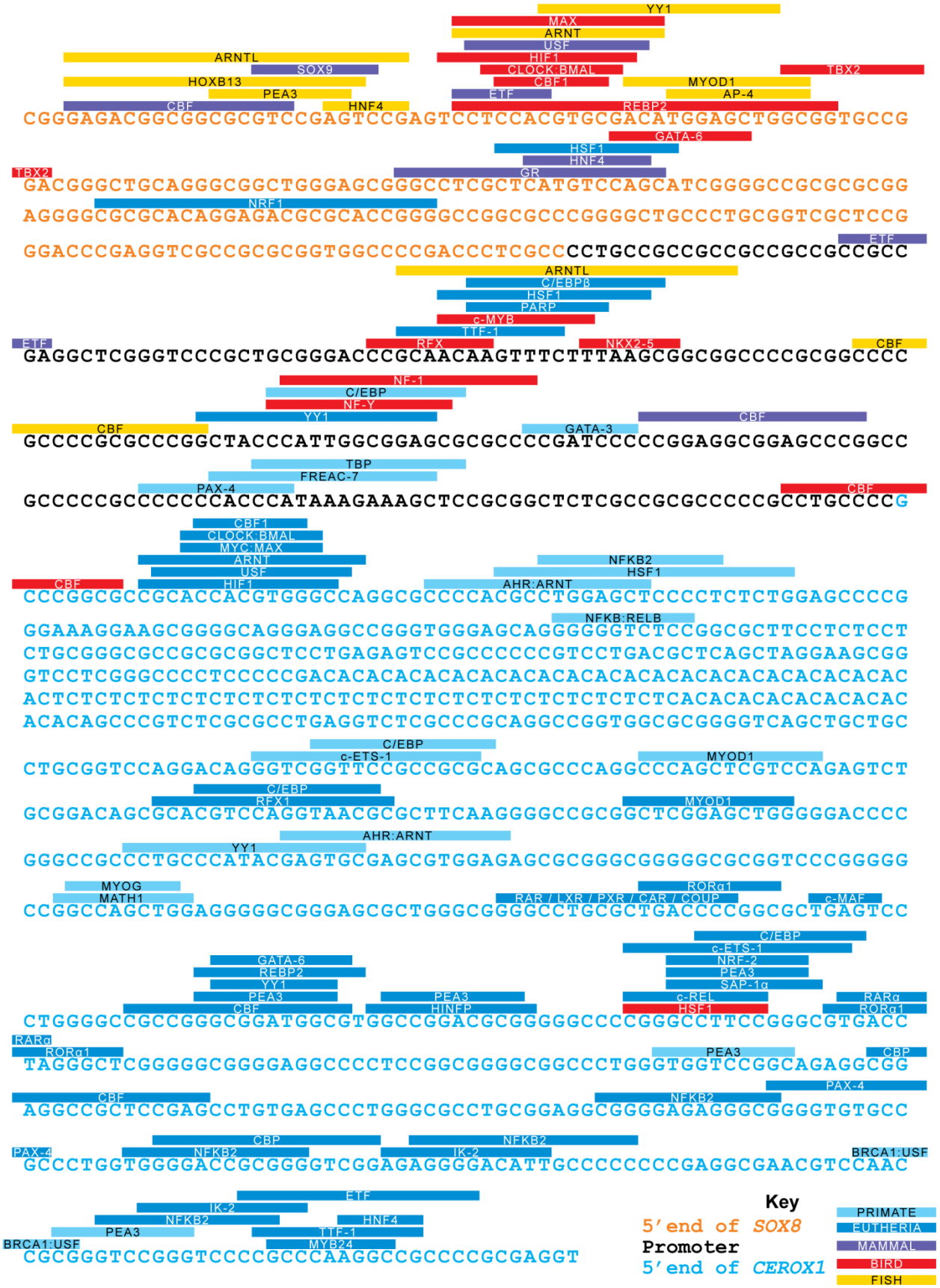


Figure 6.3: Conservation of predicted transcription factor binding sites within the *CEROX1-SOX8* promoter. Following the identification of putative transcription factors for which binding sites are present in the promoter region, ConTra v3 was used to assess conservation of each site. Displayed here above each sequence are conserved binding motifs. Each transcription factor is coloured according to the level of conservation, from primate-only and eutherian mammals-only, to all mammals, mammals and birds, or all vertebrates including fish (see key at bottom-right). In orange is the 5' sequence of *SOX8*, in blue the 5' sequence of *CEROX1*, including the three transcriptional start sites depicted in Figure 6.2.

Some motifs are highly conserved, and are consistent with the aforementioned correlation between gene and promoter conservation. For example, the sequence contains two canonical E-boxes (CACGTG), sequences containing which are responsible for binding c-MYC, MYOD, and HIF-1, and CLOCK-BMAL1, amid others. One E-box is located +200 relative to the *SOX8* TSS, and associated transcription factor sites are conserved throughout mammals and birds, and in some cases fish too. Interestingly, the other E-box is situated +15 relative to a *CEROX1* TSS, and is conserved only in eutherian mammals, like *CEROX1*, suggesting that one element may be associated with each *SOX8* and *CEROX1*. There is a general trend between length of conservation and whether elements are nearer *CEROX1* or *SOX8* (Figure 6.3). The first E-box is responsible for binding the *mrhl*-modulated binding of Max-Myc and Max-Mad complexes in mouse[26], and the presence of the alternative E-box nearer the *CEROX1* TSSs may explain why the effect of *mrhl* is specific to *Sox8*[26].

Aside from the speculative correlations mentioned here, the conclusions that may be drawn from these data alone are limited. When cross-correlated against experimental data, however, they will facilitate the formulation of hypotheses to explain the downstream mechanisms via which drugs and stimuli may affect *CEROX1* transcription.

6.4 The *CEROX1-SOX8* promoter was investigated experimentally using *in vitro* luciferase reporter assays

6.4.1 Luciferase assays provide a high-throughput measure of promoter-driven transcription

While assessing gene expression at the RNA level may provide some information about the responsivity of a given gene to stimuli, this approach does not directly quantify the transcriptional output of the promoter. RNAs are subject to numerous pathways of post-transcriptional regulation that influence stability and longevity, and so the steady-state level of any RNA in the cell is not necessarily directly proportional to the transcriptional activity at its promoter across its entire dynamic range of expression. To target this parameter more specifically, the bidirectional *CEROX1-SOX8* promoter sequence was coupled to a luciferase reporter enzyme, the biotechnologically engineered secreted NanoLuc (NLuc)[41]. Compared to traditional firefly luciferase (FLuc), this enzyme demonstrates higher signal-to-noise ratios, and crucially it is inhibited or stabilised by fewer compounds than that of almost all commercial luciferases, including firefly, *Renilla*, and *Luciola* enzymes[42].

6.4.2 The putative promoter demonstrates bidirectional transcriptional activity *in vitro*

Prior to investigating the responsivity of the putative *CEROX1-SOX8* bidirectional promoter element to cellular and biochemical stimuli, it was necessary to check that the putative sequence identified does indeed confer transcriptional activity to a chosen reporter construct. Three NLuc reporter constructs were prepared: two containing the core *CEROX1-SOX8* bidirectional promoter, in both orientations, and one empty promoter-less vector to act as a transcription negative control in case of leaky expression (Figure 6.4A). The three NLuc reporter constructs were transiently transfected into HEK293T cells, and activity of the two luciferase enzymes was measured using the Nano-Glo® Luciferase Assay system (Promega) (Figure 6.4B). Both the *CEROX1*-forward and *SOX8*-forward positions of the promoter region stimulated transcription of the NLuc gene above that of the empty vector, by factors of 26-fold and 352-fold, respectively (1-way ANOVA $F(2,15) = 3530$, $p < 10^{-5}$; Tukey's multiple comparisons test for each orientation against the empty construct, $p < 0.001$).

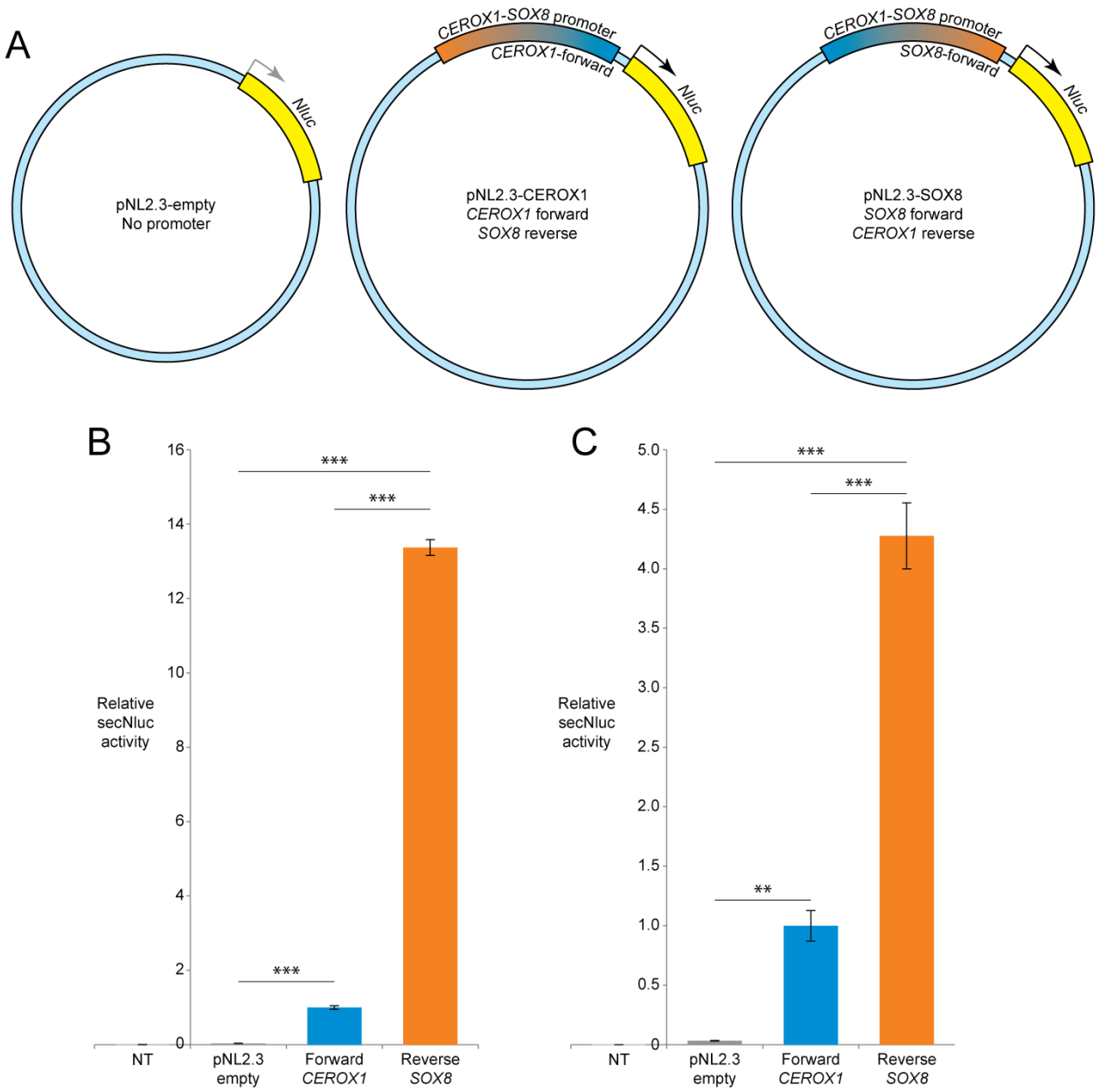


Figure 6.4: The *CEROX1-SOX8* promoter shows bidirectional transcription-promoting activity *in vitro*. (A) Three constructs were prepared, in which the NLuc reporter gene was promoter-less, or controlled by the upstream presence of *CEROX1-SOX8* promoter in either orientation. Expression of these constructs in **(B)** HEK293T cells and **(C)** SH-SY5Y cells resulted in synthesis of functional secreted secNLuc enzyme, which was detected in the extracellular medium by luminescence. In each cell line, the inclusion of the *CEROX1-SOX8* promoter region stimulated transcription significantly above the empty construct and non-transfected cells (NT). In both HEK293T and SH-SY5Y cells, transcription was promoted more strongly in the *SOX8*-forward direction, by 13.4 and 4.3 fold, respectively.

The transcriptional activity of the *SOX8*-forward (*CEROX1*-reverse) orientation was 13.4-fold higher than that of the *CEROX1*-forward orientation (Tukey's multiple comparisons test, $p < 0.001$) in this *in vitro* system (Figure 6.4B). This is at odds with the *in vivo* data presented above, which indicate that *CEROX1* is more highly expressed than *SOX8* at the mRNA level in all tissues (Figure 6.1).

6.4.2.2 SH-SY5Y cells recapitulate the transcriptional activities observed in HEK293T cells

In order to assess whether the apparent enrichment of *SOX8*-oriented transcription was cell type-specific, the constructs were also transfected into SH-SY5Y cells, and the assay repeated. Once again, the *CEROX1*-forward and *SOX8*-forward promoters elicited transcription of the luciferase construct above the empty control vector, by 30.2-fold and 129.3-fold (1-way ANOVA $F(2,12) = 158.7$, $p < 10^{-5}$; Tukey's multiple comparisons test for each promoter orientation against the empty construct, $p < 0.01$) (Figure 6.4C). In SH-SY5Y cells, the transcriptional activity of the *CEROX1*-forward orientation, relative to the empty construct, though the *SOX8*-forward orientation was still 4.3-fold higher than that of the *CEROX1*-forward orientation (Tukey's multiple comparisons test, $p < 0.001$).

6.4.2.3 The core promoter shows activity in spite of the absence of enhancers

The substantial transcription-driving activity of the core *CEROX1-SOX8* promoter *in vitro* suggests that the lack of *Sox8* promoter-driven expression observed in the absence of putative enhancer regions in mouse embryos by Guth et al.[25] is a developmental stage-specific phenomenon, consistent with the typically highly spatiotemporal specificity of enhancers[43]. This supports the inherent value of preliminarily investigating the core promoter in the absence of enhancer elements.

6.5 An initial screen of stimuli revealed several agents with putative *CEROX1-SOX8* promoter-regulating activity

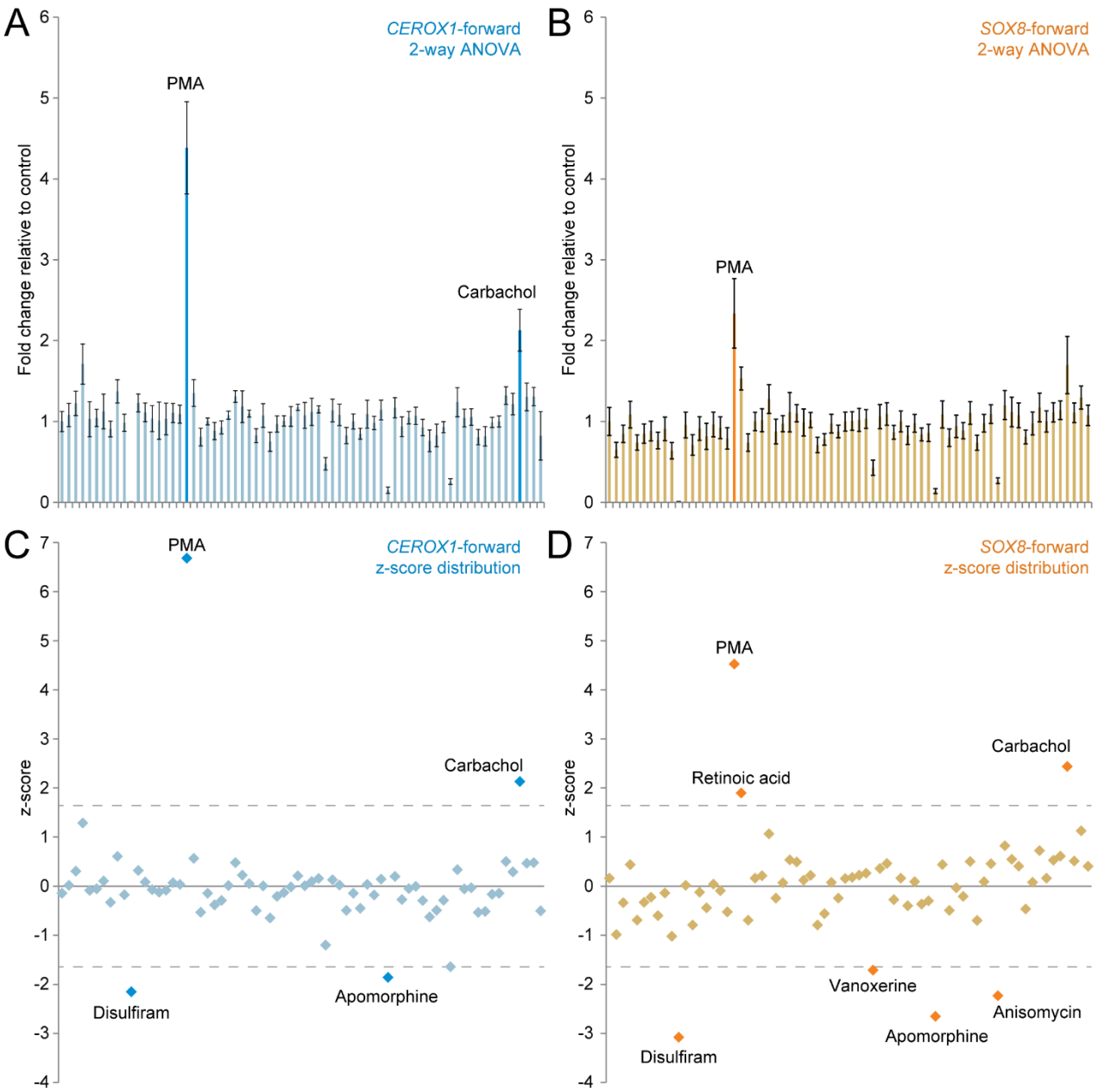
6.5.1 Choosing an *in vitro* system for studying the *CEROX1* promoter

SH-SY5Y cells, as a neuroblastoma cell line, are closer to the endogenous site of interest for *CEROX1* expression, the central nervous system. They are, therefore, more likely to host endogenous pathways of cell signalling and regulation that may influence *CEROX1* transcription, such as those transducing neurotransmitter signalling. For these reasons, SH-SY5Y cells were chosen as the model in which the regulation of the *CEROX1-SOX8* promoter should be studied.

SH-SY5Y sells were transfected with the NanoLuc® reporter constructs and then allowed to recover for 24 hours prior to treatment. Preliminarily, in order to ascertain key pathways involved in *CEROX1* regulation, the cells were exposed to a panel of 60 drugs, hormones, and growth factors, designed to target a diverse range of signalling pathways and key regulators. Concentrations were based upon EC_{50} and IC_{50} values, and examples in the literature; a full list of stimuli and concentrations may be found in Appendix A5.

6.5.2 Two statistical approaches identified significant changes in promoter activity

Following exposure to agents for 20 hours, media were assayed for secreted NLuc activity, which was normalised to that of untreated cells or cells treated with solvents only. In order to identify changes of interest in promoter activity, two metrics were used. First, a 2-way ANOVA was applied to the background-subtracted, control-normalised readings to look for specific significant differences between treatments (Figure 6.5A-B). Second, z-scores were calculated using the mean and standard deviation of all samples for each plate, and these were used to calculate the probability that an effect of that magnitude would be observed due to chance outside of an effect of the drug treatment (Figure 6.5C-D). Based upon these criteria, seven agents were identified as having a notable effect on either one or both of the forward and reverse *CEROX1-SOX8* promoter orientations (Figure 6.5E).



E

Compound	Target/effect	*CEROX1*-oriented			*SOX8*-oriented		
		Fold change	2-way ANOVA	Z-score	Fold change	2-way ANOVA	Z-score
Disulfiram	Inhibitor of aldehyde dehydrogenase	201.3	**	*	120.1		**
PMA	Activator of protein kinase C	4.4	****	***	2.3	****	***
Retinoic acid	Vitamin A metabolite; activator of retinoic acid receptors	1.4			1.5		*
Vanoxerine	Dopamine re-uptake inhibitor	2.1			2.3		*
Apomorphine	Non-specific dopamine receptor agonist	6.7		*	7.2		**
Anisomycin	Protein synthesis inhibitor, activator of JNK/SAPK and p38 MAPK	3.9			3.7		*
Carbachol	Muscarinic acetylcholine receptor agonist	2.1	***	*	1.7		**

Figure 6.5: Activity of the *CEROX1-SOX8* promoter was modulated by several exogenous agents, chiefly PMA. Following assay of secNLuc luminescence activity in the media of untreated and treated cells, relative fold changes with respect to untreated or control treated cells were calculated for both the *CEROX1*- **(A)** and the *SOX8*-orientation **(B)**, and a 2-way ANOVA applied to each. Dark blue and bright orange bars with labels indicate statistically significant changes. **(C,D)** Additionally, z-scores were calculated for each plate, converting the relative fold changes observed for each treatment into a distribution based upon overall standard deviations from the plate mean. Again, dark blue and bright orange data points represent significant z-scores, above and below the $p = 0.05$ threshold (dashed lines). Agents whose effects on *CEROX1*- or *SOX8*-oriented transcription were classified by these metrics as significant are summarised in **(E)**. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

6.5.2.2 Two candidates were excluded on the basis of conflating effects

The changes in luciferase activity displayed in Figure 6.5 do not account directly for cell number, though cells were plated at identical density in each well. In order to confirm specific effects of stimuli upon the *CEROX1-SOX8* promoter, any effects upon cell growth and number were assessed in a replica plate using a colourimetric Cell Proliferation Assay Kit that uses the rate of oxidation of the tetrazolium salt WST-1 to a formazan product as a proxy for cell number. This revealed that the dramatic decrease in promoter activity following exposure to the acetaldehyde dehydrogenase inhibitor disulfiram (120- to 200-fold) was a consequence of its toxicity, with few if any cells remaining alive. The six remaining compounds did not significantly affect proliferation. However, anisomycin, as well as stimulating the mitogen-activated kinases JNK and SAPK, is a potent inhibitor of protein synthesis, which might uncouple the process of luciferase transcription from detection by enzymatic activity. It is not possible therefore to conclude from these data whether the drop in promoter activity (3.7- to 3.9-fold) is due to a specific regulatory pathway, or simply dysfunctional luciferase synthesis.

6.5.3 Five compounds exhibited significant modulation of the *CEROX1-SOX8* promoter

Retinoic acid and vanoxerine each had a similar magnitude of effect on *CEROX1* and *SOX8*-oriented transcription (1.4- and 1.5-fold increases, and 2.1- and 2.3-fold increases, respectively), but only the effects upon *SOX8* were identified as statistically significant. However, the statistical tests here might not fully reflect the biological significance of these data, since they are also dependent upon variation in the effects exhibited across the entire screen. Three other drugs putatively demonstrated statistically significant specific effects on transcription from the *CEROX1-SOX8* locus in both orientations: phorbol 12-myristolate 13-acetate (PMA, also known as 12-O-tetradecanoylphorbol-13-acetate, TPA), carbamylcholine chloride (carbachol), and apomorphine hydrochloride.

6.6 Calcium signalling may regulate transcription of *CEROX1*

6.6.1 Two significant hits from the preliminary screen act on converging targets

Interestingly, carbachol and PMA act upon targets within the same pathway: calcium signalling. Carbachol is an agonist of muscarinic acetylcholine receptors (mAChRs), via which it stimulates phospholipase C (PLC) and the generation of the secondary messengers inositol-3,4,5-trisphosphate (IP_3) and diacylglycerol (DAG). Via the IP_3R1 receptor, IP_3 stimulates the release of calcium (Ca^{2+}) from the endoplasmic reticulum. Together, Ca^{2+} and DAG stimulate the activity of protein kinase C (PKC). PMA is an activator of PKC, and it therefore seems probable that both PMA and carbachol elicit their *CEROX1* promoter activation via signalling downstream of PKC (Figure 6.6A).

Aside from its stimulation by Ca^{2+} and DAG, PKC activity is dependent upon three activating phosphorylation events, the latter of which are autophosphorylatory. One of the kinases capable of initiating the first phosphorylation event is 3-phosphoinositide-dependent protein kinase-1 (PDK1)[44], which is one of the key regulators downstream of insulin signalling. While some activities of PDK1 are dependent on allosteric activation by 3-phosphoinositides, phosphorylation of PKC requires association with PIP_3 (phosphatidylinositol-3,4,5-trisphosphate) only in order to anchor the enzyme to the membrane, where PKC activation occurs[45]. PIP_3 is generated by phosphoinositide-3-kinase (PI3K), another key regulator downstream of insulin signalling, recruited via insulin receptor substrate 1 (IRS1) (Figure 6.6B). Incidentally, SH-SY5Y cells express the insulin receptor, and the molecule eliciting the third highest increase in *CEROX1*-oriented transcription was in fact insulin (1.7-fold increase, p = 0.15), suggesting that modulation of this pathway might also elicit a small effect on *CEROX1* via PKC or another pathway. This is of interest since insulin is an endogenous molecule, while carbachol and PMA are not (in animals).

If PMA and carbachol do act upon the *CEROX1-SOX8* promoter via a discrete common pathway, it would be expected that simultaneous treatment with both agents would elicit no additional stimulation over one alone. Indeed combinatorial treatment with PMA and carbachol was found to prompt no additional stimulation than PMA alone, suggesting that this is a maximal effect size for allosteric modulation of this signalling pathway (Figure 6.6C). PMA activates PKC by acting as a mimic of the endogenous allosteric regulator DAG, the generation of which is stimulated by carbachol. It therefore seems likely that combination of DAG and PMA does not have an additive effect.

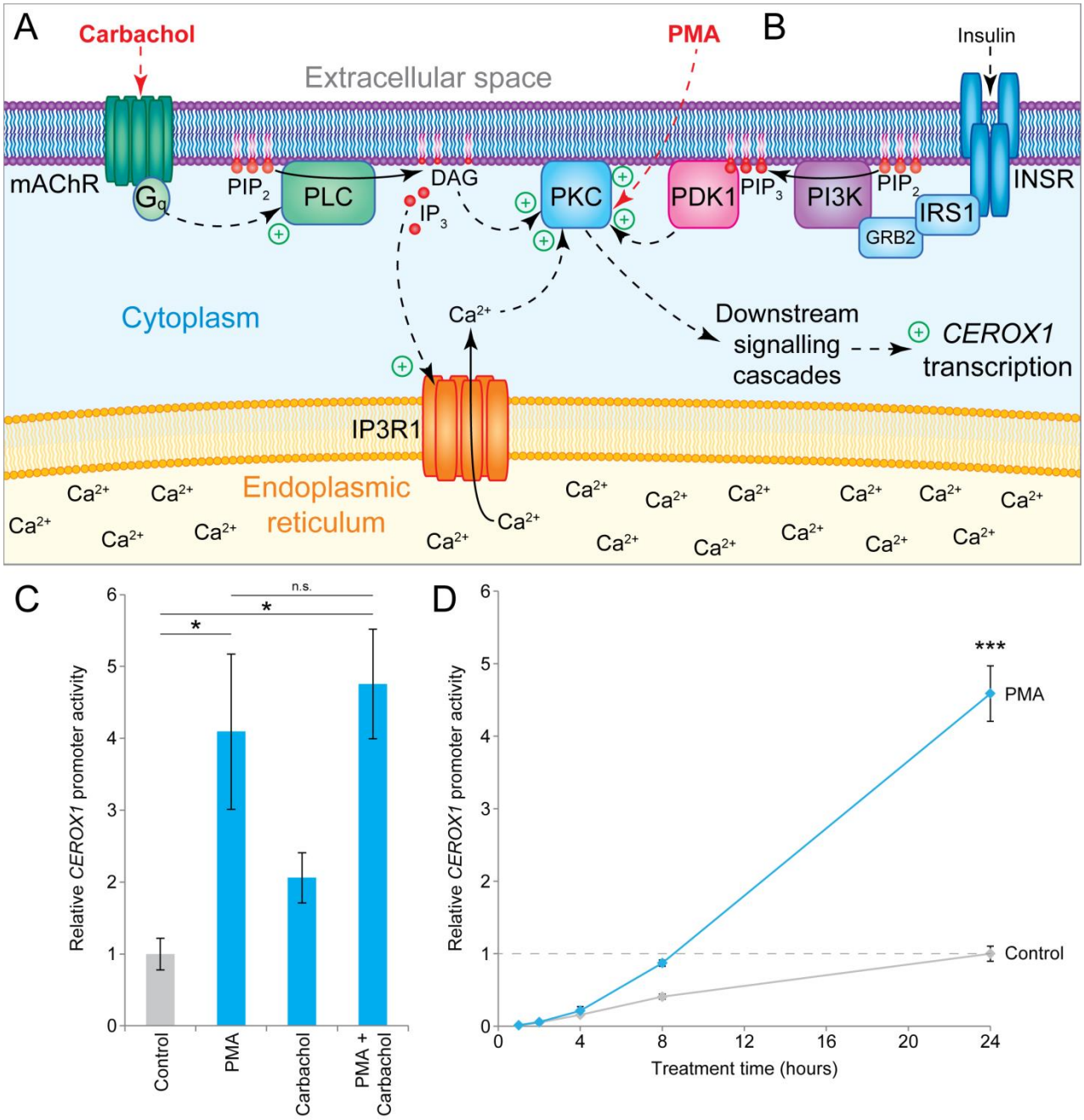


Figure 6.6: Calcium signalling mediated via PKC is capable of regulating the *CEROX1-SOX8* promoter. (A) Carbachol and PMA act upon convergent targets in the cell, with the former stimulating generation of DAG and release of calcium from the ER, both of which stimulate PKC, and the latter directly activating PKC. PKC is upstream of a multitude of signalling cascades, among which is a regulator(s) of *CEROX1* transcription. **(B)** Insulin is a physiological modulator of PKC activity, stimulating its maturation and activation via PI3K. **(C)** Combinatorial treatment of SH-SH5Y cells for 20 hours with carbachol and PMA does not result in additive upregulation of *CEROX1* transcription, supporting their targeting the same pathway. **(D)** Prolonged treatment with PMA, regardless of its subsequent induction of PKC deactivation and degradation, does not mitigate *CEROX1* transcriptional upregulation after up to 24 hours. The schematic in **(A)** was partly inspired by Becker, 2017[46].

6.6.1.2 PMA elicits sustained transcriptional modulation at the *CEROX1* promoter

Under activation by PMA, stimulation of PKC activity is rapid, achieving increased downstream signalling within 15 minutes[47]. However, activation of PKC by PMA also promotes its subsequent deactivation by de-phosphorylation and then degradation[48]. This raised the question of whether persistent PMA treatment incurs subsequent physiological feedback that tampers the effect after prolonged exposure. However, a time course of promoter activity indicates that, even after 24 hours, PMA treatment continues to activate *CEROX1* transcription (Figure 6.6D).

6.6.1.3 Transcriptional modulation at the *CEROX1* promoter is reflected by steady state RNA levels

While the luciferase assay here provides a readout of the effect of stimuli upon the transcriptional activity of the *CEROX1-SOX8* promoter, it does not account for the myriad post-transcriptional events that could dampen the effect of transcriptional modulation. In order to assess the effect of these drugs upon steady state *CEROX1* levels, treated cells were assayed by RT-qPCR, revealing a 1.9-fold increase in *CEROX1* expression following exposure to carbachol, and a 2.8-fold increase in response to PMA (Figure 6.7). As hypothesised, this indicates that post-transcriptional regulatory mechanisms partially temper the effect of transcriptional regulation and modulate RNA levels, but that sustained upregulation is still exhibited.

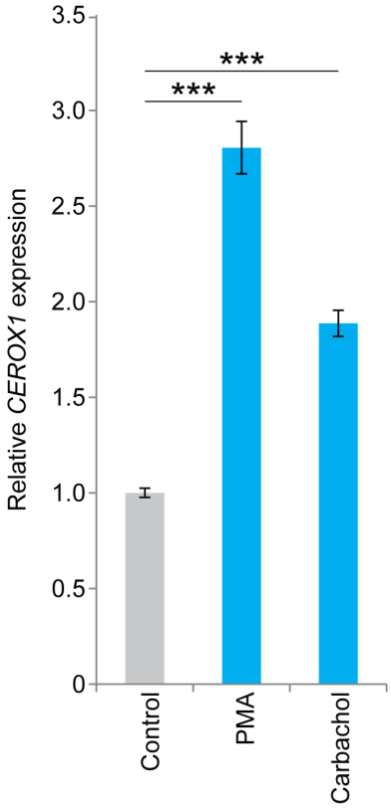


Figure 6.7: Transcription regulation identified by the luciferase reporter model reflects steady-state gene expression changes. *CEROX1* RNA levels in SH-SY5Y cells, assayed by RT-qPCR and normalised to those of *TUBA1A* and *POLR2A*, are increased 2.8- and 1.9-fold by treatment with PMA or carbachol for 20 hours, respectively.

DISCUSSION

The data generated here have confirmed facets of previous hypotheses about the *CEROX1-SOX8* promoter: it has been shown to initiate transcription bidirectionally *in vitro*, and in a manner that is responsive to cellular pathways; additionally, the conserved transcription factor binding sites predicted *in silico* are in accordance with the differential histories of the two genes, with *SOX8* being associated with more distally conserved elements. As shown by the study of CEROX1-ORF1 in Chapter 4, it is possible that *CEROX1* has very recently undergone evolution towards a bifunctional state, within the last ~10 My. Accordingly, some transcription factor binding sites, for FREAC-7 (FOXL1) for example, in the *CEROX1-SOX8* promoter are conserved only in primates, and it is tempting to speculate that these might be associated with the additional regulation of a secondary function of *CEROX1*, or even have permitted its evolution. The presence of certain transcription factor binding sites does correlate with the promoters of protein-coding and non-coding genes[49], and a role in mediating downstream translation cannot be ruled out.

In vitro modelling of promoter activities and regulation are limited, and it is notable that the failure of transfected DNA to behave as typical cellular chromatin may have effects on promoter behaviour and responsivity[50], since transcriptional modulation effected through chromatin restructuring and modification are not recapitulated. This provides one possible reason why *SOX8*-oriented transcription is promoted more heavily than the *CEROX1* direction *in vitro*: differential chromatin modifications across the locus may be required to recreate the *CEROX1*-favoured transcription seen *in vivo*. Alternatively, elements outside the region included in the experiments of this study may affect the directionality of the promoter. Overlooking the (limited) role of statistical significance in identifying biologically significant modulators, the agents whose application affected transcription at the *CEROX1-SOX8* locus in SH-SY5Y cells had a very similar magnitude of effect upon both *CEROX1*- and *SOX8*-oriented transcription, suggesting that elements involved in directional modulation are not present, or active in this system, or that the instigators of the pathways activated by the screen do not have a directional effect. The potential of the luciferase reporter would be enhanced using an episomal vector such as pREP4, which becomes chromatinized in cells[51,52]. Introduction of exogenous copies of a promoter region may not reproduce endogenous magnitudes of transcriptional regulation in response to a given stimulus or pathway, due to the introduction of competitive titration of transcription factors, typically scarce in comparison to the population of theoretical binding sites, between

the endogenous and additional binding sites[53]. However, the recapitulation of drug-induced effects on *CEROX1* promoter activity by RNA levels in non-transfected cells suggests that this is not an issue for the pathways identified here.

Some of the transcription factor binding elements predicted in the *CEROX1-SOX8* promoter sequence implicate tissue-specific factors whose regulation cannot be modelled in most cell types. For example, MyoD is a key myogenic regulatory factor involved in the differentiation of muscle tissue[54,55]. The activity of MyoD in promoting muscle differentiation may be augmented by the flavonoid hesperidin, which increases nuclear localisation and transcription factor activity of the MyoD protein[56]. Any effect of MyoD, demonstrated using hesperidin, for example, upon the *CEROX1-SOX8* promoter would therefore be absent in HEK293T or SH-SY5Y cells, which do not express MyoD endogenously (Human Protein atlas; proteinatlas.org).

Nevertheless, screening of the *CEROX1-SOX8* promoter-driven luciferase-reporter against a small drug library and subsequent validation of steady state *CEROX1* expression changes confirms its use as a tool assay for investigating the modulation of *CEROX1* transcriptionally *in vitro*. While time and resources limited the scope of the screen, in both the number of compounds included and using only one concentration and time point, the identification of PMA and carbachol as positive modulators in this assay highlights their use as tool compounds for investigating the downstream effects of *CEROX1* modulation with greater flexibility than transfection, and upon more physiological magnitudes. Additionally, these, and the other significant hits from the preliminary drug screen, in combination with the *in silico* identification of putative conserved transcription factor binding sites, have informed hypotheses of pathways that might regulate *CEROX1 in vivo*. Interestingly, intraocular application of carbachol is currently approved by the FDA for the treatment of glaucoma, as is the oral mAcR agonist pilocarpine. Highly specific novel M1 muscarinic agonists are being investigated as potential therapies targeting the cholinergic deficit that is inherent to the pathophysiology of Alzheimer's disease, and have been shown to modulate several key pathological hallmarks including tau and amyloid β plaques[57-59].

The downstream signalling of PKC is not well established, and it is possible that multiple pathways connect PKC and the *CEROX1-SOX8* promoter. There are ten isoforms of PKC, but it is likely that PKC α, encoded by *PRKCA*, is the mediator of the PMA and carbachol-induced effects seen in this study: it is modulated by calcium and DAG (and thereby the DAG mimic PMA), and is the most highly expressed isoform in SH-SY5Y

cells, by a margin of 4- to 11-fold (The SH-SY5Y Sequencing Project, systemsbiology.uni.lu/shsy5y; Human Protein Atlas, proteinatlas.org, respectively). It therefore seems logical to speculate that it might be the most important isoform in initiating downstream signalling towards the *CEROX1-SOX8* promoter. This may be investigated using an isoform-specific PKCα inhibitor such as Ro-32-043[60]. Few specific substrates for PKC phosphorylation are known, but include: RAF1, RAF kinase inhibitory protein (RKIP), GSK3β (glycogen synthase kinase 3β), and IκB kinase (IKK). Many substrates are common to some or all PKC isoforms, though the possibility of high levels of specificity have been demonstrated with synthetic peptides[61]. Pathways via which *CEROX1* might be regulated by PKC, as well as tools such as inhibitors that might be used to probe these, are described in Table 6.1, and depicted in Figures 6.8-6.9.

PKC target(s)	Pathway and possible downstream mediators at the *CEROX1-SOX8* promoter	Intervention strategies
RAF1 and RKIP	RAF1 is subject to multiple pathways of positive regulation by the PKC family[62]. PKCα directly stimulates RAF1 by phosphorylation at Ser499[63]. Additional Ser499-independent activation is also described[64]. RKIP is an endogenous binding partner of RAF1 that prevents association with its substrate MEK[65]. Phosphorylation of RKIP by PKC results in its dissociation from RAF1. PMA has been shown to induce the downstream signalling of RAF1[66] via a MAP kinase signalling cascade: RAF (MAP kinase kinase kinase) phosphorylates MEK1/2 (MAP kinase kinases), which in turn phosphorylate the MAPKs (mitogen-activated protein kinases) ERK1 and ERK2 (extracellular signal-regulated kinases). Via their target MAPKAP (MAPK-activated protein) kinases, such as MSK (mitogen- and stress-activated protein kinase) and RSK (ribosomal S6 kinase)[67], ERK1/2 regulate a broad variety of transcription factors, including CREB[68], ATF1[69], c-FOS[70], SRF[71], and c-MYC[72,73].	MEK inhibitors PD0325901[74] and PD184161[75] Mutation of transcription factor binding sites
IKK	PMA stimulates PKC phosphorylation of IKKα subunit[76]. Activation of IKK kinase activity leads to phosphorylation of IκB (inhibitor of kappa B) proteins. Subsequent degradation of IκB by the proteasome allows NF-κB to localise in the nucleus, where it acts as a transcription factor as a p52-RELB heterodimer[76]. Four consensus or near-consensus NF-κB sites (GGGRNYYYCC)[77]: GGGGGTCTCC; GGGCCCCTCC; GGGGGACCCC; and GGGGCCGGCC (non-aligning nucleotides in red).	IKK inhibitor BMS 345541[78] IKK and NF-κB inhibitor parthenolide[79] Mutation of NF-κB sites
GSK3	GSK3 has hundreds of proposed substrates[80,81] (reviewed Sutherland, 2011[82]). PMA-stimulated PKC activity results in inhibition of GSK3β via phosphorylation on Ser9[83,84]. Insulin-stimulated AKT (protein kinase B, PKB) activity phosphorylates the same residue[85,86]. GSK3β regulates the dual-function adhesion molecule and transcription factor β-catenin, and the master regulator of energy metabolism and protein synthesis the mTOR (mammalian target of rapamycin) complex.	Specific GSK3β inhibitor SB216763[87] mTOR inhibitor rapamycin

Table 6.1: Putative targets and downstream signalling of PKC, and strategies for further investigation of their relevance in the regulation of the *CEROX1-SOX8* promoter.

The identification of calcium signalling as a putative upstream regulator of *CEROX1* expression is highly consistent with the lncRNA's enrichment in the CNS, where calcium is a common second messenger, including in oligodendrocytes[88] and their precursory populations[89]. PMA and retinoic acid have been used individually and in concert to differentiate SH-SY5Y cells *in vitro*[90-92]. In OPCs, calcium signalling is modulated in response to neuronal activity, communicated via neurotransmitters including glutamate, GABA and

ATP[89]. Transient rises in intracellular calcium levels has been shown to swiftly elicit stimulation of myelination in developing oligodendrocytes[93], and approximately half of these occur in response to neuronal action potentials[93]. Glutamate activation of mGluR1 (metabotropic glutamate receptor 1) results in increased availability of calcium-permeable AMPA receptors, the increased influx of calcium through which stimulates the PI3K pathway[94]. This pathway was highlighted above as a possible mediator of the insulin-stimulated increase of *CEROX1* promoter activity, and has been shown to mediate retinoic acid-stimulated SH-SY5Y differentiation[95] and promote myelination *in vivo*[96]. Reduction of Erk1/2 and Akt/mTOR signalling has been associated with defective oligodendrocyte maturation in the *twitcher* mouse[97], which carries a mutation in the lysosomal enzyme galactosylceramidase (GALC), which is crucial in myelin metabolism. The upregulation of *CEROX1*/*Cerox1* in oligodendrocyte differentiation, and its putative role in supporting their highly energetic needs, may be regulated by these pathways, in doing so tuning neuronal needs with glial energy levels and support.

The ability of intra-mitochondrial calcium to upregulate enzymes of the TCA cycle is well characterised[98,99], and calcium signalling has been shown to upregulate the activity of the entire mitochondrial ETC by greater than 2-fold in both skeletal[100] and cardiac muscle[101]. Extra-mitochondrial calcium also stimulates the ETC, by regulation of glutamate import[102], and this is crucial in coupling neuronal stimulation to increased ATP provision by the ETC[103]. In rat neonatal myocytes, phorbol ester-stimulated PKC phosphorylation of the Cox4 subunit is associated with 2 to 4-fold activation of complex IV activity[104,105]. Activation of a putative ETC stimulant in the form of *CEROX1* would thereby be consistent with the existing role of calcium in regulating mitochondrial activity in response to cellular activity.

The possibility of regulation of *CEROX1* by β-catenin and its binding partner TCF4 was also suggested by the study that identified regulation of *Sox8* by the lncRNA *mrhl*[26]. In the testis and cell models examined, it was shown that a binding site for TCF4 upstream of the *Sox8* promoter, and in fact in the sequence of *Cerox1*, is constitutively bound by Tcf4, and that activation of Wnt signalling results in recruitment of β-catenin to that site, contributing to *Sox8* regulation. It was not assayed whether this signalling has an effect on *Cerox1* expression in this tissue via the TCF4/β-catenin element, though *Cerox1* expression is not changed by *mrhl* knockdown, unlike *Sox8* expression[27]. Likewise, examination of gene expression data in osteoblasts, in which *Sox8* is known to play a role in regulating differentiation[106], following Wnt3a stimulation also reveals a change in *Sox8* expression while *Cerox1* expression remains steady[107], suggesting that this pathway

exhibits a biased effect upon *Sox8* only. Interestingly, Wnt signalling has recently been implicated in the *in vitro* differentiation of Schwann cells[108].

Among the other three significant hits identified by the NLuc screen, hypothetical connections to the *CEROX1-SOX8* promoter may be formed based upon known pathways from the literature, and the computationally predicted transcription factor binding sites identified in this study. These are illustrated alongside putative calcium signalling pathways in Figures 6.8-6.9. For example, treatment of SH-SY5Y cells with apomorphine activates the NRF2-ARE pathway[109]. This pathway is involved in mounting an antioxidant response to oxidative stress; the ROS-diminishing activity of *Cerox1* would be consistent with activation by this pathway. NRF2 (also known as NFE2L2, nuclear factor erythroid 2-related factor 2) is among the transcription factors whose binding sites are overrepresented in bidirectional promoters[30]. The consensus sequence, based upon alignment of known elements, TMAnn**RTGAYnnnGC**Rwwww[110,111] (where the core is in bold, and essential nucleotides capitalised) is not found completely within the *CEROX1-SOX8* core promoter, though a near-match of the consensus core GTGACctaGG (non-aligning nucleotide in red) is found at +678/+12, relative to the upstream and downstream *CEROX1* transcriptional start sites, respectively. Unlike the solely exogenous vanoxerine or apomorphine, retinoic acid is an important endogenous metabolite of vitamin A. During embryogenesis, its activity as a paracrine signalling molecule plays an essential role in the regulation of HOX genes, amongst others, and subsequent embryonic anteroposterior positioning as well as development of the heart and eye[112,113]. Binding of retinoic acid to the retinoic acid receptor (RAR), which acts as a heterodimer with the retinoid X receptor (RXR), modulates its recruitment of co-activators and co-repressors at retinoic acid response elements (RAREs). Searching for the a consensus sequence RGKTSA(nnnnn)RGKTSA, derived by comparison of known DR5-RARE sites[114], within the *CEROX1-SOX8* promoter identified a near-match AGCGCA(cgtcc)AGGTAA, at +442/-209. Mutation or ablation of these putative elements would confirm whether apomorphine and retinoic acid responsivity is indeed mediated by these pathways.

The enzymatic nature of the luciferase reporter makes it susceptible to modulation by allosteric and competitive effects. High throughput screens of firefly luciferase (FLuc) activity in the presence of drug libraries shows that approximately 4-5% act as inhibitors[42,115]. Although, as mentioned in the rationale for its use as a reporter in this study, the number of compounds that act as NLuc inhibitors is fewer than that of other luciferases[42], it would be important to validate the effect of apomorphine, vanoxerine and retinoic

acid, as well as any subsequent modulators, using either an orthogonal assay with a different reporter (taking advantage of the small overlap between FLuc and NLuc inhibitors[42,115], for example), or ideally a secondary assay such as RT-qPCR. Looking ahead, these experiments are but the first step of an investigation into the regulation of *CEROX1* transcription. One point of primary importance for further studies would include the necessity of identification of distal regulatory elements, including the enhancers whose importance is highlighted by the *Sox8* promoter study of Guth et al.[25]. Highly informative experiments would include Capture-C and ChIA-PET, which would identify in high resolution the contacts that the *CEROX1-SOX8* locus makes with other loci and chromosomes, and which factors mediate these interactions.

As noted previously, the contribution of *in vivo* experimental data is essential in confirming *in silico* predictions and *in vitro* models. Following the use of ChIP and inhibitor studies to validate the predicted pathways of *CEROX1-SOX8* promoter regulation, recapitulation in a *Cerox1-Sox8* promoter reporter mouse cell line and subsequently in a mouse model or resulting primary cell culture would confirm both the evolutionary assumptions used in prioritising transcription factor binding site predictions and the *in vivo* relevance of the model here. Investigation of *CEROX1* levels in response to calcium-mediated neuron-oligodendrocyte signalling, especially, would identify whether the putative pathways identified in this chapter have relevance to the endogenous expression pattern that is so well conserved (Chapter 3).

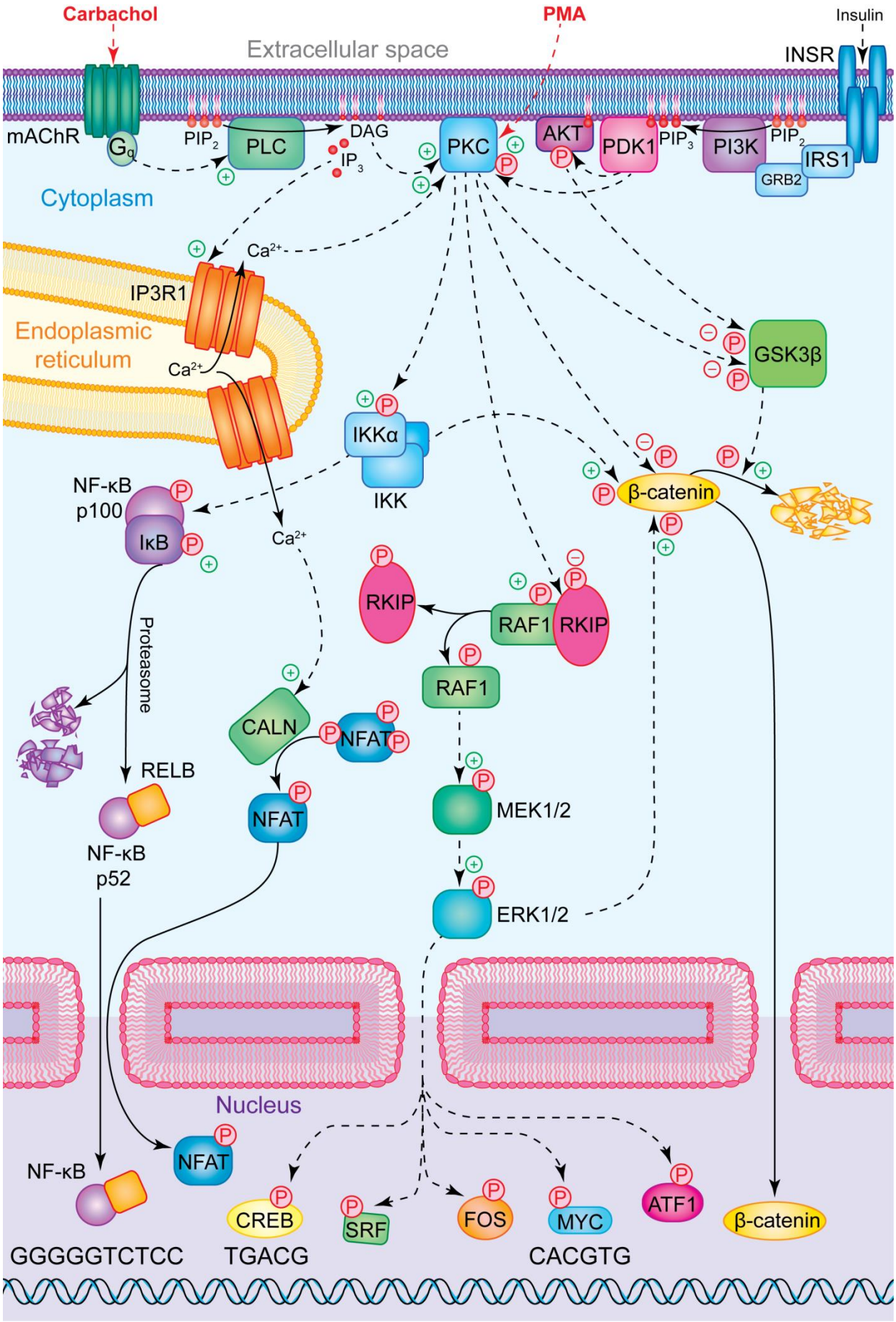


Figure 6.8: Examples of potential regulatory pathways via which calcium and PKC might act upon *CEROX1* transcription *in vitro*, part I. See text for details of NF-κB, RAF1-MEK-ERK, and GSK3β signalling pathways.

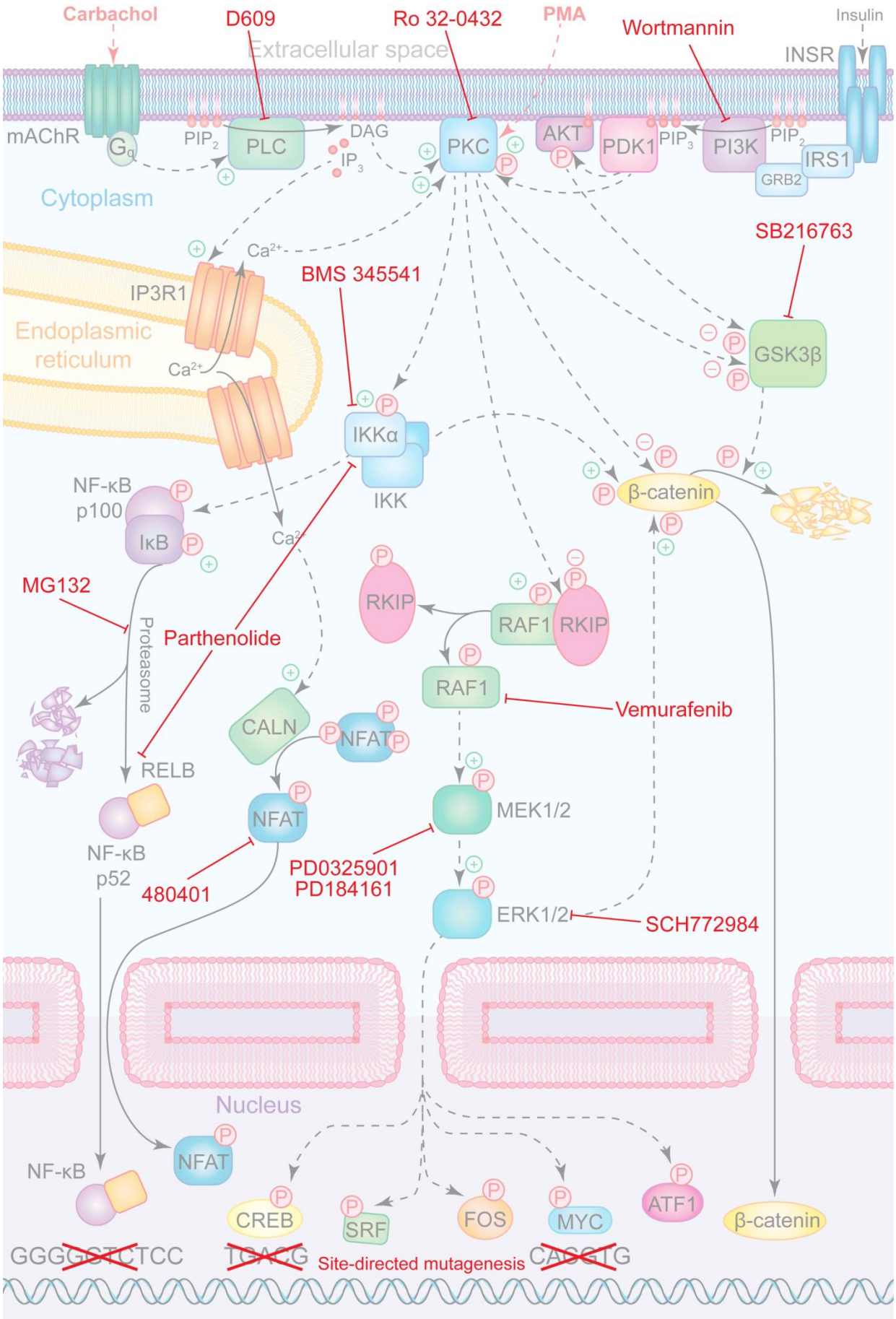


Figure 6.9: Examples of potential regulatory pathways via which calcium and PKC might act upon *CEROX1* transcription *in vitro*, part II. Many of the putative mediators of these signalling pathways have well-characterised inhibitors which may be used to interrogate their importance in the regulation of *CEROX1* transcription in this system.

REFERENCES

1. Blacker, T.S. & Duchen, M.R. Investigating mitochondrial redox state using NADH and NADPH autofluorescence. *Free Radical Biology and Medicine* **100**, 53-65 (2016).
2. Håndstad, T., Rye, M., Močnik, R., Drabløs, F. & Sætrom, P. Cell-type specificity of ChIP-predicted transcription factor binding sites. *BMC Genomics* **13**, 372-372 (2012).
3. Campbell, C.E., Piper, M., Plachez, C., Yeh, Y.T. *et al.* The transcription factor Nfix is essential for normal brain development. *BMC Dev Biol* **8**, 52 (2008).
4. Zhou, Q., Choi, G. & Anderson, D.J. The bHLH Transcription Factor Olig2 Promotes Oligodendrocyte Differentiation in Collaboration with Nkx2.2. *Neuron* **31**, 791-807 (2001).
5. Schenone, M., Dančík, V., Wagner, B.K. & Clemons, P.A. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology* **9**, 232 (2013).
6. Swinney, D.C. & Anthony, J. How were new medicines discovered? *Nature Reviews Drug Discovery* **10**, 507 (2011).
7. Wan, G.H., Hu, X.X., Liu, Y.H., Han, C. *et al.* A novel non-coding RNA lncRNA-JADE connects DNA damage signalling to histone H4 acetylation. *Embo Journal* **32**, 2833-2847 (2013).
8. Webster, G.A. & Perkins, N.D. Transcriptional Cross Talk between NF-κB and p53. *Molecular and Cellular Biology* **19**, 3485-3495 (1999).
9. Baud, V. & Karin, M. Is NF-kappaB a good target for cancer therapy? Hopes and pitfalls. *Nat Rev Drug Discov* **8**, 33-40 (2009).
10. Xia, Y.F., Shen, S. & Verma, I.M. NF-kappa B, an Active Player in Human Cancers. *Cancer Immunology Research* **2**, 823-830 (2014).
11. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497-1502 (2007).
12. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**, 11-24 (2012).
13. Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N. *et al.* Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276 (2012).
14. Hughes, J.R., Roberts, N., McGowan, S., Hay, D. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics* **46**, 205 (2014).
15. Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods* **13**, 74 (2015).
16. Li, G., Cai, L., Chang, H., Hong, P. *et al.* Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC Genomics* **15**, S11-S11 (2014).
17. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods* **10**, 1213-1218 (2013).
18. McNamara, C. & Winzeler, E.A. Target identification and validation of novel antimalarials. *Future Microbiology* **6**, 693-704 (2011).
19. Boosalis, M.S., Sangerman, J.I., White, G.L., Wolf, R.F. *et al.* Novel Inducers of Fetal Globin Identified through High Throughput Screening (HTS) Are Active In Vivo in Anemic Baboons and Transgenic Mice. *PLOS ONE* **10**, e0144660 (2016).
20. Dai, Y., Sangerman, J., Luo, H.Y., Fucharoen, S. *et al.* Therapeutic fetal-globin inducers reduce transcriptional repression in hemoglobinopathy erythroid progenitors through distinct mechanisms. *Blood Cells, Molecules, and Diseases* **56**, 62-69 (2016).
21. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell* **10**, 1453-1465 (2002).
22. Riddle, R.D., Johnson, R.L., Laufer, E. & Tabin, C. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**, 1401-16 (1993).
23. Bishop, C.E., Whitworth, D.J., Qin, Y., Agoulnik, A.I. *et al.* A transgenic insertion upstream of Sox9 is associated with dominant XX sex reversal in the mouse. *Nature Genetics* **26**, 490 (2000).
24. Jamieson, R.V., Perveen, R., Kerr, B., Carette, M. *et al.* Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum Mol Genet* **11**, 33-42 (2002).

25. Guth, S.I.E., Bösl, M.R., Sock, E. & Wegner, M. Evolutionary conserved sequence elements with embryonic enhancer activity in the vicinity of the mammalian Sox8 gene. *The International Journal of Biochemistry & Cell Biology* **42**, 465-471 (2010).
26. Kataruka, S., Akhade, V.S., Kayyar, B. & Rao, M.R.S. Mrhl Long Noncoding RNA Mediates Meiotic Commitment of Mouse Spermatogonial Cells by Regulating Sox8 Expression. *Molecular and Cellular Biology* **37**(2017).
27. Arun, G., Akhade, V.S., Donakonda, S. & Rao, M.R.S. mrhl RNA, a Long Noncoding RNA, Negatively Regulates Wnt Signaling through Its Protein Partner Ddx5/p68 in Mouse Spermatogonial Cells. *Molecular and Cellular Biology* **32**, 3140-3152 (2012).
28. Akhade, V.S., Arun, G., Donakonda, S. & Satyanarayana Rao, M.R. Genome wide chromatin occupancy of mrhl RNA and its role in gene regulation in mouse spermatogonial cells. *RNA Biology* **11**, 1262-1279 (2014).
29. Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Research* **14**, 62-66 (2004).
30. Lin, J.M., Collins, P.J., Trinklein, N.D., Fu, Y. *et al.* Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res* **17**, 818-27 (2007).
31. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580 (2013).
32. Grossman, S.R., Zhang, X., Wang, L., Engreitz, J. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci U S A* **114**, E1291-E1300 (2017).
33. Kel, A.E., Gößling, E., Reuter, I., Cheremushkin, E. *et al.* MATCH(TM): a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* **31**, 3576-3579 (2003).
34. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I. *et al.* TRANSFAC(®) and its module TRANSCompel(®): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108-D110 (2006).
35. Prestridge, D.S. Signal Scan - a Computer-Program That Scans DNA-Sequences for Eukaryotic Transcriptional Elements. *Computer Applications in the Biosciences* **7**, 203-206 (1991).
36. Prestridge, D.S. SIGNAL SCAN 4.0: additional databases and sequence formats. *Bioinformatics* **12**, 157-160 (1996).
37. Hooghe, B., Hulpiau, P., van Roy, F. & De Bleser, P. ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species. *Nucleic Acids Research* **36**, W128-W132 (2008).
38. Broos, S., Hulpiau, P., Galle, J., Hooghe, B. *et al.* ConTra v2: a tool to identify transcription factor binding sites across species, update 2011. *Nucleic Acids Research* **39**, W74-W78 (2011).
39. Kreft, Ł., Soete, A., Hulpiau, P., Botzki, A. *et al.* ConTra v3: a tool to identify transcription factor binding sites across species, update 2017. *Nucleic Acids Research* **45**, W490-W494 (2017).
40. Zhang, R., Lahens, N.F., Ballance, H.I., Hughes, M.E. & Hogenesch, J.B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences* **111**, 16219-16224 (2014).
41. Hall, M.P., Unch, J., Binkowski, B.F., Valley, M.P. *et al.* Engineered Luciferase Reporter from a Deep Sea Shrimp Utilizing a Novel Imidazopyrazinone Substrate. *ACS Chemical Biology* **7**, 1848-1857 (2012).
42. Ho, P.I., Yue, K., Pandey, P., Breault, L. *et al.* Reporter enzyme inhibitor study to aid assembly of orthogonal reporter gene assays. *ACS Chem Biol* **8**, 1009-17 (2013).
43. Visel, A., Blow, M.J., Li, Z.R., Zhang, T. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).
44. Dutil, E.M., Toker, A. & Newton, A.C. Regulation of conventional protein kinase C isozymes by phosphoinositide-dependent kinase 1 (PDK-1). *Current Biology* **8**, 1366-1375 (1998).
45. Sonnenburg, E.D., Gao, T. & Newton, A.C. The Phosphoinositide-dependent Kinase, PDK-1, Phosphorylates Conventional Protein Kinase C Isozymes by a Mechanism That Is Independent of Phosphoinositide 3-Kinase. *Journal of Biological Chemistry* **276**, 45289-45297 (2001).
46. Becker, E.B.E. From Mice to Men: TRPC3 in Cerebellar Ataxia. *The Cerebellum* **16**, 877-879 (2017).
47. Liu, M., Clarke, C.J., Salama, M.F., Choi, Y.J. *et al.* Co-ordinated activation of classical and novel PKC isoforms is required for PMA-induced mTORC1 activation. *PLOS ONE* **12**, e0184818 (2017).
48. Wang, Y., Wang, Y., Zhang, H., Gao, Y. *et al.* Sequential posttranslational modifications regulate PKC degradation. *Molecular Biology of the Cell* **27**, 410-420 (2016).

49. Alam, T., Medvedeva, Y.A., Jia, H., Brown, J.B. *et al.* Promoter Analysis Reveals Globally Differential Regulation of Human Long Non-Coding RNA and Protein-Coding Genes. *Plos One* **9**(2014).
50. Smith, C.L. & Hager, G.L. Transcriptional regulation of mammalian genes in vivo. A tale of two templates. *Journal of Biological Chemistry* **272**, 27493-27496 (1997).
51. Inayoshi, Y., Miyake, K., Machida, Y., Kaneoka, H. *et al.* Mammalian Chromatin Remodeling Complex SWI/SNF Is Essential for Enhanced Expression of the Albumin Gene during Liver Development. *The Journal of Biochemistry* **139**, 177-188 (2006).
52. Liu, R., Liu, H., Chen, X., Kirby, M. *et al.* Regulation of CSF1 Promoter by the SWI/SNF-like BAF Complex. *Cell* **106**, 309-318 (2001).
53. Brewster, Robert C., Weinert, Franz M., Garcia, Hernan G., Song, D. *et al.* The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Cell* **156**, 1312-1323 (2014).
54. Davis, R.L., Weintraub, H. & Lassar, A.B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987-1000 (1987).
55. Perry, R.L. & Rudnick, M.A. Molecular mechanisms regulating myogenic determination and differentiation. *Front Biosci* **5**, D750-67 (2000).
56. Jeong, H., Lee, J.Y., Jang, E.J., Lee, E.H. *et al.* Hesperedin promotes MyoD-induced myogenic differentiation in vitro and in vivo. *British Journal of Pharmacology* **163**, 598-608 (2011).
57. Fisher, A., Brandeis, R., Bar-Ner, R.H.N., Kliger-Spatz, M. *et al.* AF150(S) and AF267B - M1 muscarinic agonists as innovative therapies for Alzheimer's disease. *Journal of Molecular Neuroscience* **19**, 145-153 (2002).
58. Caccamo, A., Oddo, S., Billings, L.M., Green, K.N. *et al.* M1 receptors play a central role in modulating AD-like pathology in transgenic mice. *Neuron* **49**, 671-82 (2006).
59. Fisher, A. M1 muscarinic agonists target major hallmarks of Alzheimer's disease--the pivotal role of brain M1 receptors. *Neurodegener Dis* **5**, 237-40 (2008).
60. Wilkinson, S.E., Parker, P.J. & Nixon, J.S. Isoenzyme specificity of bisindolylmaleimides, selective inhibitors of protein kinase C. *Biochemical Journal* **294**, 335-337 (1993).
61. Kang, J.H., Asai, D., Yamada, S., Toita, R. *et al.* A short peptide is a protein kinase C (PKC) alpha-specific substrate. *Proteomics* **8**, 2006-11 (2008).
62. Sozeri, O., Vollmer, K., Liyanage, M., Frith, D. *et al.* Activation of the c-Raf protein kinase by protein kinase C phosphorylation. *Oncogene* **7**, 2259-62 (1992).
63. Kolch, W., Heidecker, G., Kochs, G., Hummel, R. *et al.* Protein kinase Cα activates RAF-1 by direct phosphorylation. *Nature* **364**, 249 (1993).
64. Yip-Schneider, M.T., Miao, W.Y., Lin, A., Barnard, D.S. *et al.* Regulation of the Raf-1 kinase domain by phosphorylation and 14-3-3 association. *Biochemical Journal* **351**, 151-159 (2000).
65. Yeung, K., Janosch, P., McFerran, B., Rose, D.W. *et al.* Mechanism of Suppression of the Raf/MEK/Extracellular Signal-Regulated Kinase Pathway by the Raf Kinase Inhibitor Protein. *Molecular and Cellular Biology* **20**, 3079-3085 (2000).
66. Corbit, K.C., Trakul, N., Eves, E.M., Diaz, B. *et al.* Activation of Raf-1 Signaling by Protein Kinase C through a Mechanism Involving Raf Kinase Inhibitory Protein. *Journal of Biological Chemistry* **278**, 13061-13068 (2003).
67. Carriere, A., Ray, H., Blenis, J. & Roux, P.P. The RSK factors of activating the Ras/MAPK signaling cascade. *Frontiers in Bioscience* **13**, 4258-4275 (2008).
68. Xing, J., Ginty, D.D. & Greenberg, M.E. Coupling of the RAS-MAPK Pathway to Gene Activation by RSK2, a Growth Factor-Regulated CREB Kinase. *Science* **273**, 959-963 (1996).
69. Gupta, P. & Prywes, R. ATF1 Phosphorylation by the ERK MAPK Pathway Is Required for Epidermal Growth Factor-induced c-jun Expression. *Journal of Biological Chemistry* **277**, 50550-50556 (2002).
70. Chen, R.H., Abate, C. & Blenis, J. Phosphorylation of the c-Fos transrepression domain by mitogen-activated protein kinase and 90-kDa ribosomal S6 kinase. *Proceedings of the National Academy of Sciences* **90**, 10952-10956 (1993).
71. Hauge, C. & Frödin, M. RSK and MSK in MAP kinase signalling. *Journal of Cell Science* **119**, 3021-3023 (2006).
72. Sears, R., Leone, G., DeGregori, J. & Nevins, J.R. Ras Enhances Myc Protein Stability. *Molecular Cell* **3**, 169-179 (1999).
73. Sears, R., Nuckolls, F., Haura, E., Taya, Y. *et al.* Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. *Genes & Development* **14**, 2501-2514 (2000).

74. Barrett, S.D., Bridges, A.J., Dudley, D.T., Saltiel, A.R. *et al.* The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901. *Bioorganic & Medicinal Chemistry Letters* **18**, 6501-6504 (2008).
75. Klein, P.J., Schmidt, C.M., Wiesenauer, C.A., Choi, J.N. *et al.* The Effects of a Novel MEK Inhibitor PD184161 on MEK-ERK Signaling and Growth in Human Liver Cancer. *Neoplasia (New York, N.Y.)* **8**, 1-8 (2006).
76. Leonard, B., McCann, J.L., Starrett, G.J., Kosyakovsky, L. *et al.* The PKC-NFκB Signaling Pathway Induces APOBEC3B Expression in Multiple Human Cancers. *Cancer research* **75**, 4538-4547 (2015).
77. Wan, F. & Lenardo, M.J. Specification of DNA Binding Activity of NF-κB Proteins. *Cold Spring Harbor Perspectives in Biology* **1**, a000067 (2009).
78. Burke, J.R., Pattoli, M.A., Gregor, K.R., Brassil, P.J. *et al.* BMS-345541 Is a Highly Selective Inhibitor of IκB Kinase That Binds at an Allosteric Site of the Enzyme and Blocks NF-κB-dependent Transcription in Mice. *Journal of Biological Chemistry* **278**, 1450-1456 (2003).
79. Mathema, V.B., Koh, Y.-S., Thakuri, B.C. & Sillanpää, M. Parthenolide, a Sesquiterpene Lactone, Expresses Multiple Anti-cancer and Anti-inflammatory Activities. *Inflammation* **35**, 560-565 (2012).
80. Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A.T.M. *et al.* Systematic Discovery of In Vivo Phosphorylation Networks. *Cell* **129**, 1415-1426 (2007).
81. Beurel, E., Grieco, S.F. & Jope, R.S. Glycogen synthase kinase-3 (GSK3): Regulation, actions, and diseases. *Pharmacology & Therapeutics* **148**, 114-131 (2015).
82. Sutherland, C. What are the bona fide GSK3 Substrates? *International Journal of Alzheimer's Disease* **2011**, 23 (2011).
83. Cook, D., Fry, M.J., Hughes, K., Sumathipala, R. *et al.* Wingless inactivates glycogen synthase kinase-3 via an intracellular signalling pathway which involves a protein kinase C. *The EMBO Journal* **15**, 4526-4536 (1996).
84. Chen, R.H., Ding, W.V. & McCormick, F. Wnt signaling to beta-catenin involves two interactive components - Glycogen synthase kinase-3 beta inhibition and activation of protein kinase C. *Journal of Biological Chemistry* **275**, 17894-17899 (2000).
85. Cross, D.A.E., Alessi, D.R., Cohen, P., Andjelkovich, M. & Hemmings, B.A. Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B. *Nature* **378**, 785 (1995).
86. Shaw, M., Cohen, P. & Alessi, D.R. Further evidence that the inhibition of glycogen synthase kinase-3β by IGF-1 is mediated by PDK1/PKB-induced phosphorylation of Ser-9 and not by dephosphorylation of Tyr-216. *FEBS Letters* **416**, 307-311 (1997).
87. Coghlan, M.P., Culbert, A.A., Cross, D.A.E., Corcoran, S.L. *et al.* Selective small molecule inhibitors of glycogen synthase kinase-3 modulate glycogen metabolism and gene transcription. *Chemistry & Biology* **7**, 793-803 (2000).
88. Kettenmann, H., Kirishchuk, S. & Verkhratskii, A. Calcium signalling in oligodendrocytes. *Neurophysiology* **26**, 21-25 (1994).
89. Pitman, K.A. & Young, K.M. Activity-dependent calcium signalling in oligodendrocyte generation. *The International Journal of Biochemistry & Cell Biology* **77**, 30-34 (2016).
90. Matsumoto, K., Wada, R.K., Yamashiro, J.M., Kaplan, D.R. & Thiele, C.J. Expression of Brain-Derived Neurotrophic Factor and P145(Trkb) Affects Survival, Differentiation, and Invasiveness of Human Neuroblastoma-Cells. *Cancer Research* **55**, 1798-1806 (1995).
91. Påhlman, S., Odelstad, L., Larsson, E., Grotte, G. & Nilsson, K. Phenotypic changes of human neuroblastoma cells in culture induced by 12-O-tetradecanoyl-phorbol-13-acetate. *Int J Cancer* **28**, 583-589 (1981).
92. Presgraves, S.P., Ahmed, T., Borwege, S. & Joyce, J.N. Terminally differentiated SH-SY5Y cells provide a model system for studying neuroprotective effects of dopamine agonists. *Neurotoxicity Research* **5**, 579-598 (2003).
93. Krasnow, A.M., Ford, M.C., Valdivia, L.E., Wilson, S.W. & Attwell, D. Regulation of developing myelin sheath elongation by oligodendrocyte calcium transients in vivo. *Nature Neuroscience* **21**, 24-28 (2018).
94. Zonouzi, M., Renzi, M., Farrant, M. & Cull-Candy, S.G. Bidirectional plasticity of calcium-permeable AMPA receptors in oligodendrocyte lineage cells. *Nature Neuroscience* **14**, 1430 (2011).
95. López-Carballo, G., Moreno, L., Masiá, S., Pérez, P. & Barettino, D. Activation of the Phosphatidylinositol 3-Kinase/Akt Signaling Pathway by Retinoic Acid Is Required for Neural Differentiation of SH-SY5Y Human Neuroblastoma Cells. *Journal of Biological Chemistry* **277**, 25297-25304 (2002).

96. Goebbels, S., Oltrogge, J.H., Kemper, R., Heilmann, I. *et al.* Elevated Phosphatidylinositol 3,4,5-Trisphosphate in Glia Triggers Cell-Autonomous Membrane Wrapping and Myelination. *The Journal of Neuroscience* **30**, 8953-8964 (2010).
97. Inamura, N., Kito, M., Go, S., Kishi, S. *et al.* Developmental defects and aberrant accumulation of endogenous psychosine in oligodendrocytes in a murine model of Krabbe disease. *Neurobiology of Disease* (2018).
98. McCormack, J.G. & Denton, R.M. The role of Ca2+ ions in the regulation of intramitochondrial metabolism and energy production in rat heart. *Molecular and Cellular Biochemistry* **89**, 121-125 (1989).
99. Denton, R.M. Regulation of mitochondrial dehydrogenases by calcium ions. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1787**, 1309-1316 (2009).
100. Glancy, B., Willis, W.T., Chess, D.J. & Balaban, R.S. Effect of Calcium on the Oxidative Phosphorylation Cascade in Skeletal Muscle Mitochondria. *Biochemistry* **52**, 2793-2809 (2013).
101. Covian, R., French, S., Kusnetz, H. & Balaban, R.S. Stimulation of oxidative phosphorylation by calcium in cardiac mitochondria is not influenced by cAMP and PKA activity. *Biochimica et biophysica acta* **1837**, 1913-1921 (2014).
102. Gellerich, F.N., Gizatullina, Z., Trumbeckaite, S., Nguyen, H.P. *et al.* The regulation of OXPHOS by extramitochondrial calcium. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1797**, 1018-1027 (2010).
103. Llorente-Folch, I., Rueda, C.B., Amigo, I., del Arco, A. *et al.* Calcium-Regulation of Mitochondrial Respiration Maintains ATP Homeostasis and Requires ARALAR/AGC1-Malate Aspartate Shuttle in Intact Cortical Neurons. *J Neurosci* **33**, 13957-13971 (2013).
104. Ogbi, M., Chew, Catherine S., Pohl, J., Stuchlik, O. *et al.* Cytochrome c oxidase subunit IV as a marker of protein kinase Cε function in neonatal cardiac myocytes: implications for cytochrome c oxidase activity. *Biochemical Journal* **382**, 923-932 (2004).
105. Ogbi, M. & Johnson, John A. Protein kinase Cϵ interacts with cytochrome c oxidase subunit IV and enhances cytochrome c oxidase activity in neonatal cardiac myocyte preconditioning. *Biochem J* **393**, 191-199 (2006).
106. Schmidt, K., Schinke, T., Haberland, M., Priemel, M. *et al.* The high mobility group transcription factor Sox8 is a negative regulator of osteoblast differentiation. *The Journal of Cell Biology* **168**, 899-910 (2005).
107. Sebastian, A., Hum, N.R., Murugesh, D.K., Hatsell, S. *et al.* Wnt co-receptors Lrp5 and Lrp6 differentially mediate Wnt3a signaling in osteoblasts. *PLOS ONE* **12**, e0188264 (2017).
108. Dapeng, L., Xiaojie, L., Yan, D. & Gang, S. The Role of Wnt/β-Catenin Signaling Pathway in the Transdifferentiation from Periodontal Ligament Stem Cells to Schwann Cells. *Cellular Reprogramming* **19**, 384-388 (2017).
109. Hara, H., Ohta, M. & Adachi, T. Apomorphine protects against 6-hydroxydopamine-induced neuronal cell death through activation of the Nrf2-ARE pathway. *J Neurosci Res* **84**, 860-6 (2006).
110. Wasserman, W.W. & Fahl, W.E. Functional antioxidant responsive elements. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5361-5366 (1997).
111. Nioi, P., McMahon, M., Itoh, K., Yamamoto, M. & Hayes, J.D. Identification of a novel Nrf2-regulated antioxidant response element (ARE) in the mouse NAD(P)H:quinone oxidoreductase 1 gene: reassessment of the ARE consensus sequence. *Biochemical Journal* **374**, 337-348 (2003).
112. Duester, G. Retinoic Acid Synthesis and Signaling during Early Organogenesis. *Cell* **134**, 921-931 (2008).
113. Niederreither, K., Subbarayan, V., Dolle, P. & Chambon, P. Embryonic retinoic acid synthesis is essential for early mouse post-implantation development. *Nat Genet* **21**, 444-8 (1999).
114. Lalevée, S., Anno, Y.N., Chatagnon, A., Samarut, E. *et al.* Genome-wide in Silico Identification of New Conserved and Functional Retinoic Acid Receptor Response Elements (Direct Repeats Separated by 5 bp). *Journal of Biological Chemistry* **286**, 33322-33334 (2011).
115. Auld, D.S. & Inglese, J. Interferences with Luciferase Reporter Enzymes. in *Assay Guidance Manual* (eds. Sittampalam, G.S. *et al.*) (Bethesda (MD), 2018).

CHAPTER 7: DISCUSSION

7.1 Summary of work: a holistic approach to lncRNA investigation

Throughout this thesis, I have examined many aspects of *CEROX1*, taking a multifaceted approach to build a model of its predicted functionality. In Chapter 3, I showed that *CEROX1* demonstrates a high level of evolutionary constraint, being conserved at the DNA and RNA transcript levels through Eutheria and therefore more highly than most human lncRNAs. My singularly extensive examination of its evolutionary tissue-specific expression pattern, revealing long-standing enrichment in the central nervous system, suggests an important function. The expression of *CEROX1* is highly dynamic throughout both human and murine brain development, steadily increasing in the human adult brain to become one of the most highly expressed genes in the transcriptome. Specifically, *CEROX1* is enriched in oligodendrocyte lineages, in which its expression again increases throughout differentiation, becoming exceptionally high in myelinating cells. I therefore hypothesise that *CEROX1* plays a role in the development of the central nervous system via a role in the biology of oligodendrocytes. The work of Chapter 6 provided a possible underlying pathway of regulation contributing to this dynamic expression: I discovered that *CEROX1* transcription and steady state expression are upregulated *in vitro* by activators of calcium signalling mediated by protein kinase C. The importance of calcium as a secondary messenger in the central nervous system, including in neuronal-glial interactions, is consistent with the ongoing hypothesis that *CEROX1* plays a role in regulating this tissue.

The profoundly conserved nature of *CEROX1* was supportive of the hypothesis that the human transcript, like the murine *Cerox1*, may act as a ceRNA *in vitro*. As shown in Chapter 5, this was also supported by preliminary data *in vitro*, which showed that both mouse *Cerox1* and human *CEROX1* exhibited the ability to upregulate activity of complexes of the mitochondrial electron transport chain in both their own and one another's cellular backgrounds. Further repetition of these experiments, in the process of attempting to identify the putative miRNA mediators, revealed that the overexpression phenotype was poorly reproducible: on a number of occasions, *CEROX1* modulation resulted in no change, or even a decrease, in complex I activity. Two additional assays were devised to complement the results of the original spectrophotometric enzyme assays: blue native PAGE with in-gel activity assays; and Agilent Seahorse extracellular flux analysis technology. Respectively, these supported the more recent, and the original

preliminary data and conclusions. In light of the more physiological nature of the whole cells assessed by the Seahorse technology, compared to the processed lysates used in the other assays, and the conserved nature of the original phenotype, which is consistent with the evolution of *CEROX1*, I remain confident that *CEROX1* exhibits a role in upregulating activity of the ETC *in vitro*. The reason for the apparent loss of phenotype under application of the original assay remains unclear, though I suggest that subtle differences in growth conditions as demonstrated by Wright Muelas et al.[1], as well as in sub-clones of cell lines such as HEK293(T), may both contribute.

The multifaceted approach of this thesis does, however, remain highly supportive of the hypothesis that *CEROX1* is a conserved regulator of the electron transport chain that acts via an RNA-mediated mechanism, likely as a ceRNA, to support the substantial energetic requirements of the developing central nervous system. Figure 7.1 summarises a hypothetical model incorporating the findings of this thesis.

Finally, a key characteristic difference between human *CEROX1* and mouse *Cerox1* is that the former hosts several open reading frames of substantial length. Over the course of Chapter 4, I therefore sought to establish whether *CEROX1* is truly a lncRNA, or whether recent evolution has conferred upon it a bifunctional role as a protein precursor as well as an independently functional RNA. I demonstrated unequivocally that the 234 amino acid CEROX1-ORF1 protein is translated *in vitro* following *CEROX1* overexpression, but the relevance of this finding to physiological studies of *CEROX1* remain unclear: a targeted attempt to detect the CEROX1-ORF1 protein in human brain was unsuccessful. However, I present the detection of a CEROX1-ORF1 peptide in gorilla brain as evidence supporting the possible recent evolution of protein-coding capacity in *CEROX1*. I have thereby, as I hypothesised while introducing Chapter 4, shown that confirming the non-coding status of an individual lncRNA may be less trivial than typical bioinformatic approaches would suggest. The molecular bifunctionality of *CEROX1* that is plainly observed *in vitro* must be accounted for in future work undertaken in human cell models.

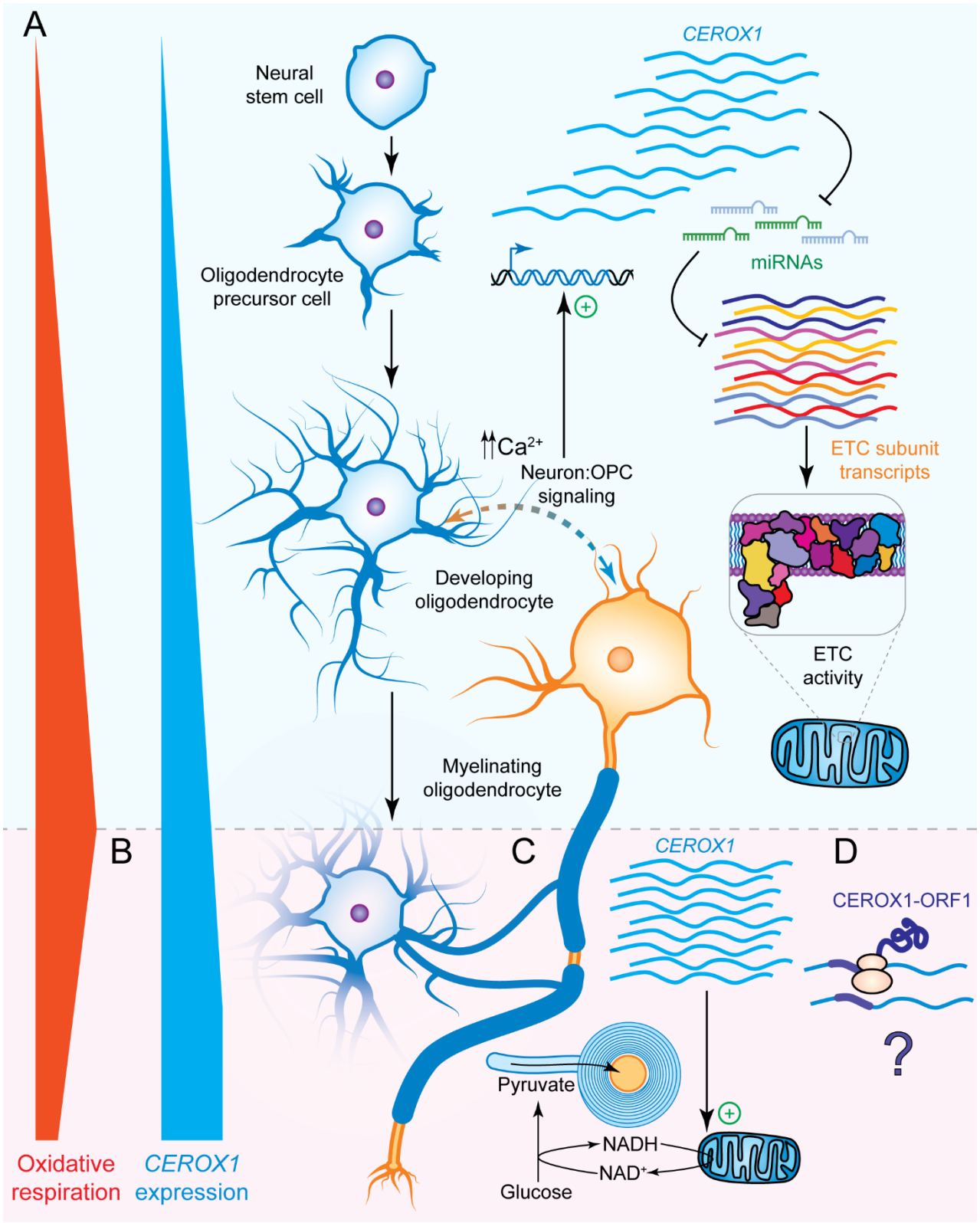


Figure 7.1: A *hypothetical* model of the physiological role of *CEROX1* in oligodendrocytes. (A) During development of oligodendrocytes, *CEROX1* expression rises in response to neuronal activity-stimulated calcium signalling[2,3]. *CEROX1* contributes to the upregulation of oxidative metabolism, supporting the energetic demands of proliferation and differentiation. **(B)** In myelinating oligodendrocytes, on the other hand, the role of the extremely high expression of *CEROX1* is less easily speculated, due to the lower dependence upon oxidative metabolism reported in this cell population[4,5]. **(C)** However, the importance of oligodendrocytes as sources of metabolites such as pyruvate transported from periaxonal myelin sheaths into neuronal axons suggests that *CEROX1*-stimulated mitochondrial metabolism might play another key role, such as promoting regeneration of NAD^+. **(D)** The possibility that *CEROX1* additionally acts as a protein precursor has not been entirely ruled out.

7.2 Moving forward with *CEROX1*

7.2.1 The ceRNA character of *CEROX1* remains unclarified experimentally

My work on *CEROX1* allows preliminary appraisal of this lncRNA as a ceRNA candidate in light of the trends and hypotheses garnered from the literature. Alongside its unusually high expression, its developmentally dynamic nature places it amongst the category of genes that I suggested, in the introduction to this thesis, might be the most appropriate ceRNA candidates. Since other approaches somewhat conflate expansion of cell populations and bona fide gene upregulation, a quantification time-course of mouse *Cerox1* copy number during oligodendrocyte development *in vivo*, for example, would provide further evidence that its expression is sufficiently dynamic to abrogate the stoichiometric arguments raised in the literature[6,7].

The difficulties faced during the phenotypic characterisation of *CEROX1* overexpression have strongly limited the conclusions of this body of work regarding ceRNA activity of *CEROX1*. While data raised in this thesis are consistent with *CEROX1* acting as a ceRNA, there is still substantial evidence required to categorise it as one *in vitro*, let alone *in vivo*. Having identified in Chapter 5 that extracellular flux analysis detects a mitochondrial phenotype following *CEROX1* overexpression, I hope that this phenotypic analysis may: a) prove to be more stable than the spectrophotometric complex I assay; and b) be used as a reproducible readout to elucidate the dependence of *CEROX1* activity upon miRNAs, generally through DICER/AGO inhibition as well as specifically via MRE ablation and assays of physical lncRNA:miRNA interaction.

In the event that further investigation does not confirm the putative ceRNA activity of human *CEROX1*, then RNA immunoprecipitation should be used to explore any direct interaction with the translation machinery, mitochondria-regulating transcription factors such as PGC-1α, or the proteins of the ETC. While its largely cytoplasmic localisation suggests a post-transcriptional mechanism of action, any interaction between *CEROX1* and, for example, promoters of nuclear-encoded mitochondrial genes, may be probed by capture hybridization analysis of RNA targets (CHART)[8].

7.2.2 *CEROX1* remains an excellent candidate for physiological lncRNA functionality

The work of this thesis strongly supports the continued investigation of *CEROX1*: despite the challenges in assessing its overexpression phenotype, it remains the first mouse and human lncRNA identified to regulate the electron transport chain *in vitro*. With a *Cerox1* knockout mouse on the horizon, the work of this thesis

will prime the assessment of the phenotype of *Cerox1* deletion *in vivo*: the hypothesis formulated above strongly points to oligodendrocyte development and function as a key starting point. As with all genes implicated in developmental processes, and especially owing to the potential temporal bifurcation of *Cerox1* functionality suggested in Figure 7.1, it will be key to examine the role of this lncRNA using inducible, cell type-specific deletion models. This will allow for its contribution to the many components of oligodendrocyte development – early precursor cell proliferation and differentiation, maturation, and myelination – to be disentangled, by depleting *Cerox1* expression at different developmental stages. If indeed the dynamic expression of *Cerox1* is key in modulating mitochondrial energy metabolism, I would expect that its deletion would: a) impair the increased mitochondrial activity that characterises oligodendrocyte differentiation; and b) impede specification, proliferation, and maturation of oligodendrocytes, leading to failure or reduction of myelination. These hypotheses could be addressed by respiratory analysis of primary cultures and by histology, respectively. The effect of *Cerox1* loss after the completion of the myelin network upon that system is less predictable, but I would hypothesise that deletion would impair oligodendrocyte function and impinge upon axonal health, possibly independently of oligodendrocyte viability similarly to monocarboxylate transporter deletions[9], or perhaps via ROS-mediated oligodendrocyte loss. Within steady-state terminally differentiated cells generally, it seems that ceRNA activity might be confined to a homeostatic role, and that a specific environmental challenge might be required to tease out a ceRNA-modulated phenotype. The role of *Cerox1* within terminally differentiated oligodendrocytes may be a useful model to investigate this theory: it would be interesting to determine the effect of *Cerox1* modulation in the cuprizone-induced demyelination-remyelination model[10]: does *Cerox1* protect against demyelination, or enhance recovery? Key within all of these *in vivo* deletion experiments will be the inclusion of RNA-mediated rescue, perhaps by lentiviral re-expression of *Cerox1*.

Oxidative status has been suggested as a link between lack of environmental stimulation and subsequent cognitive deficiency, with oxidative damage in the hippocampus and medial temporal lobe associated with spatial and emotional learning and memory in young mice[11]. Perhaps a lack of neuronal activity-stimulated induction of *CEROX1* expression leads to defective mitochondrial activity and subsequent oxidative stress. Oligodendrocyte lineage cells are especially vulnerable to oxidative damage *in vitro*[12-15] and *in vivo*[16].

The sharp correlation between the expression levels of *CEROX1* and *SOX8*, their concerted regulation at their shared promoter, at least *in vitro*, and the unclear importance of existing roles of *SOX8* in development of oligodendrocytes all suggest that any further work to understand *CEROX1 in vivo* must be mindful of disruption of *SOX8*. Indeed, a previous *Cerox1* knockout mouse model was deemed uninformative following ablation of both *Cerox1* and *Sox8* via their shared promoter (Dr Tamara Sirey, personal communication).

Looking ahead, I have shown that the expression of *CEROX1* is responsive to regulatory modulation by small molecules. Following validation of the hypothesised physiological relevance of *CEROX1* it is plausible that it might be implicated in diseases such as multiple sclerosis and ALS. Reporter constructs such as those I used in Chapter 6, alongside tool compounds such as PMA, will be a useful model in investigating the endogenous regulation of *CEROX1*, thereby providing a platform for the development of expression-modulating drugs.

7.2.3 CEROX1-ORF1: moving beyond reasonable doubt?

Advantageously, the absence of the CEROX1-ORF1 sequence in murine *Cerox1* excludes its involvement in any phenotype observed in mouse models. However, I assert that a more extensive search is required to address the continued possibility that CEROX1-ORF1 is translated in humans *in vivo*. Ideally, more tissues of interest would be subjected to a targeted search modelled upon that of Chapter 4, and all efforts taken to preserve integrity by minimising post-mortem sample delay, which may contribute to the apparent absence of CEROX1-ORF1 in the two human samples used in this study. Following its recent use in discovering several novel human proteins, it has been suggested that testis is a likely place of expression for uncharacterised proteins[17]. Since *CEROX1* is expressed in testis, most highly after the CNS, this would be a high priority tissue for a targeted proteomic study. It is unfortunate that CEROX1-ORF1 is confined to primates, making it inaccessible to *in vivo* TRAP, which has been used to investigate other ribosome-associated lncRNAs[18].

Is CEROX1-ORF1 unique? I would hypothesise that it is not: while most of the literature suggests that the majority of typical lncRNAs do not encode proteins *in vivo*[19,20], I think it unlikely that *CEROX1* is the only exceptional candidate, and the number of small to moderately sized ORFs that may be translated *in vitro* remains unaddressed. This has indirect implications for the investigation of lncRNA function since the majority of research at present is undertaken in cell culture models. A wide scale screen of epitope-tagged

lncRNA-ORF overexpression *in vitro* would indicate whether selective translation of exogenous endogenously-non-coding RNA is a widespread phenomenon.

7.3 Challenges of future ceRNA identification

My work strongly advocates the importance of taking complementary approaches, including bioinformatic and experimental methods, in investigating lncRNAs, garnering data to understand each in its own right. *CEROX1* is exceptional in many ways: extremely high levels of expression and evolutionary conservation, hosting a long ORF that is translatable *in vitro*, its GC content, and even the length of its CpG island all set it apart from the 'average' lncRNA. I hypothesise that many of the yet-unknown list of lncRNAs of physiological importance might be exceptional in their own manner within the lncRNA class, and that this will facilitate their functionality. Especially in the case of ceRNAs, which perhaps require a dynamic range of expression not observed for most lncRNAs; indeed, I propose that this may be a valuable criterion in identifying lncRNAs as priority ceRNA candidates. However, the resources to identify such cohorts are still lacking. My appraisal of the different catalogued annotations of *CEROX1* isoforms demonstrates that our knowledge of the genomic architecture of even highly expressed and conserved lncRNA transcripts is incomplete, though has been enhanced by the comprehensive nature of projects such as FANTOM5[21], which amend and improve annotation instead of fitting data to incomplete models. Likewise, while datasets such as GTEx[22] provide a unique source of averaged steady state human tissue expression data, resources are currently insufficient when it comes to identifying the highly dynamic and cell type-specific expression that I posit would characterise an efficacious ceRNA *in vivo*. The process of identifying genes with this requisite pattern will be facilitated by projects with ultra-high temporal resolution, such as the Owens et al. developing *Xenopus* transcriptome[23], though a comparable project addressing extended development of any mammalian model would be an onerous endeavour. Spatially-resolved transcriptomics projects such as the Human Cell Atlas, which seeks to map in detail the cellular milieu within complex tissues at cellular transcriptome resolution[24,25], may reveal further complexity of anatomical structure- and cell type-specific expression, especially in the brain, where cell types and gene expression are particularly diverse. Dynamic physiological human gene expression presents a further challenge due to the sporadic nature of tissue samples, but a combination of the growing technologies of hiPSC-derived cell types and organoids as well as single-cell sequencing may provide a model in which the expression programmes of differentiation may be modelled

with high spatiotemporal resolution. Perhaps advancements in pseudotemporal single cell trajectory analyses[26] might allow the inference of dynamic developmental expression profiles from temporally-limited samples such as human tissues, and extensive existing datasets such as the recently published *Tabula Muris* atlas, which includes over 100,000 cells in 20 mouse tissues[27].

High content imaging approaches using the mitochondrial membrane potential-sensitive dye TMRM have recently been shown to illustrate differences in mitochondrial activity in hiPSC-derived neurons from Parkinson's disease patients[28]. Intersecting spatial transcriptomics with single cell phenotypic analyses such as this would facilitiate a ground-breaking level of expression-function correlation analysis, which could be applied to acute brain sections as well as primary cultures. This approach may be useful in identifying specific cell sub-populations in which ceRNAs such as *CEROX1* exhibit regulatory activity, as well as new ceRNA candidates.

As noted in the introduction of this thesis, high throughput approaches have recently been applied to estimating lncRNA functionality. A CRISPRi screen of over 16,000 lncRNA loci in seven human cell lines identified ~500 "required for robust cellular growth"[29]. *CEROX1* featured in the CRISPRi-targeted lncRNA set, but the effect of its depletion in two cell lines was insufficient to pass an applied threshold; many other interesting lncRNAs were likely also omitted. While the study was a useful proof-of-principle for high-throughput CRISPRi, I strongly advocate that this approach must be extended to more physiologically relevant models such as primary and patient-derived cells to identify candidates of greater interest for future study, incorporating environmental challenges and stimuli to model disease and development states.

For those ceRNA candidates whose expression profiles are consistent with a hypothetical ability to have a meaningful effect upon miRNA levels, this investigation has illustrated a major problem in further investigation that is not unique to ceRNAs nor lncRNAs. Non-reproducibility has been highlighted as a profound problem in biomedical, and indeed all scientific, research[30]. A 2016 survey of 700 biologist *Nature* readers found that over 75% of responders had failed to reproduce an experiment of another researcher[31]. While selective reporting, misuse of statistical tests, and insufficiently detailed protocols details contribute to this statistic[32], these criteria do not so easily explain the 60% of respondents who had failed to reproduce their own experiments[31]. As discussed in Chapter 5, a complete assessment of the experimental conditions, including cell identity and growth, is required to truly assess the *CEROX1* overexpression phenotype, especially owing to the centrality (and hence interconnectedness) of the ETC within the cell.

This thesis has illustrated the process of investigating a novel ceRNA candidate, as well as many of the challenges therein. More excitingly, it has primed the further investigation of *Cerox1 in vivo*. The stage remains keenly set for the first proven example of physiological canonical ceRNA activity *in vivo*, and *CEROX1* is a promising candidate. Extensive work remains to identify other candidates and provide conclusive evidence for the decoy of miRNA activity as a physiologically relevant mechanism of action for lncRNAs.

REFERENCES

1. Wright Muelas, M., Ortega, F., Breitling, R., Bendtsen, C. & Westerhoff, H.V. Rational cell culture optimization enhances experimental reproducibility in cancer cells. *Scientific Reports* **8**, 3029 (2018).
2. Pitman, K.A. & Young, K.M. Activity-dependent calcium signalling in oligodendrocyte generation. *The International Journal of Biochemistry & Cell Biology* **77**, 30-34 (2016).
3. Krasnow, A.M., Ford, M.C., Valdivia, L.E., Wilson, S.W. & Attwell, D. Regulation of developing myelin sheath elongation by oligodendrocyte calcium transients in vivo. *Nature Neuroscience* **21**, 24-28 (2018).
4. Schoenfeld, R., Wong, A., Silva, J., Li, M. *et al.* Oligodendroglial differentiation induces mitochondrial genes and inhibition of mitochondrial function represses oligodendroglial differentiation. *Mitochondrion* **10**, 143-150 (2010).
5. Fünfschilling, U., Supplie, L.M., Mahad, D., Boretius, S. *et al.* Glycolytic oligodendrocytes maintain myelin and long-term axonal integrity. *Nature* **485**, 517 (2012).
6. Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P. & Stoffel, M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell* **54**, 766-76 (2014).
7. Denzler, R., McGeary, Sean E., Title, Alexandra C., Agarwal, V. *et al.* Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Molecular Cell* **64**, 565-579 (2016).
8. Simon, M.D. Capture Hybridization Analysis of RNA Targets (CHART). *Current Protocols in Molecular Biology* **101**, 21.25.1-21.25.16 (2013).
9. Lee, Y., Morrison, B.M., Li, Y., Lengacher, S. *et al.* Oligodendroglia metabolically support axons and contribute to neurodegeneration. *Nature* **487**, 443 (2012).
10. Torkildsen, Ø., Brunborg, L.A., Myhr, K.M. & Bø, L. The cuprizone model for demyelination. *Acta Neurologica Scandinavica* **117**, 72-76 (2008).
11. Cheng, L., Wang, S.-H., Jia, N., Xie, M. & Liao, X.-M. Environmental stimulation influence the cognition of developing mice by inducing changes in oxidative and apoptosis status. *Brain and Development* **36**, 51-56 (2014).
12. Roth, A.D. & Núñez, M.T. Oligodendrocytes: Functioning in a Delicate Balance Between High Metabolic Requirements and Oxidative Damage. in *Glial Cells in Health and Disease of the CNS* (ed. von Bernhardi, R.) 167-181 (Springer International Publishing, Cham, 2016).
13. Back, S.A., Han, B.H., Luo, N.L., Chricton, C.A. *et al.* Selective Vulnerability of Late Oligodendrocyte Progenitors to Hypoxia–Ischemia. **22**, 455-463 (2002).
14. French, H.M., Reid, M., Mamontov, P., Simmons, R.A. & Grinspan, J.B. Oxidative stress disrupts oligodendrocyte maturation. *Journal of Neuroscience Research* **87**, 3076-3087 (2009).
15. Thorburne, S.K. & Juurlink, B.H.J. Low Glutathione and High Iron Govern the Susceptibility of Oligodendroglial Precursors to Oxidative Stress. *Journal of Neurochemistry* **67**, 1014-1022 (1996).
16. Giacci, M.K., Bartlett, C.A., Smith, N.M., Iyer, K.S. *et al.* Oligodendroglia Are Particularly Vulnerable to Oxidative Damage after Neurotrauma In Vivo. **38**, 6491-6504 (2018).
17. Melaine, N., Com, E., Bellaud, P., Guillot, L. *et al.* Deciphering the Dark Proteome: Use of the Testis and Characterization of Two Dark Proteins. *Journal of Proteome Research* (2018).
18. Bazin, J., Baerenfaller, K., Gosai, S.J., Gregory, B.D. *et al.* Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. **114**, E10018-E10027 (2017).
19. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. & Lander, E.S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154**, 240-251 (2013).
20. Bánfai, B., Jia, H., Khatun, J., Wood, E. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Research* **22**, 1646-1657 (2012).
21. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* **16**, 22 (2015).
22. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580 (2013).
23. Owens, Nick D.L., Blitz, Ira L., Lane, Maura A., Patrushev, I. *et al.* Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Reports* **14**, 632-647 (2016).

24. Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A. & Teichmann, S.A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451-453 (2017).
25. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
26. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, 381 (2014).
27. Schaum, N., Karkanias, J., Neff, N.F., May, A.P. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
28. Little, D., Luft, C., Mosaku, O., Lorvellec, M. *et al.* A single cell high content assay detects mitochondrial dysfunction in iPSC-derived neurons with mutations in SNCA. *Scientific Reports* **8**, 9033 (2018).
29. Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**(2017).
30. Gannot, G., Cutting, M.A., Fischer, D.J. & Hsu, L.J. Reproducibility and transparency in biomedical sciences. *Oral Diseases* **23**, 813-816 (2016).
31. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452-4 (2016).
32. Reality check on reproducibility. *Nature* **533**, 437-437 (2016).

APPENDICES

SUPPLEMENTARY LITERATURE

A1 Supplementary literature describing *CEROX1*

Two manuscripts detail the parallel characterisation of mouse and human *CEROX1* transcripts, one a bioRχiv pre-print, the other a peer-reviewed publication in eLife; the latter is reproduced at the end of this work. This publication describes the work undertaken by the Ponting laboratory, led by Dr Tamara Sirey, into investigation of primarily mouse *Cerox1*, as well as preliminary data regarding human *CEROX1*, generated by me, Kenny Roberts. Included is detailed experimental evidence for the ceRNA activity of murine *Cerox1* in N2A cells *in vitro*, and its resulting regulation of the mitochondrial electron transport chain.

A1.1 The long non-coding RNA *Cerox1* is a post transcriptional regulator of mitochondrial complex I catalytic activity

Tamara M. Sirey, **Kenny Roberts**, Wilfried Haerty, Oscar Bedoya-Reina, Sebastian Rogatti-Granados, Jennifer Y. Tan, Nick Li, Lisa C. Heather, Roderick N. Carter, Sarah Cooper, Andrew J. Finch, Jimi Wills, Nicholas M. Morton, Ana Claudia Marques, Chris P. Ponting

Available through eLife, 2nd May 2019: doi.org/10.7554/eLife.45051[1]

Abstract: To generate energy efficiently, the cell is uniquely challenged to co-ordinate the abundance of electron transport chain protein subunits expressed from both nuclear and mitochondrial genomes. How an effective stoichiometry of this many constituent subunits is co-ordinated post-transcriptionally remains poorly understood. Here we show that *Cerox1*, an unusually abundant cytoplasmic long noncoding RNA (lncRNA), modulates the levels of mitochondrial complex I subunit transcripts in a manner that requires binding to microRNA-488-3p. Increased abundance of *Cerox1* cooperatively elevates complex I subunit protein abundance and enzymatic activity, decreases reactive oxygen species production, and protects against the complex I inhibitor rotenone. *Cerox1* function is conserved across placental mammals: human and mouse orthologues effectively modulate complex I enzymatic activity in mouse and human cells, respectively. *Cerox1* is the first lncRNA demonstrated, to our knowledge, to regulate mitochondrial oxidative phosphorylation and, with miR-488-3p, represent novel targets for the modulation of complex I activity.

A1.2 *Cerox1* and microRNA-488-3p noncoding RNAs jointly regulate mitochondrial complex I catalytic activity

Tamara M. Sirey, **Kenny Roberts**, Wilfried Haerty, Oscar Bedoya-Reina, Sebastian Rogatti-Granados, Jennifer Y. Tan, Nick Li, Lisa C. Heather, Sarah Cooper, Ana C. Marques, Chris P. Ponting

Preprint available at bioRχiv (CSHL), 16th May 2018: doi.org/10.1101/323907[2]

Abstract: To generate energy efficiently, the cell is uniquely challenged to co-ordinate the abundance of electron transport chain protein subunits expressed from both nuclear and mitochondrial genomes. How an effective stoichiometry of this many constituent subunits is co-ordinated post-transcriptionally remains poorly understood. Here we show that *Cerox1*, an unusually abundant cytoplasmic long noncoding RNA (lncRNA), modulates the levels of mitochondrial complex I subunit transcripts in a manner that requires binding to microRNA-488-3p. Increased abundance of *Cerox1* cooperatively elevates complex I subunit protein abundance and enzymatic activity, decreases reactive oxygen species production, and protects against the complex I inhibitor rotenone. *Cerox1* function is conserved across placental mammals: human and mouse orthologues effectively modulate complex I enzymatic activity in mouse and human cells, respectively. *Cerox1* is the first lncRNA demonstrated, to our knowledge, to regulate mitochondrial oxidative phosphorylation (OXPHOS) and, with miR-488-3p, represent novel targets for the modulation of complex I activity.

SUPPLEMENTARY MATERIALS AND METHODS

A2 Primer sequences

Listed here are primers used for: cloning of constructs, including the *CEROX1* overexpression construct and HA-tagged *CEROX1*; RT-qPCR; colony PCR; cloning of ISH probes; RACE; amplification of mouse and rat *Cerox1* sequences; and Sanger sequencing.

Cloning	
CEROX1 5' end cloning XhoI	TTCCTCGAGGGGCTCGGGGGCGGGGAGG
CEROX1 mid cloning Not I forward	CCAGCGGCCGCCCGCCCCCGTCC
CEROX1 mid cloning Not I reverse	CGGGCGGCCGCTGGGGGCACCAGCTCTCGG
CEROX1 5' deletion cloning XhoI	TTCCTCGAGTGAGGACCTCCGTGGAGAGC
CEROX1 3' end cloning BglII	AGATCTGACATTTTCACAGGGTGTGTTTTCATTTGC

***CEROX1* open reading frame epitope tagging**	
CEROX1 ORF1 tagging HA for	TATCCGTATGATGTGCCGGATTATGCGTGAGGACCTCCGTGGAGAG
CEROX1 ORF1 tagging HA rev	CGCATAATCCGGCACATCATACGGATACCCGGATGGCACTGTGAG
CEROX1 ORF2 tagging HA for	TATCCGTATGATGTGCCGGATTATGCGTGAGACTTGGAATCAACGCTTCCGTTG
CEROX1 ORF2 tagging HA rev	CGCATAATCCGGCACATCATACGGATAGACACCGCCCCCAGGCTG
CEROX1 ORF3 tagging HA for	TATCCGTATGATGTGCCGGATTATGCGTGATGCCCCCGACCTCCA
CEROX1 ORF3tagging HA rev	CGCATAATCCGGCACATCATACGGATAGGAGGACCAGGAGGACGTC

RT-qPCR		
Gene	**Forward**	**Reverse**
CEROX1	CAGGCACTTTCCCACACAG	CTACAGAGGCCCAGAAAGCA
Cerox1	TGGTAACTCCCGCACCTAGA	GCATGCAGGGCTACAGATG
CYC1	TTCAACCCCTACTTTCCTGGC	CTGGGGTGCCATCGTCAAA
GAPDH	CACCATCTTCCAGGAGCGAG	GACTCCACGACGTACTCAGC
LMF1	TGGGCCATGTCTGGTACTCT	ACAGGAAGATCCCCAGGAAC
MALAT1	GACCCTTCACCCCTCACC	TTATGGATCATGCCCACAAG
NDUFA1	CACTAACGGGGGCAAGGAAA	ATCAACTCCAGAGATGCGCC
NDUFA10	TTACATCAGTTCAGAGAGCTGCC	ACTTCAGCCAGATCCACTTGT
NDUFA11	CGTCGCTGGCCTGACC	TTGTCCAACCTTAGCCACTCC
NDUFA13	GGCGCCTACAAATCGAGGA	ATCTGCAAGGTCCTCCGGTC
NDUFA2	GAAGGCGAATCCCGACCTAC	CTCTTGGCCAAATGCGTAGC
NDUFA3	CGCTGTAATTCTGCCCCCAT	ACGGGCACTGGGTAGTTGTA
NDUFA4	AAGCATCCGAGCTTGATCCC	ACAATGCCAGACGCAAGAGA
NDUFA7	GATCTCCAAGCGAACTCAGC	CCATCGCGAGTGCAATAGTA
NDUFAF1	ACTGGCAGGAGGTCAAGATTC	AAGCGGAAGCTCATGCTGAA
NDUFAF2	AAGGGAAGTGAAGGAGCACG	AGTTTGTCCTCTCCAGTTCTTGT
NDUFAF3	CTAGAAGTCTGGCACCCACC	AGCTATTAAGCCGCGCTCC
NDUFAF4	TGCTCGTAAAGATGAAAAGCTGC	GCTTTTACCTGCAAGGAAGACA
NDUFB4	CACTCATGGGAGCTCTGTGT	AGTTTTTCTTTCCTATCCCTCTCA
NDUFB6	TCATGTACTTGTACCTGTCTGGA	TGGACTTCTTTTCAACTATGCCA
NDUFB9	GGACCCTACCTGACCCATCA	TTTGTCTCTCTGGACGCACC
NDUFC1	GGAGAAAGGGGCTTAACAGAGT	GGCCTCAGCCTTCTGTCGAT
NDUFC2	ACATCGGCTTCTTGGGCTAC	CTGGCGATGCAAACCAGC

RT-qPCR		
Gene	**Forward**	**Reverse**
NDUFS1	GAATAAGCGACCCGGCCTC	CCTTAACATATTGCTTCTCCCCG
NDUFS2	CCATCTGGCTGGTTTGGACA	ACTACACAATAGGCCTCGTACC
NDUFS3	TACACAGATGAGCTGACGCC	TCCAAACATGTCCCAGATCTCC
NDUFS6	AGGTCGTCAGAAAGAGGTGA	GCTATCACCCGAGTCTCCAC
NDUFS7	AACGGAGGAGGCTACTACCA	GGGCAGCCTGGGATGTAGATG
NDUFS8	CAGTGTAGCCTCCGCCTC	CATAGGCGTGGTCAGGCAG
NDUFV1	CGTGAGGTGACCCATCTGGC	TTTTCTTGGGTGCTGTCCGC
NDUFV2	CCAAAACCAGGGCCAAGGAGT	CAGGTCCCTTGGGTGGTTC
POLR2A	TTGTGCAGGACACACTCACA	CAGGAGGTTCATCACTTCACC
SOX8	ATGTGTTTCTTCTGAAAGCTG	CGCTATTGTATGCTCTATGCT
TUBA1A	CTTCGTCTCCGCCATCAG	TTGCCAATCTGGACACCA
UQCR10	TCGTGGGCGTCATGTTCTTC	TTCCACAGCTTCCCCTCGTT
UQCR11	AGCTGGTCAAGAACTGGGTC	GATCAGCCGCCAATCGGT
UQCRB	GGCCCAGCTTGTCCTCTATG	ATGCTGAAACGGCCTGCTTA
UQCRC1	CCGAGTACCTCAGCACACAT	GGCGAGGTCTAACAGTTGCT
UQCRC2	CCTGCGGGGTGATGTTGATA	GGTCAGCTACTTCCCAACGA
UQCRFS1	GCCTCAATGTCCCTGCTTCT	CTAAAACTTCAAGGCGGCGG
UQCRH	CTAAAACTTCAAGGCGGCGG	GTTTGTGGGCCACGCAATG
UQCRQ	CCGCACGTCTTCACTAAAGG	AACACTACAAACTGCGGCAC

Colony PCR		
Gene	**Forward**	**Reverse**
CEROX1	AAAGCAGTTGCATTAGGTCTG	TTAATACCTGGTGACCCTCAA
		TACAGAGGCCCAGAAAGCAC
Cerox1	GTTGTGTGTGTGGTCATGCTC	GGAGTAGGTTTCCACCCAGAG

ISH probe cloning		
Gene	**Forward**	**Reverse**
CEROX1	ACCTCCCAGAGAGCGTCACCT	GGGGCCTCATGAACTCTGTGG
Cerox1	TAAGCGGCTTTCTCTGGGTG	GTGTTTAATTTGCCCTGGACACT
SOX8	AGAACATCGACTTCAGCAACG	CAGGGTGGTGTACACCGG
Sox8	ACGACCAGTCACATGGCTCA	GAAACCACCAGTCACCTCAGA

RACE - generic primers		
Shi et al.[3]	RACE adapter Antisense	p-ACTCTGCGTTGATACCACTGCTT
	RACE adapter Sense	AAGCAGTGGTATCAACGCAGAGTGNNNNN
	UPMlong	CTAATACGACTCACTATAGGGCAAGCAGTGGTATCAACGCAGAGT
	UPMshort	CTAATACGACTCACTATAGGGC
GeneRacer™, Invitrogen	RACE Oligo dT	GCTGTCAACGATACGCTACGTAACGGCATGACAGTGTTTTTTTTTTTTTTTT
	GeneRacer™ 3’	GCTGTCAACGATACGCTACGTAACG
	GeneRacer™ 3’ nested	CGCTACGTAACGGCATGACAGTG
	GeneRacer™ RNA Oligo	CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA
	GeneRacer™ 5’	CGACTGGAGCACGAGGACACTGA
	GeneRacer™ 5’ nested	GGACACTGACATGGACTGAAGGAGTA

RACE - gene-specific	
Human *CEROX1* gene-specific 5' RACE primers	AGGCGCGAGACGGGCTGTGTG
	CACCGGCCTGCGGGCGAGAC
	CCGCAGGCAGCAGCTGACCC
	GCGGAACCGACCCTGTCCTG
	GTGCGCTGTCCGCAGACTCTGG
	GCTCTCCACGGAGGTCCTCACCCGG
	GGATGGGGAGCCAGGAAGAGGCC
	CCCTGCCTGCTCGCAGACCTAATGCAAC
	CAATGTCCCCTCTCCGACCCCGCGG
	GGCAATGTCCCCTCTCCGACCCCGC
	GCGCGCTGGTTGAAAGGATCCGAGC
	CGCGCTGGTTGAAAGGATCCGAGCCACC
	CAGGGCGGCACACCCCGCCCTCTC
	CCACCTCGCGGGGCGGCCTT
	CCTCCAGCGCGCTGGTTGAAA
	GCGTTGGCTGCCTGACCCGG
	ACCGCGTGGCTCAGTGGCTGC
	GCCAATGCTTCCTCCAAAGGCCC
	CTACGATGCGTGGAGCCTCCCAACT
	GCATGGGGATCCTCACCCACATACC
Human *CEROX1* gene-specific 3' RACE primers	GTGAACCCGCGCTGGCGGGAG
	CCGCTGGGGACCCCTGCCCAGG
	CCCCTGCCCAGGCTGTTCGCATGTGG
	GGATTCAGGGAGCAGGCGATGCCC
	GGGAGCAGGCGATGCCCAAAGC
	GGCCCCTTGCCCAGAATGTCCACAG
	CGCCCACTCGGGCTCACAGTGCCATCC
	GGTCTGCGAGCAGGCAGGGTGG
	CCTCTTCCTGGCTCCCCATCCACC
	GAGGTCTCGCCCGCAGGCCGGTG
	CGCGGGGTCAGCTGCTGCCTGC
	CCAGGACAGGGTCGGTTCCGCC
	CGCCCAGGCCCAGCTCGTCCAG
	GGGCTCTCTCAGCCCATCAGCAGGAG
	GCCGTGTACAGGAGGGTGCTGTCTGACG
	GGGAGCGGGTCCCCAAGCTTCTGGC
	GCCCTGCAGACGCTGGCCTTGTGGAG
	GCCCACAAACCCCCCTTCCAGAG
	CCACCTTGAAGTCCCCAGGAAAATGAC
	CGTCCTCCGCGGGTGTGAGCCAC
	GGACGCCGGGACGGAGCTTTCAG
	CCCAACACCGAGCCCCATTTCA
	CCGTCCTTCCAAGCTCAACCTTCAGAG
	GGCTGAGGTCCTCTGAGAAGGGACAC
	TTCAGCCCTTGGGCCCTTGTAGGAT
	GCCGTGTACAGGAGGGTGCTGTCTG
Mouse *Cerox1* gene-specific 3' RACE primers	GGGGACTCACTGTTGGGCCACCTG
	CGGCTGTTGATCCCCTCGGAAGTG
	GCATGGGCTACGGTGGACGGGG

RACE - gene-specific	
Macaque *Cerox1* gene-specific 3' RACE primers	GGCTGAGGTCCTCTGAGAAGGGACACG
	CAGCCCTTGGGCCCTTGTAGGATGC
	CGGGCGTCATCTCCTGCCGAG
	GGAAAGGATTCGGGGAGCGAGCG
Rat *Cerox1* gene-specific 3' RACE primers	CCTCCCAGGCTCGGCGTCTCAGA
	GGGGGACTCCACACCACCCTGATCC
	CCCTGGTGACATCTGTGACTGGGATGG

Amplification of rodent *Cerox1*		
Gene	**Forward**	**Reverse**
Mouse and rat *Cerox1*	GGAAAGGGGCTCTATTTAACCC	CTGTCGAATGCCAAGGAGGA
	GCTGTTCTCAGTCTGAAAGGATTC	CCAAGACCCACTGTATACTGTTGG
	CAGTGGAAGCTGCAGTTGGG	GAATCCTTTCAGACTGAGAACAGC
	CTCCAATGTGGGGCTCGAAT	GGGTTAAATAGAGCCCCTTTCC

Sequencing	
M13F	TGTAAAACGACGGCCAGT
M13R	AGGAAACAGCTATGACCAT
pCAGGS	GAGCCTCTGCTAACCATGTT
	GGGCTTCATGATGTCCCCAT
pNL2.3	GGCCTAACTGGCCGGTACCT
	GATCTTGATATCCTCGAGGC
	GTACCGGATTGCCAAGCTTGG
	GAAGGCGCTTGTGGAGAAGGAGT
	AGGAGCAGGCCCAGGGAGAA

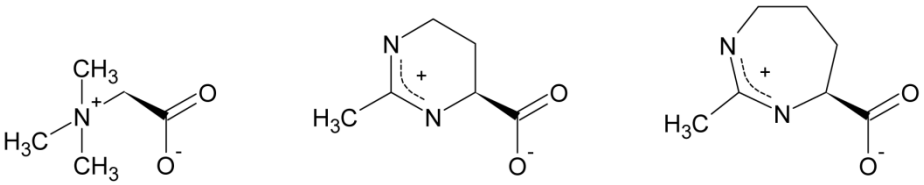
Supplemental Figure S1: Primers used in various experiments throughout this project, including those for cloning, colony PCR, RT-qPCR, RACE, generation of probes for ISH, and sequencing. All primers were synthesised by Sigma or Invitrogen, and subject to regular desalting purification prior to delivery. N – random nucleotides; p – 5' phosphorylated. N.B. human *CEROX1* gene-specific 3' RACE primers were also used in 3' RACE of gorilla *Cerox1*.

A3 Performing RACE upon human *CEROX1* was extremely challenging

Initially, to circumvent the GC-richness of *CEROX1*, RACE-PCR was repeated using Phusion High-Fidelity DNA Polymerase along with its GC-enhancer buffer and the addition of 5% DMSO. More products were obtained, but still none represented *CEROX1* amplicons. DMSO is a common PCR additive, used for amplification of difficult sequences. A number of other PCR adjuvants have been identified, and a number were trialled in attempting to amplify *CEROX1* products from the RLM-RACE-derived cDNA derived above. Betaine (N,N,N-trimethylglycine) is an osmolyte and cofactor found in some plants and invertebrates. It has been shown to promote PCR of difficult GC-rich sequences[4] by easing DNA melting, reducing the dependence of melting temperature upon the base composition of a sequence[5]. Trehalose is another osmolyte from non-vertebrates that has also been used in optimising difficult PCRs[6]. Based upon the findings of Zhang et al.[7] and Mousavian et al.[8], the organic solvent 1,2-propanediol was also included in a panel of PCR optimisation attempts, alongside new *CEROX1* 5' RACE primers and Phusion® High-Fidelity DNA Polymerase. Neither of these additives, nor any combination thereof, yielded a *CEROX1* amplicon.

The failure to amplify *CEROX1* 5' end amplicons led increasingly to the conclusion that the GC-richness of *CEROX1*, or a putative inherent difficulty in its secondary structure, was likely interfering with an earlier stage of the experiment, the reverse transcription reaction. Even though the SuperScript™ III Reverse Transcriptase protocol includes a five minute RNA denaturation step at 65°C, this might be insufficient to make the 5' end of *CEROX1* accessible. There would be no question of detecting CEROX1 RACE-PCR products downstream if first strand cDNA of *CEROX1* were not being produced.

While attempting to RACE a transcript containing stretches of extremely high GC content from the cabbage looper moth *Trichoplusia ni*, Shi et al.[3] developed an optimised RACE protocol using a number of specialised modifications designed to ease melting of difficult GC-rich or complexly structured DNA sequences. One of these important modifications, among only a few within the literature that target the reverse transcription cDNA synthesis step, was the addition of the molecule homoectoine. Homoectoine (4,5,6,7-tetrahydro-2-methyl-1H-[1,3]-diazepine-4-carbonic acid) is a synthetic derivative[9] of a bacterial 'compatible solute' called ectoine (1,4,5,6-tetrahydro-2-methyl-4-pyrimidinecarboxylic acid), which is found as an endogenous osmolyte in certain halophilic bacteria. Ectoine and homoectoine bear some chemical resemblance to betaine (Supplemental Figure S2), another well-known PCR-promoting osmoprotectant that was utilised above. Along with ectoine and betaine, homoectoine has the ability to facilitate melting of GC-rich DNA, and to a greater degree while acting at lower concentrations and stimulating the formation of fewer additional products[9]. Shi et al. combined use of the stably engineered reverse transcriptase ThermoScript™, which is active up to 70°C, with homoectoine to great success in reverse transcribing a sequence with stretches of up to 89% GC content[3].



Betaine	Ectoine	Homoectoine
N,N,N-trimethylglycine	1,4,5,6-tetrahydro-2-methyl-4-pyrimidinecarboxylic acid	4,5,6,7-tetrahydro-2-methyl-1H-[1,3]-diazepine-4-carbonic acid

Supplemental Figure S2: Osmolytes used in *CEROX1* 5' RACE. Betaine, ectoine, and the synthetic homoectoine share some common chemical features: they all exist in aqueous solution as zwitterions, with carboxyl anions and quaternary ammonium cations.

Beginning afresh from HEK293T total RNA, the protocol of Shi et al.[3] was utilised. First, mRNA was enriched from total RNA by polyA$^+$ selection. Initially this was attempted using a modified biotin-oligo-dT capture protocol inspired by that of the PolyATtract® mRNA Isolation System (Promega). Lack of reproducible purification led to the use of the commercial GenElute™ mRNA Miniprep Kit (Sigma). PolyA$^+$ RNA was isolated from HEK293T total RNA and this was used for cDNA synthesis. Briefly, RNA was reverse transcribed using ThermoScript™ with a gene-specific primer, in the presence of homoectoine. In contrast to the RNA-ligase mediated approach taken with the GeneRacer™ Kit, this protocol takes the first strand cDNA from the reverse transcription reaction, and subjects it to ligation to a double-stranded universal DNA adapter afterwards. An overhanging guanosine residue was used to increase specificity for full-length RNAs, targeting the reverse complement of the previously 5' methyl-guanosine-capped RNA.

Following cDNA synthesis and 5' end adapter ligation, 15 *CEROX1*-complementary primers spanning over 1 kb at the 5' end of the *CEROX1* locus were used in combination with the universal primer UPM long (Clontech, TaKaRa) for PCR with Phusion® High-Fidelity DNA Polymerase, DMSO, and betaine. Two different touchdown amplification patterns were trialled, one inspired by the GeneRacer™ Kit suggested protocol for Platinum® Taq, with temperatures modified for Phusion®, and the other based upon that published by Shi et al.[3]. Only three primer combinations derived any products with the GeneRacer™ Kit cycle protocol. Semi- and double-semi-nested reactions, using the UPM long primer with other gene-specific primers, produced a wider range of products but sequencing revealed that these were still not *CEROX1*-derived.

The level of expression of *CEROX1* in the HEK293T cells used above is relatively low, on par with the more lowly-expressing human tissues and around one-fiftieth of the expression level of whole brain (RT-qPCR). Brain tissue was therefore chosen for a final attempt at 5' RACE. cDNA was produced from total human brain RNA, using the ThermoScript™ and homoectoine protocol above. In order to account for possible problems with priming the high-temperature reverse transcription reaction, a cocktail of three *CEROX1*-specific primers was used. RACE-PCR with gene-specific primers and the UPM long primer followed by two semi-nested reactions produced products that sequencing revealed to be the furthest upstream 5' end (Chapter 3, Figure 3.6A).

A4 Species and loci used for analysis of *CEROX1* orthologous sequences

A4.1 Identification of transcribed *CEROX1* orthologues

The following species were used to generate the phylogenetic tree of *CEROX1* evolution:

Species		PLAR analysis?	Genomic locus between *Sox8* and *Lmf1* (strand)	Transcriptomics datasets	*Cerox1* locus (strand)
Human	*Homo sapiens*	Yes	hg38:chr16:971,056-981,743(-)		hg38:chr16:975,756-981,600(-)
Gorilla	*Gorilla gorilla*	-	gorGor5:chr16:1,003,627-1,014,618(-)	SRX217685 SRX217686	gorGor5:chr16:1,008,111-1,014,386(-)
Chimpanzee	*Pan troglodytes*	-	panTro4:chr16:976,518-986,854(-)	SRX1038914 SRX1038915	panTro4:chr16:981,525-985,687
Baboon	*Papio anubis*	-	papAnu2:chr20:830,251-841,394(-)	SRX843132 SRX843133 SRX843134 SRX843135	papAnu2:chr20:830,263-840,330(-)
Rhesus macaque	*Macaca mulatta*	Yes	rheMac8:chr20:925,526-956,960(-)		rheMac8:chr20:931,447-936,498(-)
Marmoset	*Callithrix jacchus*	Yes	calJac3:chr12:994,124-1,007,001(-)		calJac3:chr12:998,363-1,005,960(-)
Guinea pig	*Cavia porcellus*	-	cavPor3:scaffold_4:20,928,629-20,937,393(-)	ERX138087 ERX138111 ERX138133 ERX240886 ERX240876 ERX240893 SRX211681 SRX211682 SRX211683	cavPor3:scaffold_4:20,932,406-20,937,134
Rat	*Rattus norvegicus*	-	rn6:chr10:14,937,109-14,945,120(+)	SRX211710 SRX211711 SRX211712	rn6:chr10:14,937,232-14,941,536(+)
Mouse	*Mus musculus*	Yes	mm10:chr17:25,570,743-25,578,940(+)		mm10:chr17:25,707,756-25,711,988(+)
Cow	*Bos taurus*	-	bosTau8:chr25:770,133-787,327(-)	SRX211674 SRX211675	bosTau8:chr25:776,709-786,044(-)
Sheep	*Ovis aries*	Yes	oviAri4:chr24:708,811-725,749(-)		oviAri3:chr24:725,924-735,580(-)
Goat	*Capra hircus*	-	hub.capHir1:chr25:745,500-765,277(-)	SRX1936141 SRX1936142 SRX1936143 SRX211679 SRX211680	hub.capHir1:chr25:752,444-761,627(-)
Horse	*Equus caballus ferus*	-	equCab2:chr13:41,704,683-41,728,647(+)	SRX3744617 SRX3744618 SRX3744619 SRX3744620 SRX3744621 SRX3744622 SRX3744623 SRX3744624 SRX3744625 SRX3744626 SRX3744627 SRX3744628 SRX3744629 SRX3744630 SRX3744631	equCab2:chr13:41,706,438-41,713,855(+)
Cat	*Felis catus*	-	felCat8:chrE3:40,370,438-40,382,188(+)	SRX211688 SRX211689 SRX211690	felCat8:chrE3:40,373,308-40,381,258(+)
Dog	*Canis lupus familiaris*	Yes	canFam3:chr6:39,672,654-39,681,813(+)		canFam3:chr6:39,673,347-39,675,456(+)
Ferret	*Mustela putorius furo*	Yes	musFur1:GL896925:14,805,919-14,815,672(+)		musFur1:GL896925:14,807,197-14,811,494(+)

Species		PLAR analysis?	Genomic locus between *Sox8* and *Lmf1* (strand)	Transcriptomics datasets	*Cerox1* locus (strand)
Giant panda	*Ailuropoda melanoleuca*	-	ailMel1:GL192733.1:760,723-782,721(-)	SRX1208734 SRX1208761 SRX1208729	ailMel1:GL192733.1:777,130-782,447 (-)
Nine-banded armadillo	*Dasypus novemcinctus*	-	dasNov3:JH562917:40,227-58,460(+)	SRX146634	dasNov3:JH562917:41,493-44,433(+)

Supplemental Figure S3: Species whose transcriptomes were analysed in search of *CEROX1* orthologues. Also indicated is whether the species have been analysed by PLAR[10,11], and, if not, then which datasets were used to annotate a transcript manually using SRA-BLAST. Genomic positions describe the syntenic *SOX8-LMF1* intergenic regions (strand denotes that opposite to *SOX8* orthologues, which would be expected to host *CEROX1* orthologues), and putative *CEROX1*-orthologous transcripts identified.

A4.2 *CEROX1* orthologue tissue expression profiling

Species		Transcriptomics datasets and sources	
Human	*Homo sapiens*	GTEx project (gtexportal.org)[12]	
Rhesus macaque	*Macaca mulatta*	Kunming Institute of Zoology Chinese Academy of Sciences	
		Callosum	SRX2577753
		Globus pallidus	SRX2577757
		Putamen	SRX2577755
		Caudate nucleus	SRX2577756
		Substantia nigra	SRX2577759
		Hypothalamus	SRX2577767
		Mesencephalon	SRX2577770
		Medulla oblongata	SRX2577760
		Olfactory bulb	SRX2577758
		Hippocampus	SRX2577752
		Frontal lobe	SRX2577764
		Inferior colliculus	SRX2577762
		Temporal lobe	SRX2577765
		Parietal lobe	SRX2577763
		Superior colliculus	SRX2577761
		Hypophysis / pituitary	SRX2577751
		Conarium	SRX2577754
		Thalamus	SRX2577768
		Occipital lobe	SRX2577766
		Cerebrum	SRX2577771
		Retina	SRX2577772
		Cerebellum	SRX2577769
		Epididymis	SRX2577773
		Lung	SRX2577777
		Spleen	SRX2577776
		Liver	SRX2577780
		Testis	SRX2577774
		Kidney	SRX2577775
		Muscle	SRX2577779
		Heart	SRX2577778
Mouse	*Mus musculus*	FANTOM5 Project, Riken Institute, Japan[13]	
		Spinal cord	DRX009083
		Medulla oblongata	DRX009046
		Diencephalon	DRX008949
		Corpora quadrigemina	DRX008944
		Corpus striatum	DRX008945
		Cortex	DRX008947
		Hippocampus	DRX008986
		Olfactory brain	DRX009049
		Cerebellum	DRX008941
		Pituitary gland	DRX009066
		Urinary bladder	DRX009127
		Eyeball	DRX008954
		Adrenal gland	DRX008927
		Ovary	DRX009050
		Lung	DRX009030
		Tongue	DRX009126
		Uterus	DRX009129
		Testis	DRX009101
		Aorta	DRX008935
		Skin	DRX009077
		Spleen	DRX009084
		Epididymis	DRX008953
		Prostate	DRX009076
		Colon	DRX008943
		Accessory axillary lymph node	DRX008926
		Bone (os femoris)	DRX008936
		Submandibular gland	DRX009100
		Intestinal mucosa	DRX008988
		Cecum	DRX008940
		Pancreas	DRX009054

Species		Transcriptomics datasets and sources	
		Stomach	DRX009090
		Thymus	DRX009112
		Vagina	DRX009130
		Vesicular gland	DRX009131
Rat	*Rattus norvegicus*	Department of Biology, MIT, MA, US[14]	
		Brain	SRX196290; SRX196299; SRX196308
		Colon	SRX196291; SRX196300; SRX196309
		Heart	SRX196292; SRX196301; SRX196310
		Kidney	SRX196293; SRX196302; SRX196311
		Liver	SRX196294; SRX196303; SRX196312
		Lung	SRX196295; SRX196304; SRX196313
		Skeletal muscle	SRX196296; SRX196305; SRX196314
		Spleen	SRX196297; SRX196306; SRX196315
		Testes	SRX196298; SRX196307; SRX196316
Cow	*Bos taurus*	University of Missouri, MO, US / United States Department of Agriculture	
		Thalamus	SRX2658621
		Medulla oblongata	SRX2658629
		Hypothalamus	SRX2658634
		Temporal cortex	SRX2658623
		Cerebellum	SRX2658637
		Frontal cortex	SRX2658635
		Subcutaneous fat	SRX2658624
		Atrium	SRX2658639
		Anterior pituitary	SRX2658641
		Ventricle	SRX2658619
		Aorta	SRX2658640
		Nasal mucosa	SRX2658628
		Testis	SRX2658622
		Kidney, pelvic, heart fat	SRX2658633
		Bone marrow	SRX2658638
		Mammary gland	SRX2658630
		Uterine myometrium	SRX2658620
		Duodenum	SRX2658636
		Lymph nodes	SRX2658631
		Omasum	SRX2658627
		Lung	SRX2658632
		Rumen	SRX2658625
		Abomasum	SRX2658642
		Reticulum	SRX2658626
Cat	*Felis catus*	99 Lives Cat Genome Sequencing Initiative (felinegenetics.missouri.edu)	
		Ear cartilage	SRX1610308
		Bone marrow	SRX1610312
		Cerebellum	SRX1610309; SRX1625949
		Ear	SRX1610307
		Heart	SRX1610324
		Hippocampus	SRX1610305
		Kidney	SRX1610313; SRX1610326
		Liver	SRX1610306
		Lung	SRX1610302
		Muscle	SRX1610304
		Occipital lobe	SRX1610316
		Pancreas	SRX1610322
		Parietal lobe	SRX1610325
		Retina	SRX1610310
		Salivary gland	SRX1610318
		Skin	SRX1610323; SRX1625943; SRX1610320
		Spinal cord	SRX1610319
		Spleen	SRX1625944; SRX1625946
		Temporal lobe	SRX1610314
		Testes	SRX1610315
		Thymus	SRX1625945
		Uterus	SRX1610311
Dog	*Canis lupus familiaris*	Broad Institute, Cambridge, MA, US	
		Kidney	SRX111061
		Liver	SRX111062
		Brain	SRX111063

Species		Transcriptomics datasets and sources	
		Skin	SRX111064
		Ovary	SRX111066
		Heart	SRX111067
		Smooth muscle	SRX111068
		Testis	SRX111069
		Blood	SRX111070
		Lung	SRX111071
Giant panda	*Ailuropoda melanoleuca*	Institute of Zoology, Chinese Academy of Sciences[15]	
		Stomach	SRX1208633
		Skin	SRX1208682; SRX1208685
		Small intestine	SRX1208706
		Colon	SRX1208728
		Pallium	SRX1208729
		Testis	SRX1208734
		Pituitary	SRX1208761
		Tongue	SRX1208772
		Ovary	SRX1208780
		Liver	SRX1208869
		Skeletal muscle	SRX1208767
Nine-banded armadillo	*Dasypus novemcinctus*	Broad Institute, Cambridge, MA, US	
		Cerebellum	SRX146635
		Spleen	SRX146629
		Lung	SRX146634
		Kidney	SRX146633
		Quadricep	SRX146631
		Heart	SRX146630
		Liver	SRX146628
		Colon	SRX146632

Supplemental Figure S4: RNA-seq data sets used in tissue expression profiling of mammalian *CEROX1*-orthologous transcripts. All datasets were obtained via the NCBI Sequence Read Archive (ncbi.nlm.nih.gov/sra).

A5 Drugs used in screening activity of the CEROX1-SOX8 promoter

Stimulus or condition	Concentration
Insulin	5 µg/ml
Glucagon	500 ng/ml
Glucagon-like peptide (GLP1)	50 nM
C-peptide	50 nM
Glutamatic acid	5 µM
Brain-derived neurotrophic factor (BDNF)	50 ng/ml
Calcium chloride	10 mM
Disulfiram	20 µM
Basic fibroblast growth factor (FGF2)	50 ng/ml
Fibroblast growth factor 19 (FGF19)	100 ng/ml
Tumour necrosis factor alpha (TNF-α)	10 ng/ml
β-Estradiol	50 pM
Progesterone	20 nM
Citalopram hydrobromide	100 nM
Acetylsalicylic acid	1 mM
Phorbol 12-myristolate 13-acetate (PMA/TPA)	20 nM
Retinoic acid	50 nM
Phenobarbital	50 µM
DL-AP5	1 µM
Sodium ascorbate	100 µM
Allopurinol	10 µM
Ciliary neurotrophic factor (CNTF)	1 ng/ml
Heparin sodium	800 µM
SB431542	10 µM
ROCK inhibitor Y-27632	50 µM
Cytosine-1-β-D-arabinofuranoside (AraC)	500 nM
Apo-transferrin	15 µg/ml
Neurotrophin-3 (NT3)	50 ng/ml
T0070907	1 nM
Neoruscogenin	150 nM
Compound C	250 nM
N-ethyllidocaine (QX-314) chloride	2 µM
GBR 12909 dihydrochloride	50 µM
5-aminoimidazole-4-carboxamide ribonucleotide (AICAR)	1 µM
Forskolin	2 µM
PD0325901	1 nM
Idazoxan hydrochloride	1 µM
Homovanillic acid	10 µM
Apomorphine hydrochloride	50 µM
Diprenorphine	5 nM
Isoproterenol hydrochloride	2 µM
Naloxone hydrochloride	5 nM
Resveratrol	10 µM
Dihydro-β-erythroidine hydrobromide	0.5 µM
Lidocaine	50 µM
Mecamylamine hydrochloride	10 µM
Benserazide hydrochloride	50 µM
Anisomycin	100 nM
5-tetradecyloxy-2-furoic acid (TOFA)	1 µM
Etomoxir	10 nM
PD-184161	100 nM
KN-62	1 µM
SR1001	1 µM
SR1078	5 µM
Hydrocortisone (cortisol)	700 nM
Dopamine hydrochloride	200 nM
Carbamylcholine chloride (carbachol)	10 µM
Glial cell line-derived neurotrophic factor (GDNF)	5 ng/ml
Chemokine (C-C motif) ligand 21 (CCL21)	15 ng/ml
Interferon β-1b (IFNB-1b)	100 pg/ml

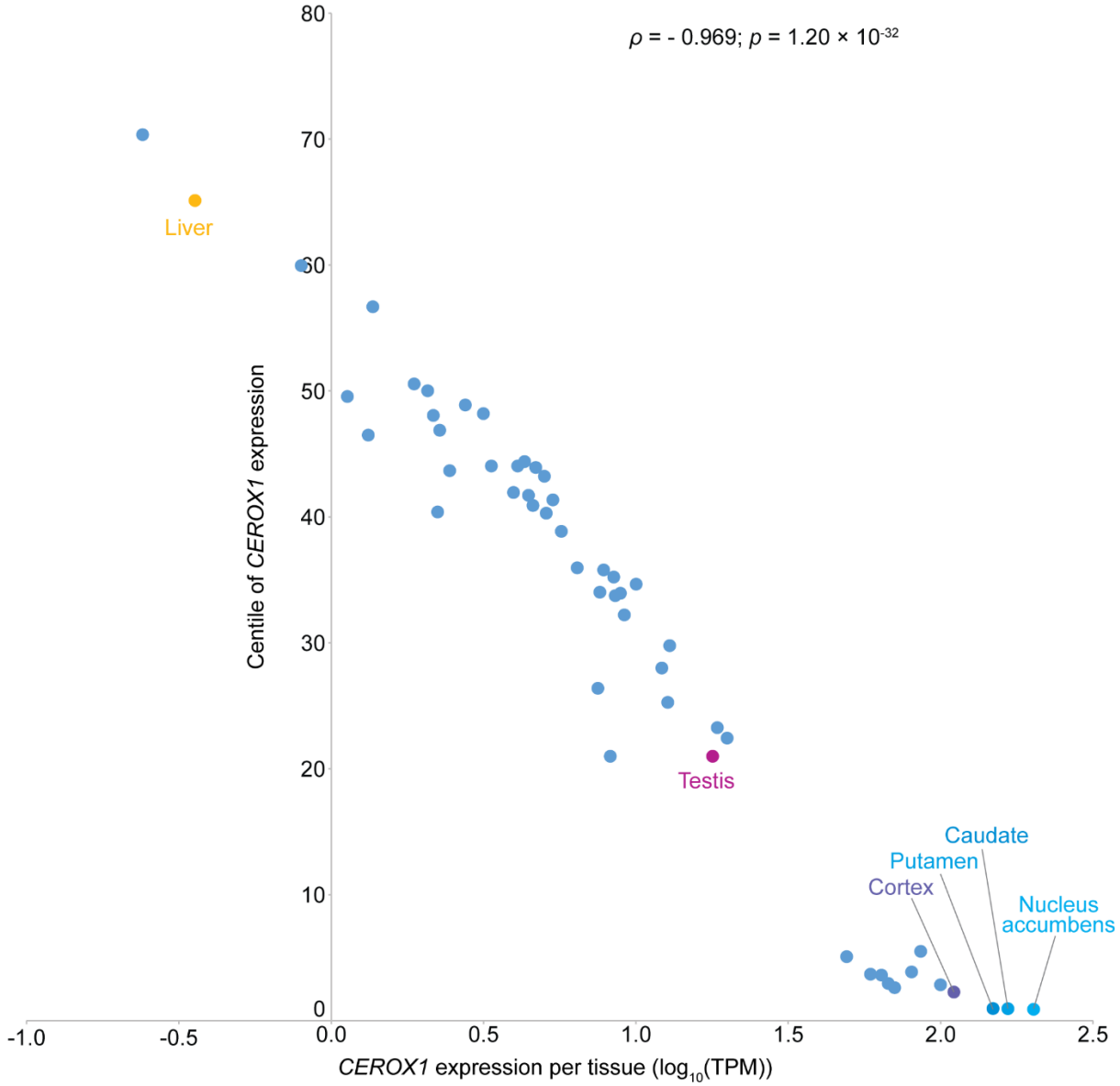
Supplemental Figure S5: Drugs and hormones used in the SH-SY5Y *CEROX1-SOX8* Nluc promoter screen. Shading colours denote non-aqueous solvents: blue – DMSO; orange – ethanol (both limited to 0.1% final concentration).

SUPPLEMENTARY DATA

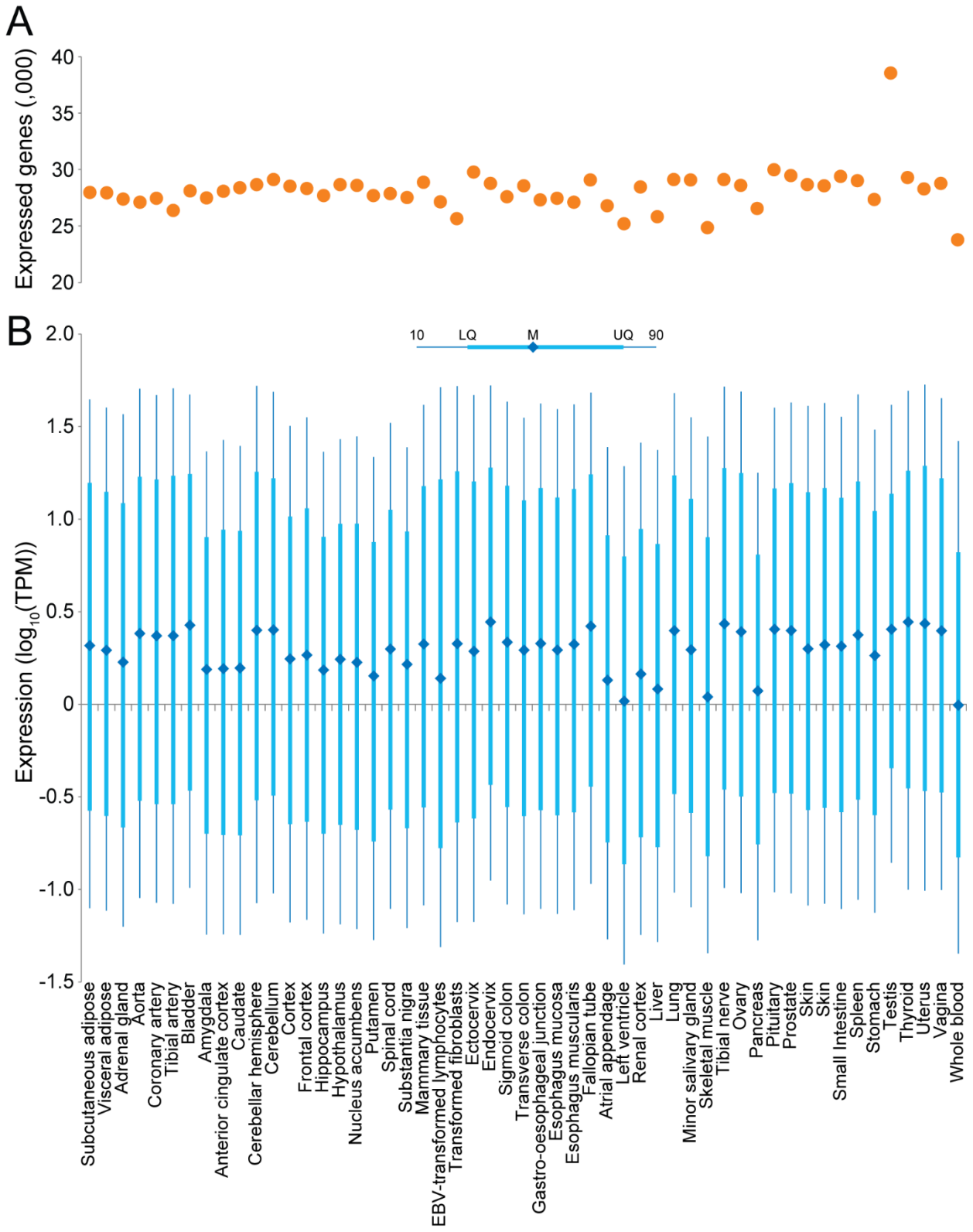
A6 Tissue-specific transcriptome analyses of *CEROX1* expression

A6.1 *CEROX1* transcriptomic rank correlates with expression level

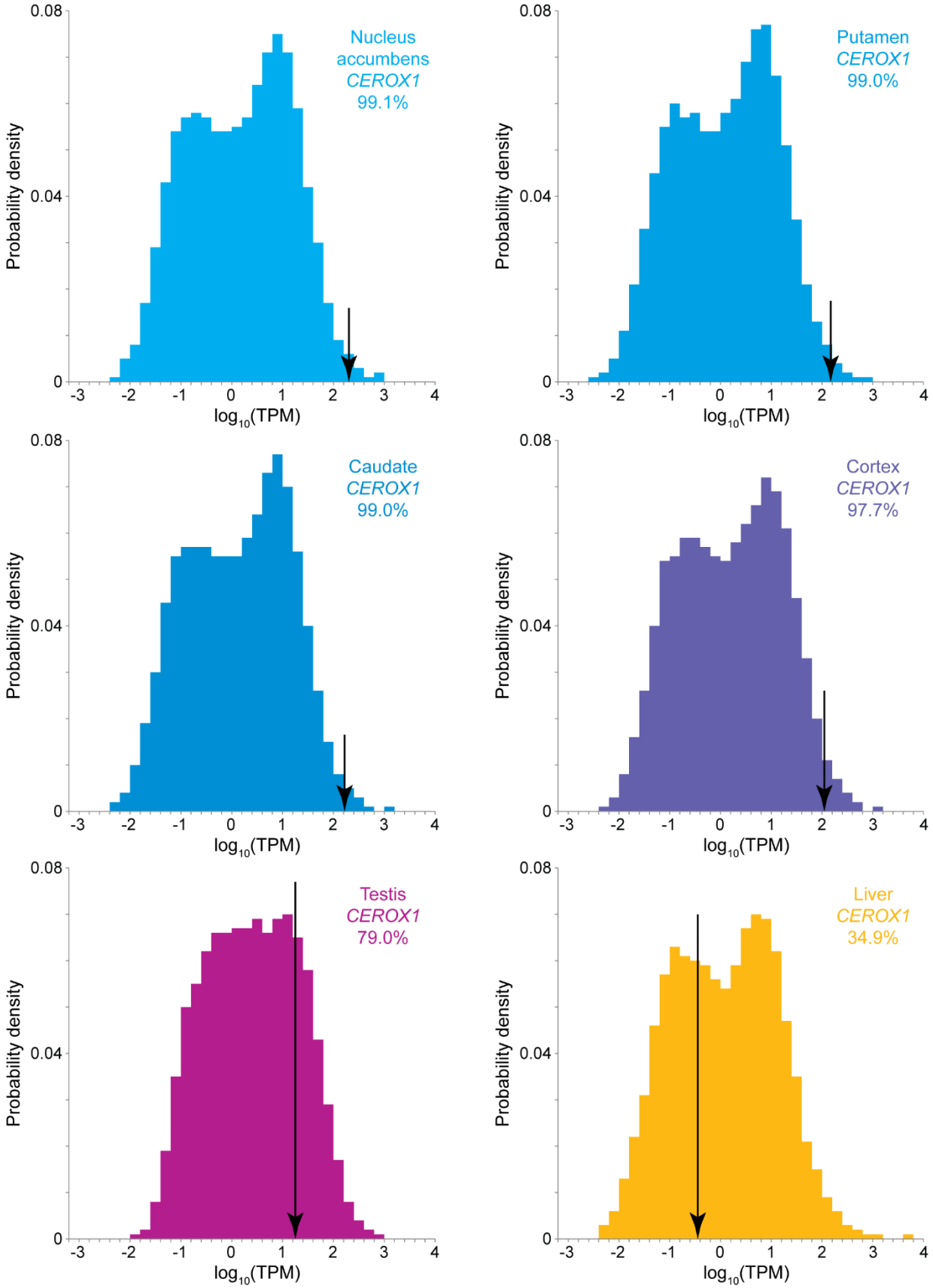
Across the 53 human tissues represented by the GTEx project[12], there is a strong correlation between the expression level of *CEROX1* and its relative rank within the transcriptome of a given tissue (Spearman's ρ = -0.97, $p = 1.2 \times 10^{-32}$) (Supplemental Figure S6). This is to be expected since a similar number of genes are expressed in each tissue, and the median expression level is relatively constant (Supplemental Figure S7), and not correlated with *CEROX1* expression (Spearman's ρ = 0.07, p = 0.61). Individual examples of tissues from high, medium, and low expression are shown in Supplemental Figure S8.



Supplemental Figure S6: *CEROX1* expression within a given tissue is strongly correlated with expression rank in that tissue. Spearman's ρ = -0.97, $p = 1.2 \times 10^{-32}$. Transcriptome analyses of liver, testis, cortex, putamen, caudate, and nucleus accumbens are shown in Supplemental Figure S7. All data were obtained from the GTEx project[12].



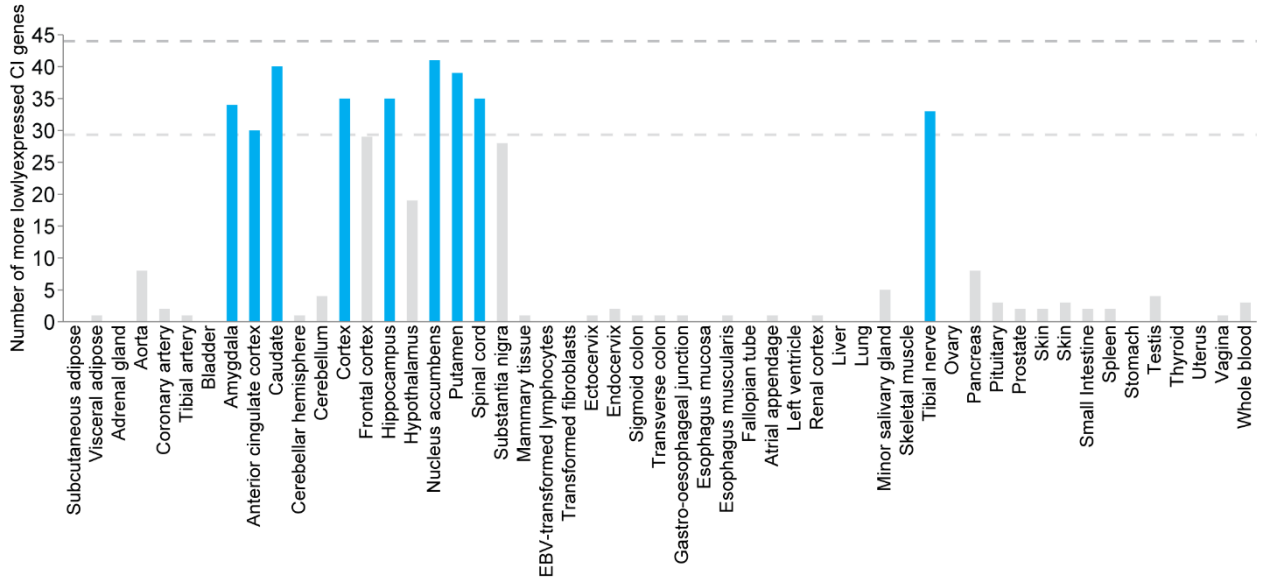
Supplemental Figure S7: Expression distributions for all 53 GTEx tissues. (A) Total number of genes expressed per tissue. **(B)** As shown by the key at the top, dark blue markers represent the median expression, blue bars the interquartile range, and dark blue lines the 2nd to 9th decile range. All data were obtained from the GTEx project[12].



Supplemental Figure S8: *CEROX1* expression within a given tissue is strongly correlated with expression rank in that tissue. In the most highly expressing tissues such as the nucleus accumbens, *CEROX1* is within the top <1% of expressed genes. Rank decreases sharply with overall expression, placing *CEROX1* more lowly expressed than 65% of genes in the liver.

A6.2 Expression rank relative to mitochondrial genes is also tissue-specific

In the tissues in which *CEROX1* is most highly expressed, such as the basal ganglia, it is more highly expressed than ~40 of 44 complex I-associated genes. In most other tissues, *CEROX1* is more highly expressed than only a couple of the more lowly expressed assembly factors. Supplemental Figure S9 shows the number of complex I-associated genes whose expression is outranked in each tissue. In 9 of 52 tissues, *CEROX1* is more highly expressed than two-thirds of complex I-associated genes; all 9 are in the CNS.



Supplemental Figure S9: *CEROX1* is more highly expressed than a majority of complex I-associated genes in the CNS. *CEROX1* is more highly expressed than two-thirds of 44 nuclear-encoded complex I-associated genes (37 subunits and 7 assembly factors) in only 9 of 53 tissues (blue bars). All 9 are within the CNS.

A7 Further investigation of *Cerox1* in the striatum

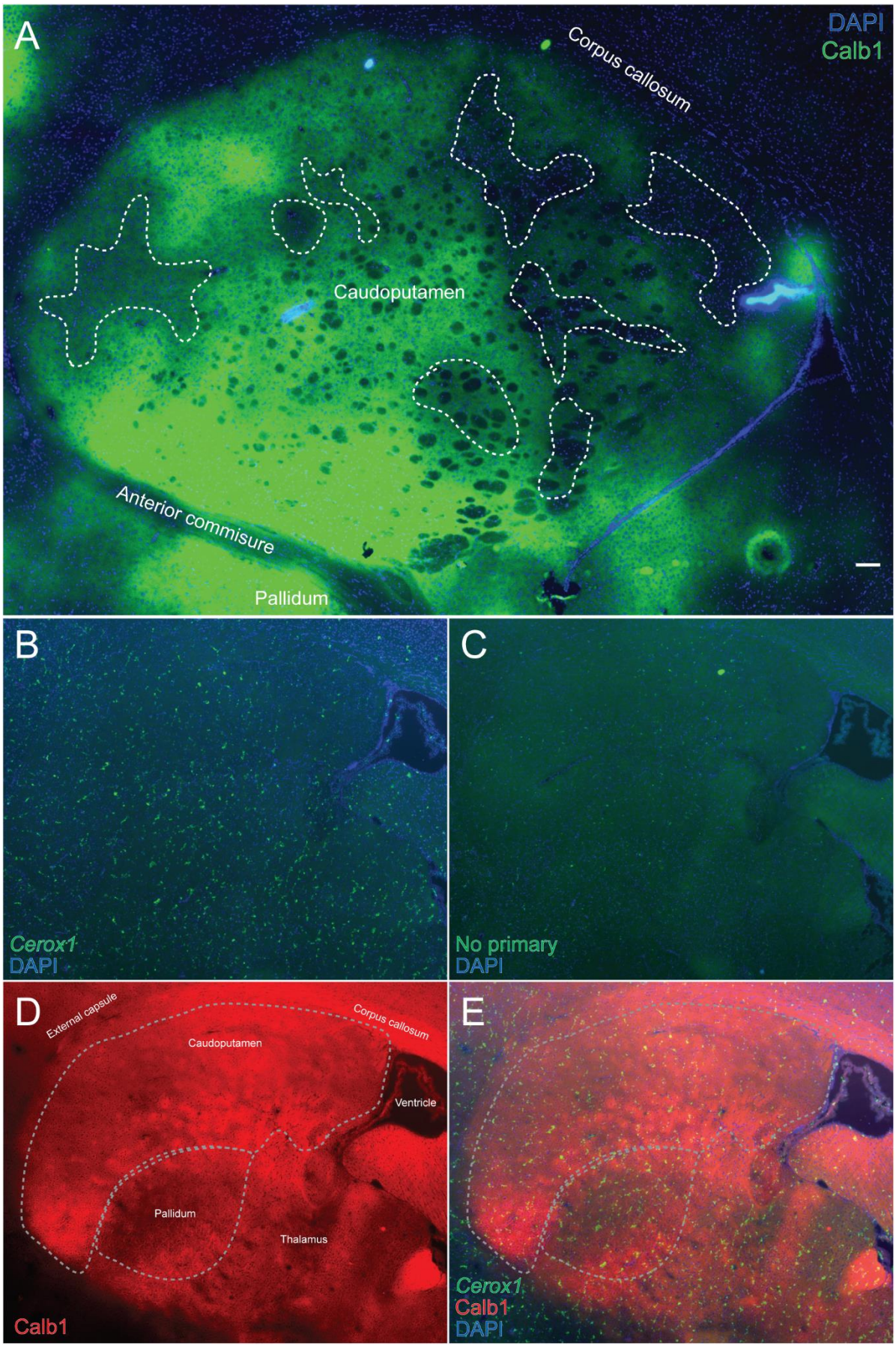
Following the putative association of human *CEROX1* with striosomes, IHC and FISH were applied in parallel to further investigate this in mouse striatum.

IHC for the striosome-depleted marker calbindin (Calb1) was carried out as described for general IHC in Chapter 2, and initially identified striosomes in a coronal section of adult mouse brain (Supplemental Figure S10A).

For fluorescent detection of in situ hybridisation, slides were washed in 100 mM Tris-HCl pH 7.6, 150 mM sodium chloride three times for five minutes each, and then equilibrated in 20 mM Tris-HCl pH 7.6, 150 mM sodium chloride (TBS), for five minutes. Slides were incubated overnight at 4°C with fluorescently labelled anti-sheep secondary in in antibody carrier solution (1% BSA in TBS with 0.1% Tween-20). Following three washes of five minutes in TBS, slides were mounted with VECTASHIELD™ Antifade Mounting Medium with DAPI (Vector Laboratories) and imaged with an Axioplan 2 upright microscope and an AxioVision HRc CCD camera (both Zeiss).

Direct co-staining failed due to loss of the FISH signal during IHC. Therefore as a preliminary experiment, the two techniques were applied to adjacent sections of the same mouse brain. FISH for mouse *Cerox1* labelled a number of cells throughout the striatum (Supplemental Figure S10B), and specificity was supported by a lack of comparable staining in a no anti-DIG primary antibody control (Supplemental Figure S10C). Unfortunately, staining of an adjacent section with anti-Calb1 as above did not provide such clear demarcation of striosomes, with extensive cross-reaction of the antibody with myelin bundles (Supplemental Figure S10D). This precluded clear identification of the striosome network in a composite image (Supplemental Figure S10E). It remains unclear, therefore, based upon these few sections, whether *Cerox1* is associated with striosomes here.

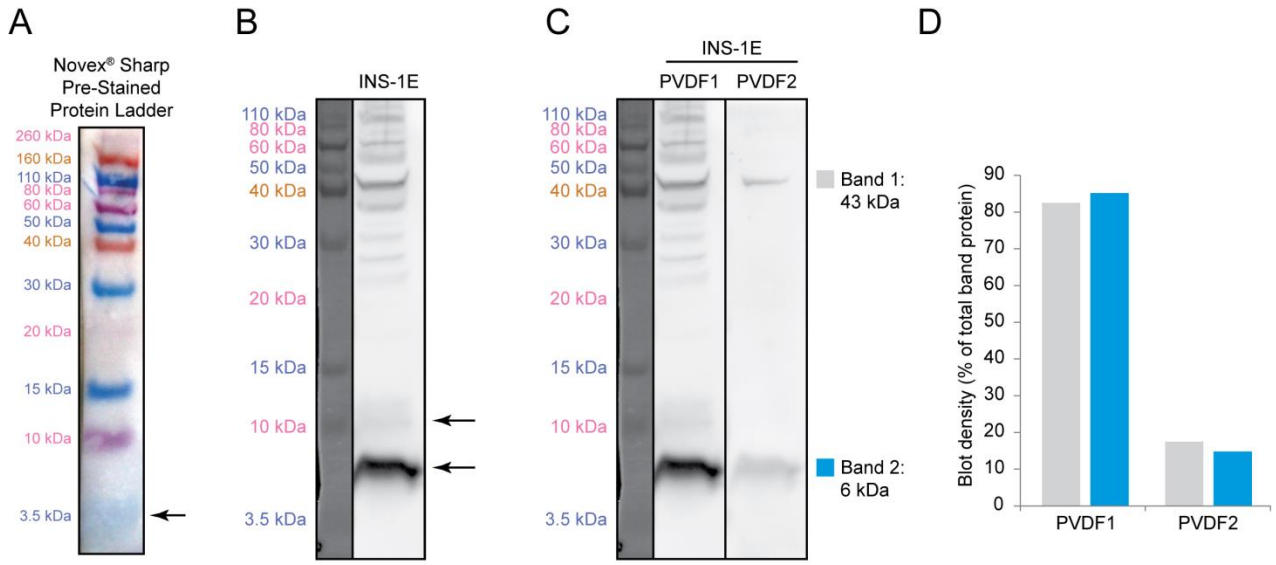
Guidance and antibodies for Oprm1 and Calb1 were provided by Dr Katherine Brimblecomb of the Cragg laboratory, DPAG.



Supplemental Figure S10: Investigating the *Cerox1*-striosome relationship. (A) A coronal section of adult mouse brain stained for Calb1 reveals depleted patches consistent with striosomes (white dashed lines). FISH of *Cerox1* **(B)** stained a number of cells specifically compared to a no anti-DIG primary control **(C)**. **(D)** Calb1 staining of an adjacent section did not stain as well as previously, with a lot of cross-reaction with myelin bundles. **(E)** Composite of **(B)** and **(D)**. Grey dashed lines outline the caudoputamen and pallidum, by comparison to the Allen Brain Reference Atlas.

A8 Optimised western blotting permits detection of very small proteins

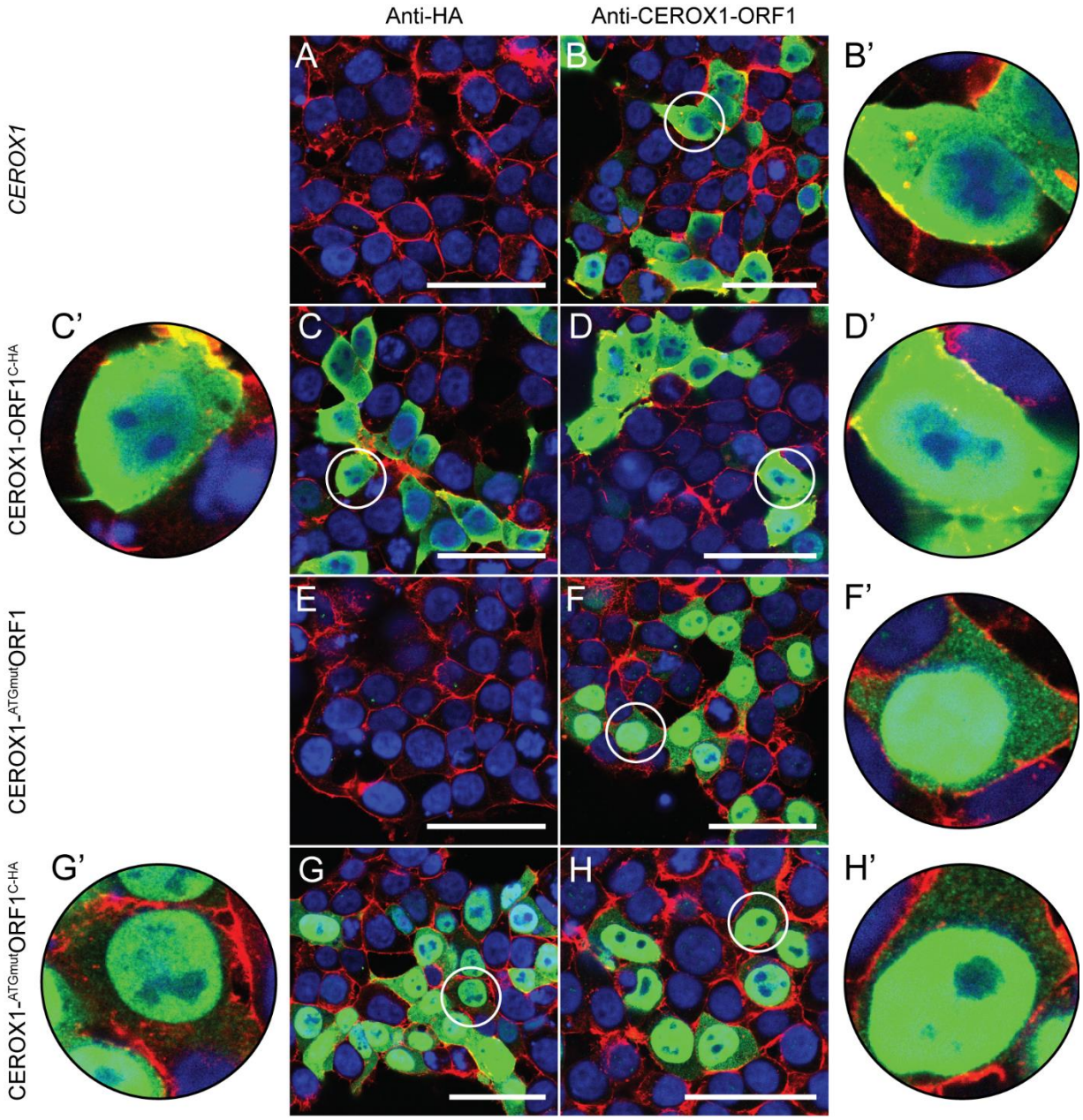
An optimised protocol was devised in order to minimise any loss of very small proteins. Namely, using high percentage (12%) NuPAGE gels with slow-running MES SDS running buffer to maximise protein separation in the small range. Proteins were transferred onto 0.2 µm PVDF, alongside a sample of INS-1 rat pancreatic β-cell protein, such that the small protein insulin may be used as a small protein retention control (Supplemental Figure S11B-C). Short, low voltage transfer conditions ensured minimal bleed-through of small proteins, as demonstrated using layered PVDF membranes (Supplemental Figure S11D).



Supplemental Figure S11: Small proteins are retained and detectable during the western blotting protocol used in the experiments of Chapter 4. (A) The smallest band of the Novex® Sharp Pre-stained Protein Standard ladder is 3.5 kDa, and is visible by its blue colour on a PVDF membrane following transfer (arrow). **(B)** Blotting total INS-1 protein with an anti-insulin antibody reveals a clear strong band for insulin at around 6 kDa, and a much fainter band at around 10 kDa representing pro-insulin (arrows). **(C)** Inclusion of a second membrane during transfer and subsequent parallel blotting demonstrates that, while some protein is transferred through the first PVDF membrane, the majority is retained. **(D)** Densitometry analysis of the bands on the respective membranes indicated that 15-17% of the total protein detected in the strongest bands has transferred onto the second membrane.

A9 CEROX1-ORF1 may have cryptic start codons

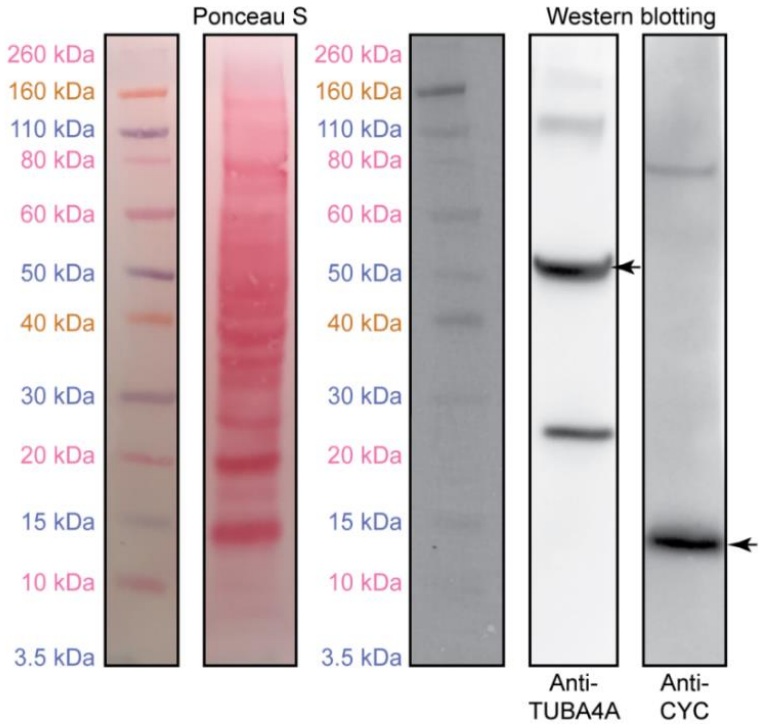
Following mutation of the predicted start codon for CEROX1-ORF1, both western blotting (Chapter 4, Figure 4.9) and ICC (Supplemental Figure S12) indicated a reduction, but not complete loss, of signal. Bands remained on the western blot, and fluorescence was still observed in transfected cells. Intriguingly, it was found with both the anti-HA and the anti-CEROX1-ORF1 antibodies that ablation of the CEROX1-ORF1 ATG start codon depleted most cytoplasmic CEROX1-ORF1 signal, but that nuclear signal strongly remained (consider B vs. F, C vs G, and D vs H, in Supplemental Figure S12). It may be the case that a cryptic upstream start codon is responsible for this leaky expression, and the inclusion of a nuclear retention signal in the extended ORF, for example.



Supplemental Figure S12: Ablation of the CEROX1-ORF1 start codon does not remove all signal associated with the ORF. Transfected HEK293T cells were stained by ICC using anti-HA (left) or anti-CEROX1-ORF1 (right) antibodies (green). Strong generalised cytoplasmic signal was observed for *CEROX1* **(A-B)** and CEROX1-ORF1$^{C\text{-}HA}$ **(C-D)**. Following deletion of the ATG start codon, however, cytoplasmic signal was largely lost, leaving only concentrated nuclear fluorescence **(E-H)**. All scale bars represent 50 µm. Cells were counterstained with DAPI (blue) for the nucleus, and phalloidin (red), for the actin cytoskeleton (membrane).

A10 Integrity of gorilla cortex protein

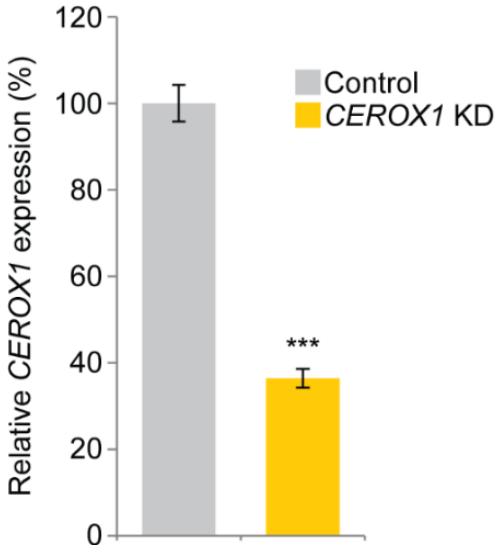
In order to confirm the integrity of protein extracted from frozen gorilla cortex, western blotting and Ponceau S staining were used to demonstrate discrete and specific bands of protein, suggesting minimal overall degradation (Supplemental Figure S13).



Supplemental Figure S13: High integrity protein was extracted from frozen gorilla cortex for use in IP-MS experiments. Following homogenisation of tissue and extraction of protein, integrity was evaluated by SDS-PAGE. Staining with the total protein stain Ponceau S indicated distinct bands of protein across the full size range (left), while western blotting with two antibodies, against α-tubulin (TUBA4A) and cytochrome *c* (CYC), revealed clear bands of correct sizes (right).

A11 *CEROX1* knock-down was achieved

Following the poor results obtained using shRNAs, *CEROX1* knock-down was attempted using locked nucleic acid (LNA™) GapmeRs (Exiqon, Denmark). Transfection of the most efficacious GapmeR resulted in a reproducible decrease in *CEROX1* levels of 64% (p = 2.71 × 10^{-6}) (Figure 5.16). Owing to the recent poor reproducibility of the overexpression phenotype and the expense of the GapmeRs, assay of complex I activity following *CEROX1* knock-down was not completed.



Supplemental Figure S14: *CEROX1* knockdown was produced using LNA™ GapmeRs. Following transfection of HEK293T cells with Exiqon GapmeRs at 50 nM, *CEROX1* RNA levels were assayed by RT-qPCR and normalised to those of *POLR2A* and *TUBA1A*.

SUPPLEMENTARY REFERENCES

1. Sirey, T.M., Roberts, K., Haerty, W., Bedoya-Reina, O. *et al.* The long non-coding RNA Cerox1 is a post transcriptional regulator of mitochondrial complex I catalytic activity. *eLife* **8**, e45051 (2019).
2. Sirey, T.M., Roberts, K., Haerty, W., Bedoya-Reina, O. *et al.* Cerox1 and microRNA-488-3p noncoding RNAs jointly regulate mitochondrial complex I catalytic activity. *bioRxiv* (2018).
3. Shi, X. & Jarvis, D.L. A new RACE method for extremely GC-rich genes. *Analytical biochemistry* **356**, 222-228 (2006).
4. Henke, W., Herdel, K., Jung, K., Schnorr, D. & Loening, S.A. Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Research* **25**, 3957-3958 (1997).
5. Rees, W.A., Yager, T.D., Korte, J. & Von Hippel, P.H. Betaine can eliminate the base pair composition dependence of DNA melting. *Biochemistry* **32**, 137-144 (1993).
6. Horáková, H., Polakovičová, I., Shaik, G.M., Eitler, J. *et al.* 1,2-propanediol-trehalose mixture as a potent quantitative real-time PCR enhancer. *BMC Biotechnology* **11**, 41 (2011).
7. Zhang, Z., Yang, X., Meng, L., Liu, F. *et al.* Enhanced amplification of GC-rich DNA with two organic reagents. *Biotechniques* **47**, 775-9 (2009).
8. Mousavian, Z., Sadeghi, H., Sabzghabaee, A. & Moazen, F. Polymerase chain reaction amplification of a GC rich region by adding 1,2 propanediol. *Advanced Biomedical Research* **3**, 65-65 (2014).
9. Schnoor, M., Voß, P., Cullen, P., Böking, T. *et al.* Characterization of the synthetic compatible solute homoectoine as a potent PCR enhancer. *Biochemical and Biophysical Research Communications* **322**, 867-872 (2004).
10. Hezroni, H., Koppstein, D., Schwartz, Matthew G., Avrutin, A. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11**, 1110-1122 (2015).
11. Hezroni, H., Perry, R.B.T., Meir, Z., Housman, G. *et al.* A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biology* **18**(2017).
12. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580 (2013).
13. Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
14. Merkin, J., Russell, C., Chen, P. & Burge, C.B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593-9 (2012).
15. Chen, M., Hu, Y., Liu, J., Wu, Q. *et al.* Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome. *Sci Rep* **5**, 18019 (2015).

The publication aforementioned in Appendix A1 is reproduced hereon in full:

The long non-coding RNA Cerox1 is a post transcriptional regulator of mitochondrial complex I catalytic activity. Tamara M. Sirey, Kenny Roberts, Wilfried Haerty, Oscar Bedoya-Reina, Sebastian Rogatti-Granados, Jennifer Y. Tan, Nick Li, Lisa C. Heather, Roderick N. Carter, Sarah Cooper, Andrew J. Finch, Jimi Wills, Nicholas M. Morton, Ana Claudia Marques, Chris P. Ponting.

Available through eLife, 2nd May 2019: doi.org/10.7554/eLife.45051

© 2019, Sirey et al.

"This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited."

eLIFE elifesciences.org

RESEARCH ARTICLE

The long non-coding RNA *Cerox1* is a post transcriptional regulator of mitochondrial complex I catalytic activity

Tamara M Sirey[1,2]*, Kenny Roberts[2], Wilfried Haerty[2†], Oscar Bedoya-Reina[1,2‡], Sebastian Rogatti-Granados[1,2], Jennifer Y Tan[2§], Nick Li[2], Lisa C Heather[3], Roderick N Carter[4], Sarah Cooper[5#], Andrew J Finch[1], Jimi Wills[1], Nicholas M Morton[4], Ana Claudia Marques[2§], Chris P Ponting[1,2]*

[1]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom; [2]MRC Functional Genomics Unit, University of Oxford, Oxford, United Kingdom; [3]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom; [4]University/British Heart Foundation Centre for Cardiovascular Science, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, United Kingdom; [5]Department of Biochemistry, University of Oxford, Oxford, United Kingdom

***For correspondence:** tamara.sirey@igmm.ed.ac.uk (TMS); Chris.Ponting@igmm.ed.ac.uk (CPP)

Present address: †Earlham Institute, Norwich, United Kingdom; ‡Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Solna, Sweden; §Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; #Wellcome Sanger Institute, Cambridge, United Kingdom

Competing interest: See page 25

Funding: See page 25

Received: 10 January 2019
Accepted: 02 May 2019
Published: 02 May 2019

Reviewing editor: Detlef Weigel, Max Planck Institute for Developmental Biology, Germany

© Copyright Sirey et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Abstract To generate energy efficiently, the cell is uniquely challenged to co-ordinate the abundance of electron transport chain protein subunits expressed from both nuclear and mitochondrial genomes. How an effective stoichiometry of this many constituent subunits is co-ordinated post-transcriptionally remains poorly understood. Here we show that *Cerox1*, an unusually abundant cytoplasmic long noncoding RNA (lncRNA), modulates the levels of mitochondrial complex I subunit transcripts in a manner that requires binding to microRNA-488-3p. Increased abundance of *Cerox1* cooperatively elevates complex I subunit protein abundance and enzymatic activity, decreases reactive oxygen species production, and protects against the complex I inhibitor rotenone. *Cerox1* function is conserved across placental mammals: human and mouse orthologues effectively modulate complex I enzymatic activity in mouse and human cells, respectively. *Cerox1* is the first lncRNA demonstrated, to our knowledge, to regulate mitochondrial oxidative phosphorylation and, with miR-488-3p, represent novel targets for the modulation of complex I activity.
DOI: https://doi.org/10.7554/eLife.45051.001

Introduction

In eukaryotes, coupling of the mitochondrial electron transport chain to oxidative phosphorylation (OXPHOS) generates the majority of ATP that fulfils cellular energy requirements. The first enzyme of the electron transport chain, NADH:ubiquinone oxidoreductase (complex I), catalyses the transfer of electrons from NADH to coenzyme Q10, pumps protons across the inner mitochondrial membrane and produces reactive oxygen species (ROS). Mammalian mitochondrial complex I dynamically incorporates 45 distinct subunits into a ~ 1 MDa mature structure (***Vinothkumar et al., 2014***; ***Guerrero-Castillo et al., 2017***). It is known that oxidatively damaged subunits can be exchanged in the intact holo-enzyme (***Dieteren et al., 2012***), but how this process may be regulated is poorly understood. The efficiency and functional integrity of OXPHOS are thought to be partly maintained through a combination of tightly co-ordinated transcriptional and post-transcriptional regulation

eLife digest Animal cells generate over 90% of the energy they need within small structures called mitochondria. Converting food into energy requires many different proteins and cells control the relative amounts of the proteins in mitochondria to ensure this process is efficient. To make more of a given protein, the cell must copy the DNA of the gene that encodes it into another molecule known as a messenger RNA, before reading the instructions in the messenger RNA to build the protein. However, this is not the only way that a cell uses molecules of RNA.

A second group of RNAs called long non-coding RNAs (or lncRNAs) can help regulate the production of proteins in complex ways, and each lncRNA can have an effect across multiple genes. Some lncRNAs, for example, stop a third group of RNAs – microRNAs – from blocking certain messenger RNAs from being read. Sirey et al. set out to answer whether a lncRNA might help to co-ordinate the production of the many proteins needed by mitochondria.

In experiments with mouse cells grown in the laboratory, Sirey et al. identified a lncRNA called *Cerox1* that can co-ordinate the levels of at least 12 mitochondrial proteins. A microRNA called miR-488-3p suppresses the production of many of these proteins. By binding to miR-488-3p, *Cerox1* blocks the effects of the microRNA so more proteins are produced. Sirey et al. artificially altered the amount of *Cerox1* in the cells and showed that more *Cerox1* leads to higher mitochondria activity. Further experiments revealed that this same control system also exists in human cells.

Mitochondria are vital to cell survival and changes that affect their efficiency can be fatal or highly debilitating. Reduced efficiency is also a hallmark of ageing and contributes to conditions including cardiovascular disease, diabetes and Parkinson's disease. Understanding how mitochondria are regulated could unlock new treatment methods for these conditions, while a better understanding of the co-ordination of protein production offers other insights into some of the most fundamental biology.

DOI: https://doi.org/10.7554/eLife.45051.002

(***Mootha et al., 2003***; ***van Waveren and Moraes, 2008***; ***Sirey and Ponting, 2016***) and specific sub-cytoplasmic co-localisation (***Matsumoto et al., 2012***; ***Michaud et al., 2014***). The nuclear encoded subunits are imported into the mitochondria after translation in the cytoplasm and their complexes assembled together with the mitochondrially encoded subunits in an intricate assembly process (***Perales-Clemente et al., 2010***; ***Lazarou et al., 1793***; ***Vogel et al., 2007***). Mitochondrial biogenesis is co-ordinated first transcriptionally from both genomes (***Scarpulla et al., 2012***), and then post-transcriptionally by regulatory small noncoding RNAs such as microRNAs (miRNAs) (***Dumortier et al., 2013***; ***Li et al., 2012***). Recently, SAMMSON a long noncoding RNA (lncRNA) was found to bind p32 and, within mitochondria, enhanced the expression of mitochondrial genome-encoded polypeptides (***Leucci et al., 2016***).

Nuclear-encoded and cytosol-located lncRNAs have not yet been implicated in regulating mitochondrial OXPHOS (***Vendramin et al., 2017***) despite being surprisingly numerous and often found localised to mitochondrion- and ribosome-adjacent portions of the rough endoplasmic reticulum (***van Heesch et al., 2014***). It is here, on the ribosome, that turnover of miRNA-targeted mRNAs frequently occurs during their translation (***Tat et al., 2016***). Here we describe a novel mammalian conserved lncRNA, termed *Cerox1* (cytoplasmic endogenous regulator of oxidative phosphorylation 1). *Cerox1* regulates complex I activity by co-ordinately regulating the abundance of at least 12 complex I transcripts via a miRNA-mediated mechanism. *Cerox1* knockdown decreases the enzymatic activities of complexes I and IV. Conversely, elevation of *Cerox1* levels increases their enzymatic activities, halves cellular oxidative stress, and protects cells against the cytotoxic effects of the complex I inhibitor rotenone. To our knowledge, *Cerox1* is the first lncRNA modulator of normal mitochondrial energy metabolism homeostasis and cellular redox state. The miRNA-dependency of *Cerox1* and the regulation of associated OXPHOS transcripts are supported by: (i) direct physical interaction of miR-488–3p with *Cerox1* and complex I transcripts; (ii) decrease or increase in *Cerox1* and complex I transcripts following miR-488–3p overexpression or inhibition, respectively; (iii) miR-488–3p destabilisation of wildtype *Cerox1*, but not a *Cerox1* transcript containing a mutated miR-488–3p miRNA recognition element (MRE) seed region; and, (iv) absence of the OXPHOS

phenotypes either in cell lines deficient in microRNA biogenesis or when *Cerox1*'s predicted miR-488–3p response element is mutated. The miRNA-dependent role of *Cerox1* illustrates how RNA-interaction networks can regulate OXPHOS and that lncRNAs represent novel targets for modulating OXPHOS enzymatic activity.

Results

Cerox1 is a conserved, highly expressed long noncoding RNA

Cerox1 was selected for further investigation from among a set of central nervous system-derived polyadenylated long non-coding RNAs identified by cDNA sequencing (GenBank Accession AK079380, 2810468N07Rik) (**Carninci et al., 2000**; **Ponjavic et al., 2007**). Mouse *Cerox1* is a 1.2 kb, two exon, intergenic transcript which shares a bidirectional promoter with the SRY (sex determining region Y)-box 8 (*Sox8*) gene (**Figure 1A**). A human orthologous transcript (*CEROX1*, GenBank Accession BC098409) was identified by sequence similarity and conserved synteny (60–70% nucleotide identity within alignable regions, **Figure 1B,C**). Both mouse and human transcripts have low protein coding potential (Materials and methods, **Figure 1—figure supplement 1A**) and no evidence for translation from available proteomic datasets.

Four human or mouse data types supported *CEROX1* as having an important organismal role. First, its promoter shows a greater extent of sequence conservation than the adjacent *SOX8* promoter and its exons are conserved among eutherian mammals (**Figure 1B**). Second, from expression data, its levels in primary tissues and cells are exceptionally high, within the top 13% of a set of 879 lncRNAs with associated cap analysis of gene expression (CAGE) clusters (**Figure 1D,E**). Expression is particularly high in neuroglia, neural progenitor cells and oligodendrocyte progenitors (**Bergmann et al., 2015**; **Mercer et al., 2010**; **Zhang et al., 2014**) (**Figure 1—figure supplement 1B**). *Cerox1* is notable in its higher expression in the adult brain than 64% of all protein coding genes. *Cerox1* expression is also developmentally regulated. For example, it is a known marker gene for type one pre-haematopoietic stem cells in the dorsal aorta (**Zhou et al., 2016**) and its expression is high in mouse embryos beyond the 21 somite stage (https://dmdd.org.uk/). High expression of *Cerox1* in the brain was confirmed using quantitative real-time PCR (qPCR) for both mouse and human orthologous transcripts (**Figure 1—figure supplement 1C,D**).

Third, single nucleotide variants are significantly associated both with *CEROX1* expression and anthropomorphic traits, measured in the UK Biobank. These lie within an 85 kb interval encompassing the *CEROX1* locus, and 5' regions of *SOX8* and *LMF1* genes. For example, rs3809674 is significantly associated with standing and sitting heights; arm fat-free mass (left or right); arm predicted mass (left or right); trunk fat-free or predicted mass; and whole body fat-free mass ($p<5\times10^{-8}$, **Supplementary file 1**); this variant is also a *CEROX1* expression quantitative trait locus (eQTL) for 30 GTEx tissues ($p\leq0.05$), some with absolute normalised effect sizes reaching 0.83 (**Figure 1—figure supplement 1E**). Genetically determined expression change in *CEROX1* therefore could explain, in part, variation in these anthropomorphic traits. These variations affect a large fraction of the human population (minor allele frequency of 28% in 1000 Genomes data). An alternative interpretation that rs3809674 (and linked variants) act on anthropomorphic traits through *LMF1*, rather than *CEROX1*, is consistent with this variant being an eQTL for *LMF1*, but is not consistent with the known protein function of LMF1, a lipase maturation factor, because the associated anthropomorphic traits relate only to fat-free mass. A summary-data-based Mendelian randomization analysis that uses rs3809674 as an instrumental variable, predicts an effect of *CEROX1* gene expression on glioma risk (**Melin et al., 2017**). Finally, it has recently been demonstrated in a mouse model of haematopoietic lineage differentiation that *Cerox1* depletion may impair the contributions of stem cell or B cell differentiation to haematopoiesis (**Delás et al., 2019**).

In mouse neuroblastoma (N2A) cells the most highly expressed *Cerox1* transcript (**Figure 1—figure supplement 1F,G**) is enriched in the cytoplasmic fraction (**Figure 1F**) with a short half-life of 36 ± 16 mins (**Figure 1—figure supplement 1H**) and is mainly associated with the ribosome-free fraction (**Figure 1—figure supplement 1I**).

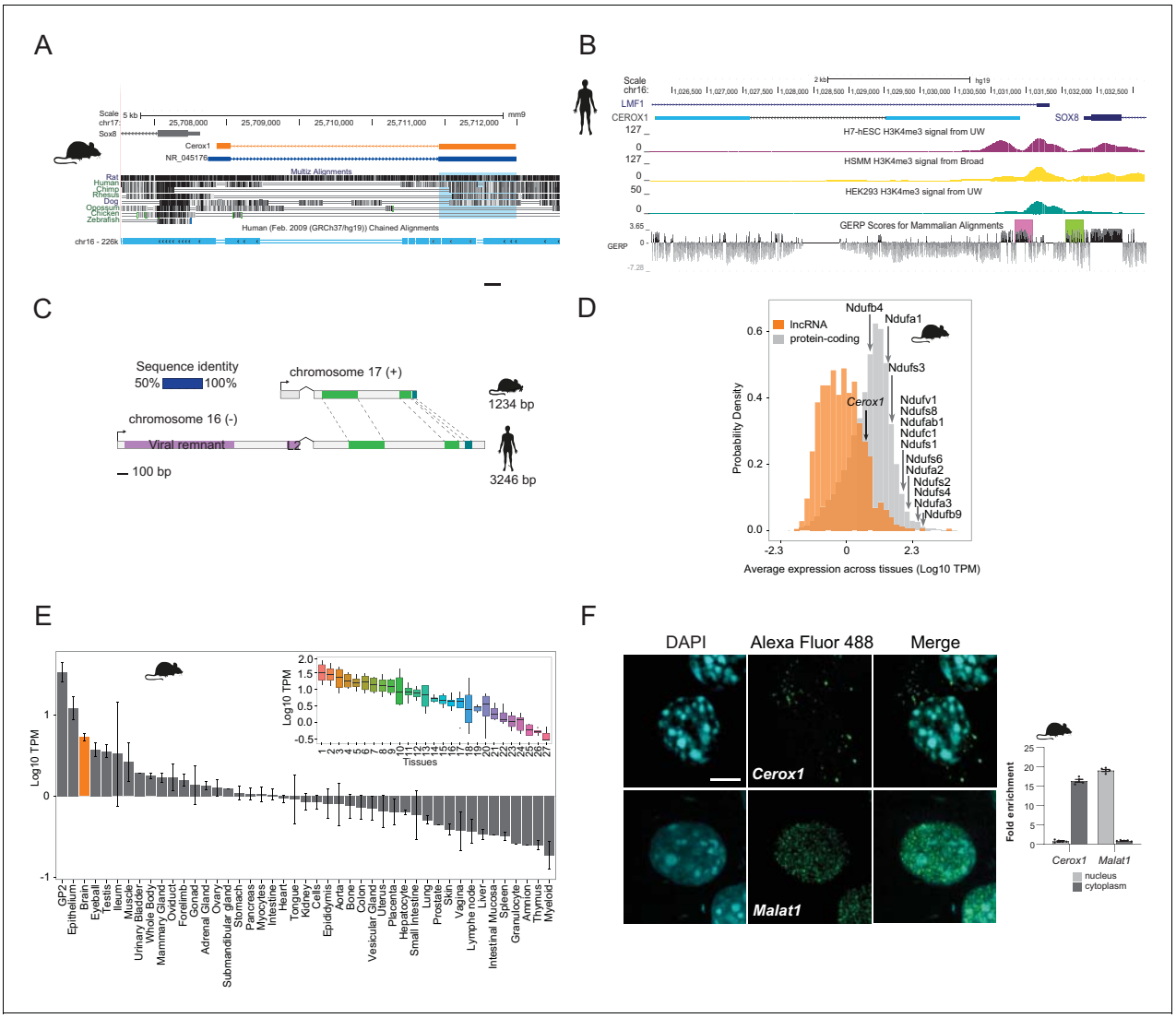


Figure 1. *Cerox1* is an evolutionarily conserved, highly expressed and predominantly cytoplasmic lncRNA. (**A**) The mouse *Cerox1* locus (mm9 assembly). Sequence shaded in blue highlights conservation within exon two among eutherian mammals, but not in non-mammalian vertebrates such as chicken and zebrafish. (**B**) Syntenic human locus (hg19). This transcript was previously identified on the minus strand as *LMF1* non-coding transcript 4, and is located within the intron of *LMF1* non-coding transcript 2. *LMF1* non-coding transcript two is annotated as a *nonsense mediated decay* biotype. A prominent peak of H3K4me3 (tri-methylation of histone H3 lysine 4 often found near promoters) modification marks the *CEROX1* transcriptional start site. H3K4me3 peaks for H7-human embryonic stem cells (H7-hESC), human skeletal muscle myoblasts (HSMM) and human embryonic kidney 293 cells (HEK293) are depicted. The genomic evolutionary rate profiling (GERP) score indicates a higher extent of conservation of the *CEROX1* promoter (shaded pink) than the adjacent *SOX8* promoter (shaded green). (**C**) Schematic representation of mouse *Cerox1* transcript and the human orthologous sequence. Exon two contains blocks of 60–70% sequence identity; human *CEROX1* has an additional 1235 bases of retrotransposed insertions at the 5' end. (**D**) Distributions of lncRNAs' and protein-coding genes' average expression levels across tissues in mouse. Average expression levels of representative mitochondrial complex I subunits' mRNAs are indicated. TPM = tags per million. (**E**) Average expression levels of *Cerox1* across mouse tissue samples. The orange bar highlights nervous system tissue samples whose values for replicates among neurological tissues are shown in the inset panel: 1- Medulla oblongata, 2– Spinal cord, 3– Diencephalon, 4– Substantia nigra, 5– Microglia, 6– Raphe, 7– Dorsal spinal cord, 8– Corpora quadrigemina, 9– Cortex, 10– Corpus striatum, 11- Visual cortex, 12– Olfactory brain, 13– Cerebellum, 14– Neurospheres sympathetic neuron derived, 15– Neurospheres parasympathetic neuron derived, 16– Neurospheres enteric neuron derived, 17– Astrocytes (cerebellar), 18– Hippocampus, 19– Hippocampal, 20– Ventral spinal cord, 21– Astrocytes, 22– Pituitary gland, 23– Astrocytes (hippocampus), 24– Cortical neurons, 25– Striatal neurons, 26– Schwann cells, 27– Meningeal cells. Error bars indicate s.e.m. (**F**) Cytoplasmic localisation of mouse *Cerox1* compared to a nuclear retained lncRNA,

Figure 1 continued on next page

Figure 1 continued

Malat1, as demonstrated by fluorescent in situ hybridization and cell fractionation followed by quantitative PCR. By fractionation, mouse *Cerox1* is 15-fold enriched in the cytoplasm of N2A cells (n = 5; error bars s.e.m.). Scale bar = 5 μm.

DOI: https://doi.org/10.7554/eLife.45051.003

The following figure supplement is available for figure 1:

Figure supplement 1. Transcript characterisation.

DOI: https://doi.org/10.7554/eLife.45051.004

Cerox1 expression modulates levels of oxidative phosphorylation transcripts

Expression of *Cerox1* was manipulated by transient overexpression or shRNA-mediated knockdown. Twelve shRNAs were tested for the ability to knock-down *Cerox1* (***Figure 2—figure supplement 1A***). One of these (sh92) decreased expression levels by greater than 60%, with the next best shRNA (sh1159) decreasing expression by approximately 40%. As expected from their sharing of the *Cerox1-Sox8* bidirectional promoter, CRISPR-mediated activation and inhibition of the *Cerox1* locus was not specific as it led also to changes in *Sox8* expression (***Figure 1A***, ***Figure 2—figure supplement 1B***). However, decreasing *Cerox1* levels in N2A cells by shRNA or transient overexpression had no effect on the expression of neighbouring genes (***Figure 2—figure supplement 1C***). In contrast, *Cerox1* overexpression led to differential expression of 286 distal genes (q < 0.05, Bonferroni multiple testing correction; ***Supplementary file 2***), of which an unexpected and large majority (83%; 237) were upregulated ($p<10^{-6}$; binomial test). Our attention was immediately drawn to the considerable (≥20 fold) enrichment of the mitochondrial respiratory chain gene ontology term among upregulated genes (***Figure 2A***).

The mitochondrial electron transport chain (ETC) consists of five multi-subunit complexes encoded by approximately 100 genes of which only 13 are located in the mitochondrial genome. The 15 ETC transcripts that show statistically significant differential expression after *Cerox1* overexpression are nuclear encoded (***Figure 2B,C***) with the greatest changes observed by qPCR for complex I subunit transcripts (***Figure 2—figure supplement 1D***). Twelve of 35 nuclear encoded complex I subunits or assembly factors transcripts increased substantially and significantly (>40%) following *Cerox1* overexpression; we consider these to be gene expression biomarkers for *Cerox1* activity in the mouse N2A system (***Figure 2C***). In the reciprocal *Cerox1* knock-down experiment, all 12 were reduced in abundance using sh92, three significantly, with a concordant pattern observed for the less effective shRNA, sh1159 (***Figure 2—figure supplement 1E,F***). Taken together, these results indicate that *Cerox1* positively and co-ordinately regulates the levels of many mitochondrial complex I transcripts.

Increased abundance of OXPHOS subunit transcripts, following *Cerox1* overexpression, was found to elevate protein levels. Western blots using reliable antibodies for the key complex I catalytic core proteins NDUFS1 and NDUFS3 showed approximately 2.0-fold protein level increases that surpassed their ~ 1.4 fold transcript level changes (median 2.4 and 1.4-fold increases [p=0.0013 and 0.002], respectively; ***Figure 2D***). *Cerox1* transcript abundance is thus coupled positively to OXPHOS transcript levels and to their availability for translation, resulting in an amplification of the amount of protein produced. In summary, protein subunits of the same complex (Complex I) that are sustained at high abundance and with long half-lives (***Dörrbaum et al., 2018***; ***Mathieson et al., 2018***; ***Schwanhäusser et al., 2011***) (***Figure 2—figure supplement 1G***), and whose transcripts are stable (***Schwanhäusser et al., 2011***; ***Friedel et al., 2009***; ***Sharova et al., 2009***; ***Tani et al., 2012***) (***Figure 2—figure supplement 1H***) and also have very high copy numbers in the cell (***Schwanhäusser et al., 2011***; ***Cao et al., 2017***), can be increased two-fold, and co-ordinately, by the simple expediency of increasing the level of this single abundant lncRNA.

These large effects on protein and mRNA copy number are associated with both metabolic and cellular phenotypes. Ten metabolites are significantly different in N2A cells overexpressing *Cerox1* (***Figure 2E***). These cells show a significant increase in the reduced glutathione to oxidized glutathione ratio (GSH:GSSG, $p=1.3\times10^{-3}$, ***Figure 2D*** Inset) indicative of a more favourable cellular redox state. *Cerox1*-overexpressing cells also exhibited a 43% reduction in cell cycle activity, yet without a

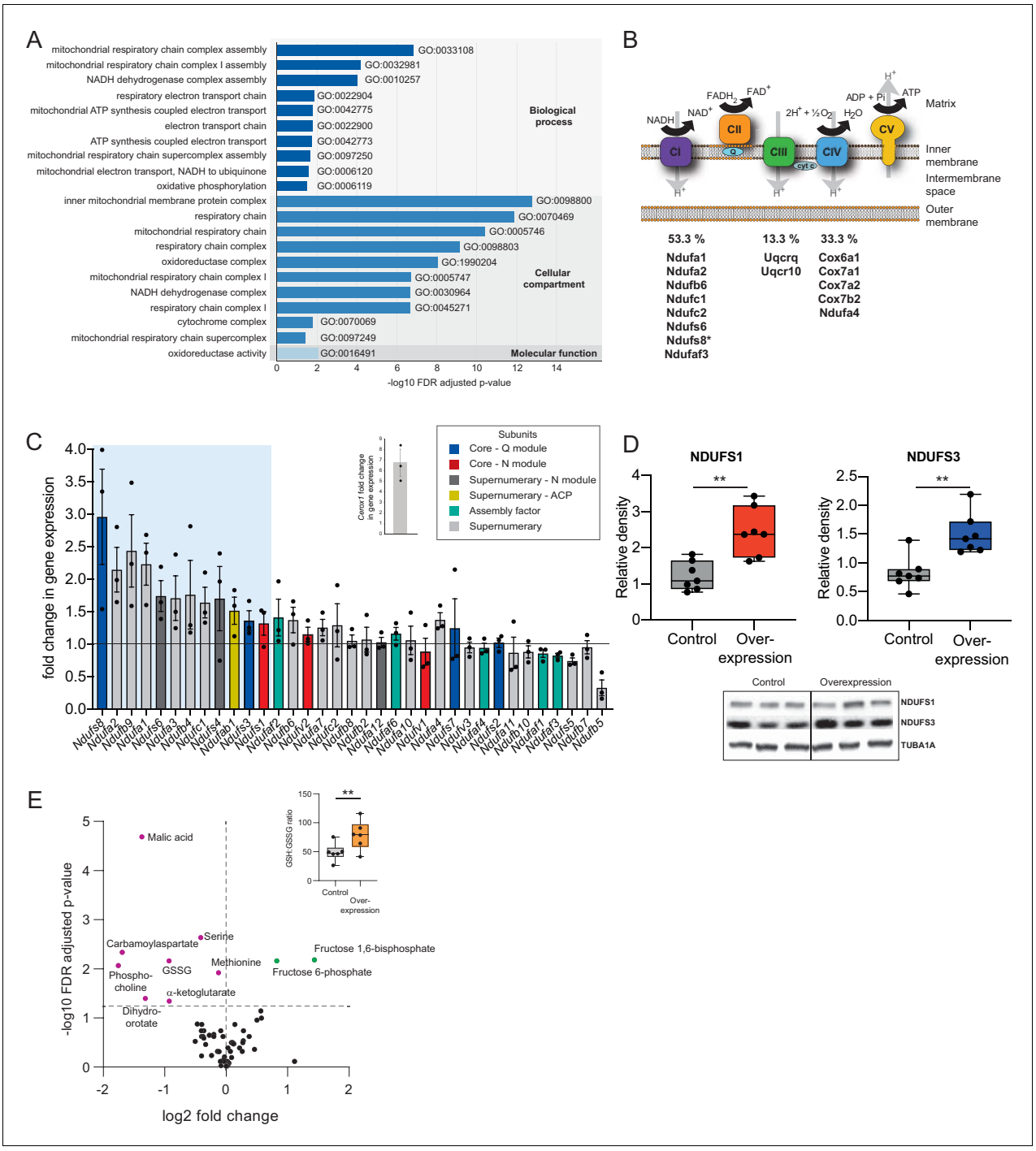


Figure 2. *Cerox1* overexpression elevates levels of OXPHOS transcripts and their encoded proteins. (**A**) Gene ontology analysis indicates a significant enrichment of upregulated genes involved in mitochondrial electron transport, energy production and redox reactions. (**B**) Four membrane bound multi-subunit complexes (CI, CII, CIII, CIV) are embedded in the inner mitochondrial membrane and facilitate transfer of electrons; three of these subunits are also proton pumps which create the chemiosmotic gradient required for ATP synthase activity, with complex V being ATP synthase. The subunits vary in size and complexity with Complex I (NADH:ubiquinone oxidoreductase) consisting of 45 subunits, Complex II (succinate

Figure 2 continued on next page

Figure 2 continued

dehydrogenase) four subunits, Complex III (Ubiquinol:cytochrome *c* oxidoreductase) 11 subunits and Complex IV (Cytochrome *c* oxidase) 13 subunits. Of 15 oxidative phosphorylation genes whose transcripts were up-regulated following *Cerox1* overexpression 53% were subunits of Complex I, 13% were subunits of Complex III and 33% were subunits of Complex IV. * indicates core subunits that are essential for activity. Note: subunit NDUFA4 has recently been reassigned to mitochondrial complex IV (***Balsa et al., 2012***). (**C**) qPCR profiling of 35 complex I subunits and assembly factors (30 nuclear encoded complex I subunits and five assembly factors). Transcripts showing a 1.4 fold, or greater, change in expression after overexpression of *Cerox1* are present within the boxed shaded area. Fold change of wild-type *Cerox1* compared to the control are indicated in the inset panel. The transcripts profiled can be characterised into six categories: Core–Q module, subunits responsible for the electron transfer to ubiquinone; Core–N module, subunits responsible for the oxidation of NADH; Supernumerary subunits– those that are additional to the core subunits required for the catalytic role of complex I, but do not play a catalytic role themselves. Many of these subunits may be performing a structural role, but the majority are of unknown function. The supernumerary subunits can be further subdivided into supernumerary – N module, those accessory subunits associated with the NADH oxidation module of CI; supernumerary ACP (acyl carrier protein) – in addition to being a non-catalytic subunit of CI, NDUFAB1 is also a carrier of the growing fatty acid chain in mitochondrial fatty acid biosynthesis; assembly factor, proteins that are required for the correct assembly and integration of CI. Error bars s.e.m. (*n* = 3 biological replicates). (**D**) Overexpression of *Cerox1* results in large increases in the total protein levels of two core subunits for which high quality antibodies exist, normalised to the loading control α-tubulin (TUBA1A). NDUFS1 is one of three (NDUFS1, NDUFV1, NDUFV2) core components of the N-module of Complex I. NDUFS3 is one of four (NDUFS2, NDUFS3, NDUFS7, NDUFS8) core components of the Complex I Q-module. *n* = 7 biological replicates for control and overexpression. 2-sided *t*-test; **p<0.01. (**E**) Overexpression of *Cerox1* results in a change in the metabolite profile of N2A cells, with 10 of 66 metabolites measured demonstrating a significant change in the experimental sample after multiple testing correction (q < 0.05; *n* = 6 biological replicates for pCAG-control and pCAG *Cerox1* overexpression). N2A cells overexpressing *Cerox1* show an increased GSH:GSSG ratio (figure inset, 2-sided *t*-test; **p<0.01).

DOI: https://doi.org/10.7554/eLife.45051.005

The following source data and figure supplement are available for figure 2:

Source data 1. N2A metabolomics profiling - ***Figure 2E***.

DOI: https://doi.org/10.7554/eLife.45051.007

Figure supplement 1. Manipulation of *Cerox1* expression levels.

DOI: https://doi.org/10.7554/eLife.45051.006

change in the proportion of live/dead cells or a deviation from normal cell cycle proportions (***Figure 2—figure supplement 1I,J,K***). *Cerox1* levels thus affect the overall timing of cell division.

Cerox1 can regulate mitochondrial OXPHOS enzymatic activity

Increased translation of some complex I transcripts leads to increased respiration (***Shyh-Chang et al., 2013***) and, more specifically, to an increase in the enzymatic activity of complex I (***Alvarez-Fischer et al., 2011***). To address this hypothesis we used oxidative phosphorylation enzyme assays to investigate whether changes in expression to a subset of subunits lead to a change in enzyme activity and oxygen consumption. Indeed, complex I and complex IV enzymatic activities increased substantially after *Cerox1* overexpression (by 22%, p=0.01; by 50%, p=0.003, respectively; 2-tailed Student's *t*-test; ***Figure 3A***). Such rate increases for two of the eukaryotic cell's most abundant and active enzymes were unexpected.

We next measured oxygen consumption under these conditions using a Seahorse XF^{e}24 Analyzer. These complexes' more rapid catalytic rates resulted, unexpectedly, in large increases in: (i) overall basal oxygen consumption (by 85%), (ii) ATP-linked oxygen consumption (by 107%) and, (iii) maximum uncoupled respiration (by 59%; $p=5\times10^{-4}$, $p=1\times10^{-4}$, $p=4\times10^{-3}$, respectively, 2-tailed Student's *t*-test; ***Figure 3B***). These increases in enzyme activities and mitochondrial respiration are expected to produce persistent and substantial increases beyond the already very high basal rate of ATP formation (***Rich, 2003***), due to the long-half lives of the *Cerox1* sensitive complex I protein subunits (***Dörrbaum et al., 2018***; ***Mathieson et al., 2018***; ***Schwanhäusser et al., 2011***).

Conversely, after sh92-mediated *Cerox1* knockdown complex I and complex IV enzymatic activities decreased significantly (by 11%, p=0.03% and 19%, p=0.02, respectively; ***Figure 3C***), with concomitant large decreases in basal oxygen consumption (53%), ATP linked oxygen consumption (52%) and maximal uncoupled respiration (61%; p=0.011, p=0.034, p=0.042 respectively, 2-tailed Student's *t*-test; ***Figure 3D***).

These observed changes in enzymatic activity were not due to changes in mitochondria number because the enzymatic activities of complexes II, III and citrate synthase (***Figure 3A,B***), and the mitochondrial-to-nuclear genome ratio (***Figure 3—figure supplement 1A,B***), each remained unaltered

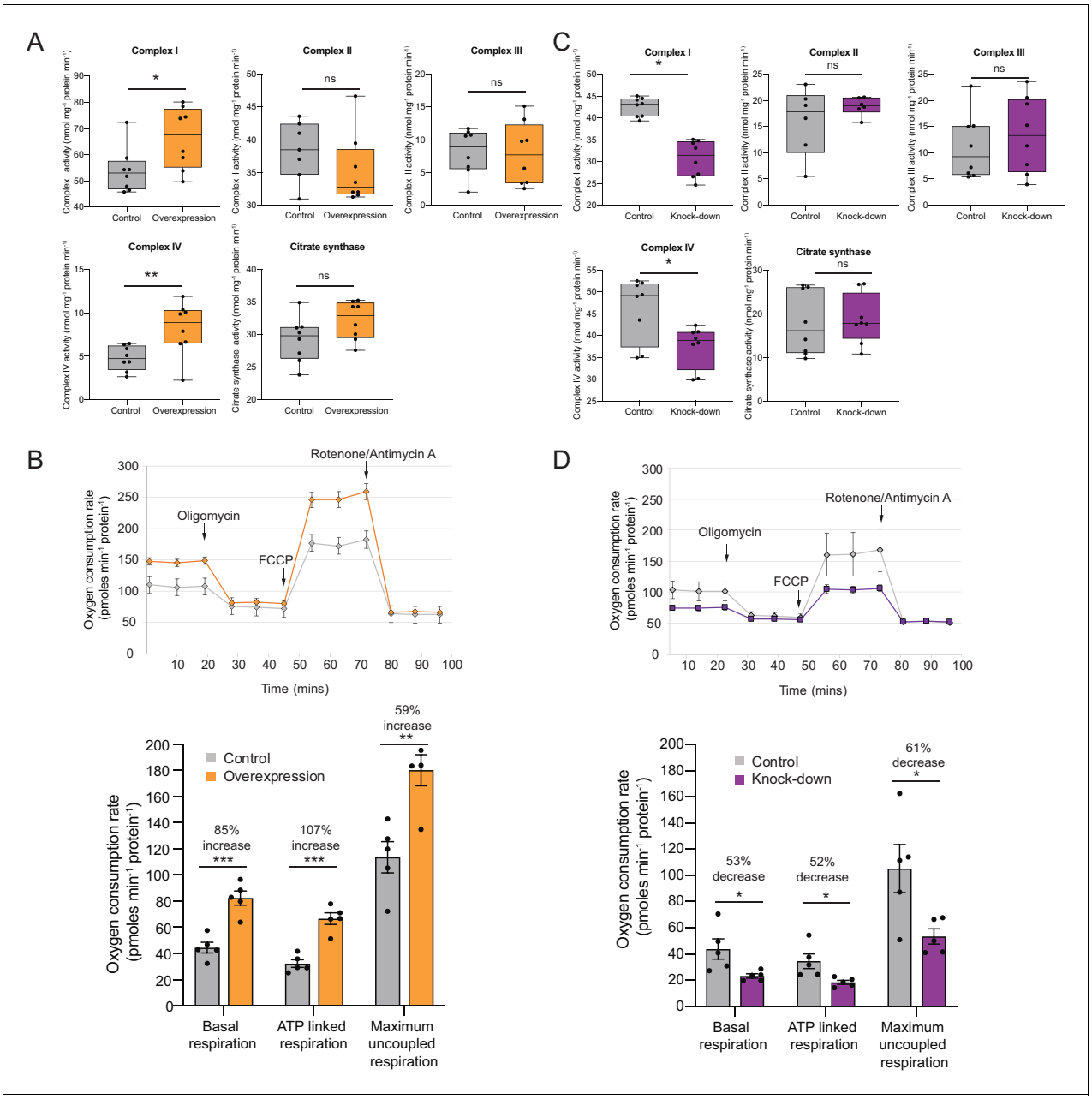


Figure 3. OXPHOS enzyme activity and oxygen consumption change concordantly and substantially with *Cerox1* level alteration. (**A**) Enzyme activities in mouse N2A cells 72 hr post-transfection of *Cerox1* overexpression construct. Mouse *Cerox1* overexpression in N2A cells results in significant increases in the catalytic activities of complexes I (22% increase) and IV (50% increase). Complexes II, III and citrate synthase show no significant change in activity. *n* = 8 biological replicates for control and overexpression. (**B**) Oxygen consumption, by N2A cells overexpressing *Cerox1*. Top: normalised real time oxygen consumption rate in basal conditions and after sequential injections of oligomycin, FCCP and rotenone/antimycin A. Bottom: changes in basal, ATP-linked and maximum uncoupled respiration respectively. Error bars s.e.m. (*n* = 5 biological replicates). (**C**) sh92-mediated knockdown of *Cerox1* results in significant decreases of Complexes I and IV enzymatic activities 72 hr post transfection; no significant changes were observed for complexes II, III or the citrate synthase control. *n* = 8 biological replicates for control and knockdown. The less effective shRNA (sh1159) also decreased oxygen consumption yet not significantly relative to the control (data not shown). (**D**) Oxygen consumption, by *Cerox1* knockdown N2A cells. Top: normalised real time oxygen consumption rate in basal conditions and after sequential injections of oligomycin, FCCP and rotenone/antimycin A.

Figure 3 continued on next page

Figure 3 continued

Bottom: changes in basal, ATP-linked and maximum uncoupled respiration respectively. Error bars s.e.m. (n = 5 biological replicates). 2-tailed Student's *t*-test: ***p<0.001, **p<0.01, *p<0.05, ns not significant.

DOI: https://doi.org/10.7554/eLife.45051.008

The following source data and figure supplement are available for figure 3:

Source data 1. N2A Cerox1 overexpression specific enzyme assays - ***Figure 3A***.

DOI: https://doi.org/10.7554/eLife.45051.010

Source data 2. N2A Cerox1 overexpression seahorse bioanalyzer - ***Figure 3B***.

DOI: https://doi.org/10.7554/eLife.45051.011

Source data 3. N2A Cerox1 knock down specific enzyme assays - ***Figure 3C***.

DOI: https://doi.org/10.7554/eLife.45051.012

Source data 4. N2A Cerox1 knock down seahorse bioanalyzer - ***Figure 3D***.

DOI: https://doi.org/10.7554/eLife.45051.013

Figure supplement 1. Increases in mitochondrial complex I and complex IV activities are not due to an increase in mitochondrial copy number.

DOI: https://doi.org/10.7554/eLife.45051.009

by changes in *Cerox1* levels. These data indicate that *Cerox1* can specifically and substantially regulate oxygen consumption and catalytic activities of complex I and complex IV in mouse N2A cells.

Cerox1 expression can protect cells from oxidative stress

Complex I deficient patient cells experience elevated ROS production (***Pitkanen and Robinson, 1996***). In *Cerox1* knockdown N2A cells ROS levels were increased significantly, by almost 20% ($p=4.2\times10^{-6}$; ***Figure 4A***). Conversely, in cells overexpressing *Cerox1*, ROS production was nearly halved ($p=3.5\times10^{-7}$; ***Figure 4A***), and protein carbonylation, a measure of ROS-induced damage, was reduced by 35% ($p=1\times10^{-3}$; ***Figure 4B***). Knock-down of *Cerox1* resulted in a 6.6% increase in protein carbonylation compared to the control (p=0.05, data not shown). The observed *Cerox1*-dependent reduction in ROS levels is of particular interest because mitochondrial complex I is a major producer of ROS which triggers cellular oxidative stress and damage, and an increase in ROS production is a common feature of mitochondrial dysfunction in disease (***Murphy, 2009***).

We next demonstrated that the increased activities of complex I and complex IV induced by *Cerox1* protect cells against the deleterious effects of specific mitochondrial complex inhibitors, specifically rotenone and sodium azide (complex I and complex IV inhibitors, 37% and 58% respectively, p<0.01); conversely, *Cerox1*-knockdown cells were significantly more sensitive to rotenone and exposure to heat (12%, p<0.001% and 11%, p<0.01 respectively ***Figure 4C***). Cells overexpressing *Cerox1* and treated with rotenone, a complex I inhibitor, exhibited no significant difference in protein carbonylation (data not shown). Taken together, these results indicate that elevation of *Cerox1* expression leads to decreased ROS production, decreased levels of oxidative damage to proteins and can confer protective effects against complex I and complex IV inhibitors.

Increased OXPHOS enzymatic activity is dependent upon miRNA binding to *Cerox1*

Due to their positive correlation in expression and cytoplasmic localisation we next considered whether *Cerox1* regulates complex I transcripts post-transcriptionally by competing with them for the binding of particular miRNAs. To address this hypothesis, we took advantage of mouse Dicer-deficient ($Dicer^{\Delta/\Delta}$) embryonic stem cells that are deficient in miRNA biogenesis (***Nesterova et al., 2008***). We first tested *Cerox1* overexpression in wildtype mouse ES cells and showed that this, again, led to an increase in transcript levels, specifically of six complex I subunits (***Figure 5A***), of which four had previously shown significant changes in N2A cells after *Cerox1* overexpression (***Figure 2C***). In contrast, overexpression in $Dicer^{\Delta/\Delta}$ cells failed to increase levels of these transcripts (***Figure 5A***). These results indicate that *Cerox1*'s ability to alter mitochondrial metabolism is miRNA-dependent.

Four miRNA families (miR-138–5p, miR-28/28–5p/708–5p, miR-370–3p, and miR-488–3p) were selected for further investigation based on the conservation of their predicted binding sites (MREs) in both mouse *Cerox1* and human *CEROX1* (***Figure 5B***). All five MREs conserved in mouse *Cerox1* and human *CEROX1* for N2A-expressed miRNAs (***Figure 5B Figure 5—figure supplement 1A***)

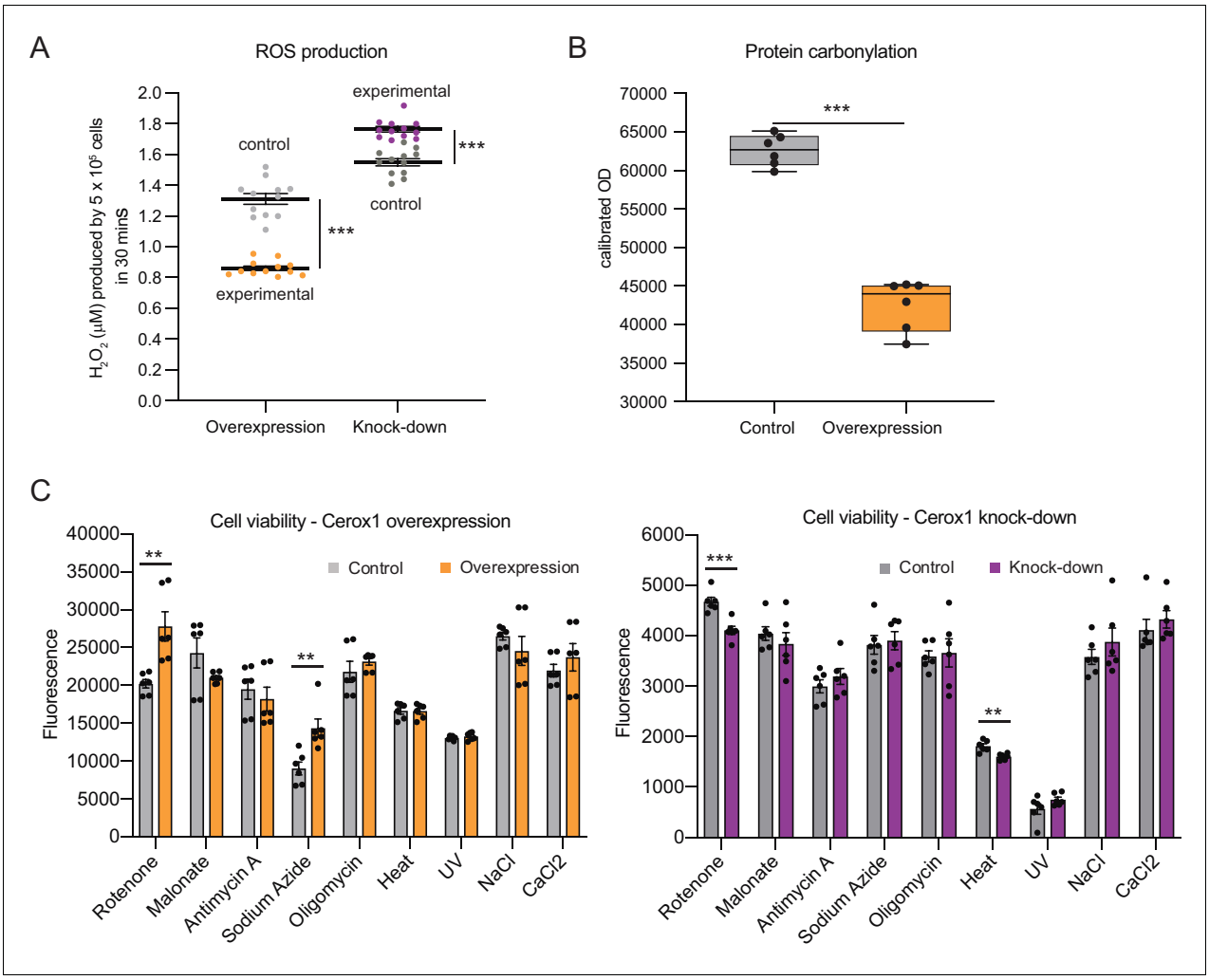


Figure 4. Cellular oxidative stress and viability depend on *Cerox1* levels. (**A**) *Cerox1* knockdown increases the production of reactive oxygen species by 20%, whilst *Cerox1* overexpression decreases it by 45% (error bars s.e.m., *n* = 12 biological replicates). (**B**) Protein oxidative damage also decreases in the overexpression condition compared to the control, as measured by densitometry on western blots against carbonylation of amino acid side chains. *n* = 6 biological replicates. (**C**) Viability of *Cerox1* overexpressing and knock-down cells when stressed. N2A cells were stressed by addition of electron transport chain (ETC) inhibitors (rotenone, CI inhibitor; malonate, competitive inhibitor of CII; antimycin A, CIII inhibitor; sodium azide, CIV inhibitor; oligomycin, ATP synthase inhibitor), exposure to environmental stress (heat, ultraviolet radiation), or manipulation of extracellular osmolarity (NaCl) or extracellular calcium ($CaCl_2$) concentration, for 1 hr and then the viability of the cells measured using the fluorescent indicator Alamar Blue. Error bars s.e.m. (*n* = 6 biological replicates for overexpression control, overexpression, knock-down control and knock-down). 2-tailed Student's *t*-test: ***p<0.001, **p<0.01.

DOI: https://doi.org/10.7554/eLife.45051.014

The following source data is available for figure 4:

Source data 1. N2A Reactive oxygen species production - ***Figure 4A***.

DOI: https://doi.org/10.7554/eLife.45051.015

Source data 2. N2A Cerox1 overexpression protein carbonylation - ***Figure 4B***.

DOI: https://doi.org/10.7554/eLife.45051.016

Source data 3. N2A cell viability Cerox1 overexpression and knockdown - ***Figure 4C***.

DOI: https://doi.org/10.7554/eLife.45051.017

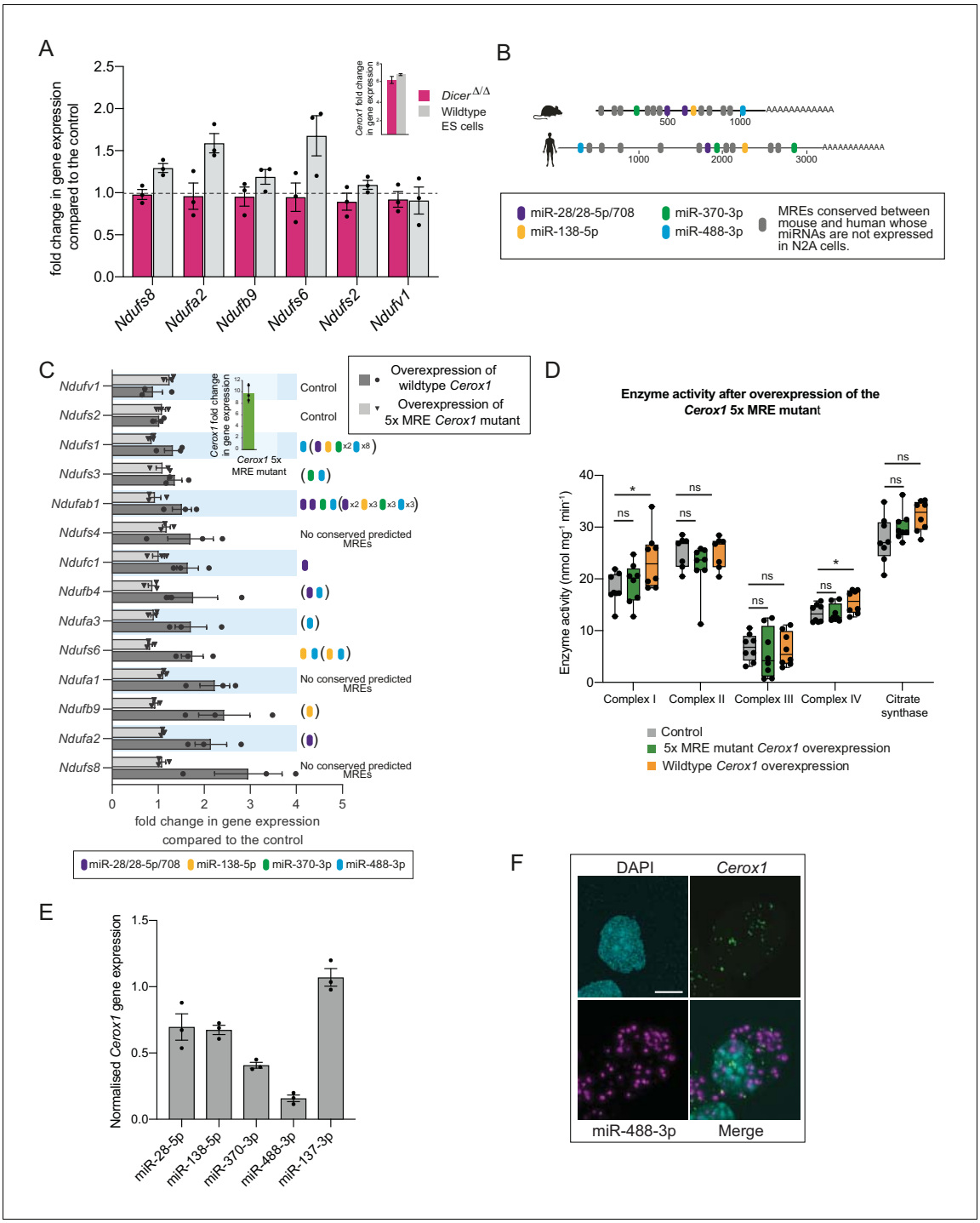


Figure 5. The effect of *Cerox1* on complex I transcript levels is miRNA-dependent. (**A**) Overexpression of *Cerox1* in mouse wildtype and *Dicer*$^{\Delta/\Delta}$embryonic stem (ES) cells (inset graph). The overexpression of *Cerox1* in wildtype mouse embryonic stem cells results in an increase in complex I subunit transcripts, with no observed change in expression of two control subunits (*Ndufs2, Ndufv1)* that were also unaffected in N2A cells. Overexpression of *Cerox1* in *Dicer*$^{\Delta/\Delta}$embryonic stem cells results in no increase in the expression of any complex I subunit. 2-sided *t*-test; **p<0.01,

Figure 5 continued on next page

Figure 5 continued

*p<0.05, ns = not significant. Error bars s.e.m. *n* = 3 biological replicates. (**B**) Predicted MREs whose presence is conserved in both the mouse and human *Cerox1*. Coloured MREs indicate those MREs whose presence is conserved between mouse and human and whose miRNAs are expressed in N2A cells. miRNA site types are as follows: miR-28–5p, 8mer-A1; miR-138–5p, 6mer; miR-370–3p, 7mer-m8; miR-488–3p, 7mer-m8; miR-708–5p, 7mer-A1. The grey predicted MREs represent those that are conserved, but whose miRNAs are not expressed in N2A cells (miR-125a-3p, miR-199/199–5p, miR-302ac/520f, miR-485/485–5p, miR-486/486–5p, miR-501/501–5p, miR-654–3p, miR-675/675–5p). (**C**) Overexpression of the 5xMRE mutant failed to alter expression of complex I subunit transcripts that otherwise all increase in abundance following wild-type *Cerox1* overexpression. Fold changes of wildtype *Cerox1* or the 5xMRE *Cerox1* mutant compared to the control are indicated in the inset panel. The numbers of MREs predicted by TargetScan v7.0 (***Agarwal et al., 2015***) in these transcripts' 3'UTRs for the four conserved, N2A expressed miRNA families are indicated (see also ***Supplementary file 3***). Due to known widespread noncanonical miRNA binding (***Helwak et al., 2013***), predictions were also extended across the gene body (bracketed MREs). 2-sided *t*-test; **p<0.01, *p<0.05, ns = not significant. Error bars s.e.m. *n* = 3 biological replicates. (**D**) Overexpression of the 5xMRE mutant failed to alter OXPHOS enzymatic activity compared to the control for any of the complexes measured. A one-way ANOVA was applied to test for differences in activities of the mitochondrial complexes between a control and overexpression of wildtype *Cerox1* and the 5xMRE mutant. A post-hoc Dunnett's test indicated that the overexpression of wildtype *Cerox1* resulted, as expected, in significantly increased complex I and IV activities of 30% and 17% respectively (F [2, 21]=4.9, p=0.017; F[2, 20]=4.6, p=0.033), while comparisons for the 5xMRE mutant with the control were not significant. There was no significant difference in the activities of complex II (F[2,19]=3.5, p=0.26), complex III (F[2,19]=0.08, p=0.5) or citrate synthase (F [2,20]=2.6, p=0.42). *n* = 6 biological replicates. Significance levels, one-way ANOVA, Dunnett's post hoc test *p<0.05. (**E**) Four to six fold overexpression of each of four miRNAs with predicted MREs whose presence is conserved in both mouse and human *Cerox1* resulted in a decrease in *Cerox1* transcript level, with overexpression of miR-488–3p resulting in >90% knock down of *Cerox1*. This was not observed when the miRNA miR-137–3p, which has no predicted MREs in *Cerox1*, was similarly overexpressed. Error bars s.e.m. *n* = 3 biological replicates. (**F**) Fluorescent in situ hybridisation detection of miR-488–3p (magenta) and *Cerox1* (green) in N2A cells. Scale bar = 5 μm. A no probe control (***Figure 5—figure supplement 1D***) indicated some background Fast Red signal (miRNA detection) localised to the nucleus, but no background for Alexa Fluor-488 (lncRNA detection).

DOI: https://doi.org/10.7554/eLife.45051.018

The following source data and figure supplement are available for figure 5:

Source data 1. N2A wildtype and MRE mutant Cerox1 overexpression specific enzyme assays - ***Figure 5D***.

DOI: https://doi.org/10.7554/eLife.45051.020

Figure supplement 1. Expression profile of mmu-miR-488-3p.

DOI: https://doi.org/10.7554/eLife.45051.019

were mutated by inversion of their seed regions. This mutated *Cerox1* transcript failed to alter either complex I transcript levels or enzyme activities when overexpressed in mouse N2A cells (***Figure 5C, D***). This indicates that *Cerox1*'s molecular effects are mediated by one or more of these MREs.

If so, then the overexpression of these miRNAs in turn would be expected to deplete *Cerox1* RNA levels. Indeed, overexpression of each miRNA reduced *Cerox1* levels (***Figure 5E***). Overexpression of the tissue-restricted miRNA miR-488–3p (***Figure 5—figure supplement 1B,C***; ***Landgraf et al., 2007***; ***Isakova and Quake, 2018***) caused the greatest depletion of the *Cerox1* transcript (***Figure 5E***) indicating that this MRE is likely to be physiologically relevant. Dual fluorescent RNA in situ hybridization of miR-488–3p and *Cerox1* indicates the proximity of these non-coding RNAs within the N2A cell (***Figure 5F***) and that both *Cerox1* and miR-488–3p are localised in the cytoplasm (***Figures 1F*** and ***5F***). *CEROX1* transcripts are predominantly (94%) localised to ribosomes (***van Heesch et al., 2014***), as is the destabilisation by microRNAs of mRNAs as they are being translated (***Tat et al., 2016***). Together, these observations imply that *Cerox1* and mitochondrial protein mRNAs are targets of miR-488–3p on ribosomes within the rough ER that forms a network around mitochondria (***Wu et al., 2017***; ***Eskelinen, 2008***).

Cerox1 activity is mediated by miR-488–3p

Our previous results showed that *Cerox1* abundance modulates complex I activity and transcripts (***Figures 2–4***) and that miR-488–3p has the greatest effect in decreasing *Cerox1* transcript levels (***Figure 5E***). To determine whether miR-488–3p modulates complex I transcript levels we overexpressed and inhibited miR-488–3p in N2A cells (***Figure 6A,B***). Results showed that miR-488–3p modulates these transcripts' levels, with overexpression leading to a significant downregulation of all 12 *Cerox1*-sensitive complex I transcripts (***Figure 6A***), whilst, conversely, miR-488–3p inhibition leads to increased expression for 10 of 12 transcripts, of which 4 (*Ndufa2, Ndufb9, Ndufs4* and *Ndufs1*) were significantly increased (***Figure 6B***).

To determine whether the single predicted miR-488–3p MRE in *Cerox1* is required to exert its effects on complex I we created a *Cerox1* transcript containing three mutated nucleotides within this

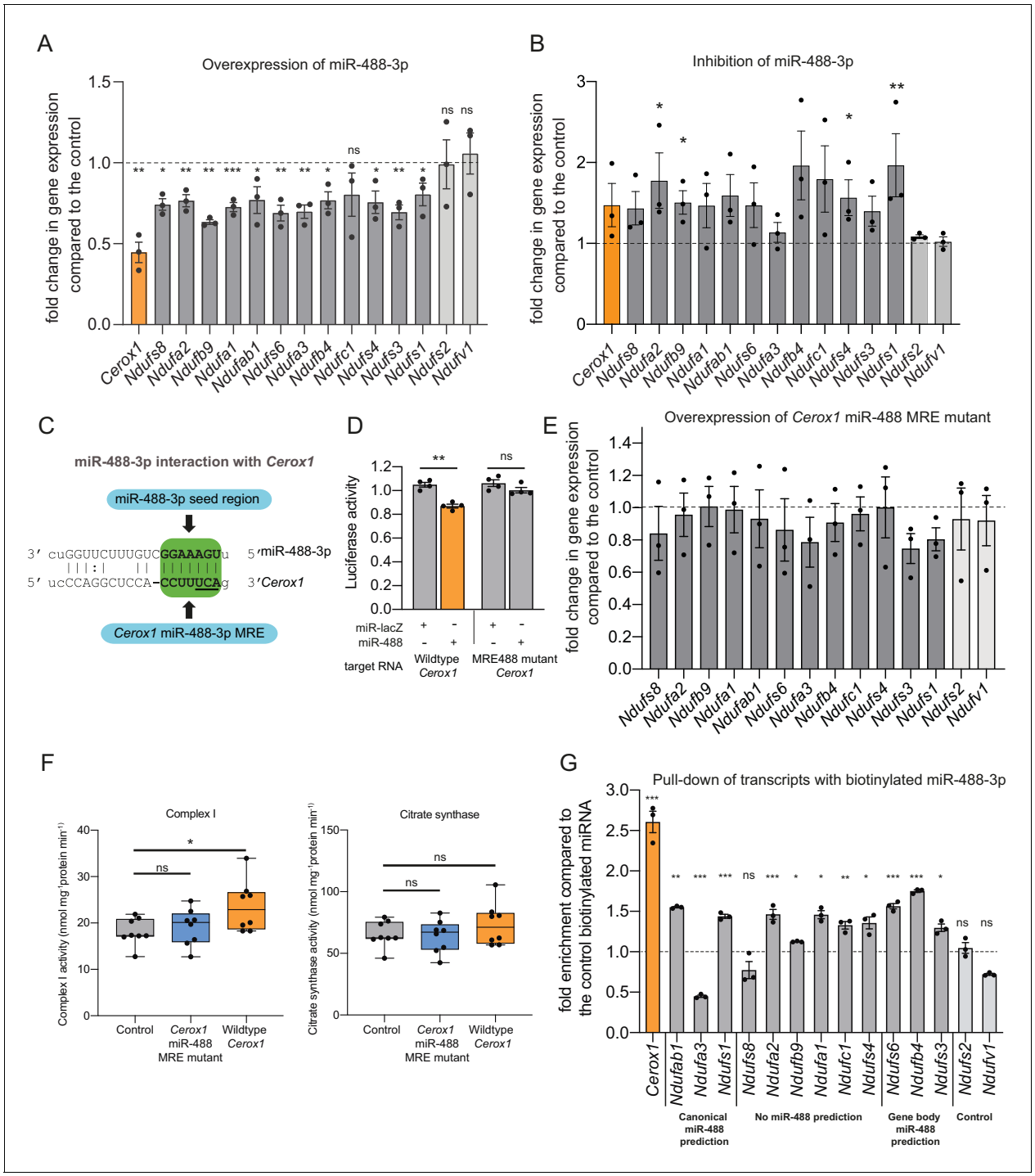


Figure 6. An intact miR-488–3p response element site is required for the effect of *Cerox1* on complex I catalytic activity. (**A**) Overexpression of miR-488–3p knocks down all *Cerox1*-sensitive subunit transcripts. Error bars s.e.m. (n = 3 biological replicates for control and overexpression of miR-488–3p). 2-sided t-test; ***p<0.001, **p<0.01, *p<0.05 (**B**) Inhibition of miR-488–3p increases the expression of most (10/12) target transcripts compared to the control. Error bars s.e.m. (n = 3 biological replicates for control and inhibition of miR-488–3p). 2-sided t-test; **p<0.01, *p<0.05. (**C**) Schematic of the

Figure 6 continued on next page

Figure 6 continued

predicted miR-488–3p miRNA recognition element in *Cerox1*. The interaction of miR-488–3p with *Cerox1* is predicted to involve a 7mer-8m seed site, with the heptamer sequence of the seed being complementary to nucleotides 2–8 of the miRNA. Underlined residues indicate the location of the *Cerox1* miR-488-3p MRE mutation. (**D**) Luciferase destabilisation assay for both wildtype *Cerox1* and *Cerox1* mutated within the miR-488–3p MRE. Error bars s.e.m. (n = 4 biological replicates for each condition). 2-sided t-test; **p<0.01. (**E, F**) Overexpression of *Cerox1* mutated within a single miR-488–3p MRE (**E**) failed to alter expression levels of complex I subunits that increase in expression with wild-type *Cerox1* overexpression (error bars s.e.m, n = 3 biological replicates) and (**F**) failed to recapitulate the increase in complex I catalytic activity observed for the wildtype transcript. Fold change of wildtype *Cerox1* compared to the control is indicated on the left of panel E. As expected, wildtype enzymatic activity was significantly different for complex I (F [2, 21]=4.944 P=0.019). A post-hoc Dunnett's test indicated that the overexpression of wildtype *Cerox1* resulted in significantly increased complex I activity, while the comparison of the *Cerox1* miR-488–3p MRE mutant with the control was not significant. There was no significant difference in the activity of citrate synthase (F[2,21]=1.4, p=0.28). n = 8 biological replicates. Significance levels, one-way ANOVA, Dunnett's post hoc test *p<0.05, ns = not significant. (**G**) Enrichment of 8 *Cerox*1 sensitive transcripts that do not have predicted canonical 3'UTR miR-488–3p MREs using biotinylated miR-488–3p as bait as compared to the control biotinylated miRNA. Error bars s.e.m. (n = 3 biological replicates). 2-sided t-test; ***p<0.001, **p<0.01, *p<0.05, ns – not significant.

DOI: https://doi.org/10.7554/eLife.45051.021

The following source data is available for figure 6:

Source data 1. N2A wildtype and miR-488–3 p mutant Cerox1 overexpression complex I and citrate synthase assays - ***Figure 6F***.

DOI: https://doi.org/10.7554/eLife.45051.022

MRE (***Figure 6C***). As expected for a bona fide MRE, these substitutions abrogated the ability of miR-488–3p to destabilise *Cerox1* transcript in a luciferase assay (***Figure 6D***). Importantly, these substitutions also abolished the ability of *Cerox1*, when overexpressed, to elevate complex I transcript levels (***Figure 6E***), and to enhance complex I enzymatic activity (***Figure 6F***). The latter observation is important because not all bona fide miRNA-transcript interactions are physiologically active (***Bassett et al., 2014***).

Finally, direct physical interaction between *Cerox1* and miR-488–3p was confirmed by pulling-down transcripts with biotinylated miR-488–3p (***Figure 6G***). This experiment also identified 10 of 12 complex I transcripts tested as direct targets of miR-488–3p binding. These included transcripts not predicted as containing a miR-488–3p MRE, as expected from the high false negative rate of MRE prediction algorithms (***Mazière and Enright, 2007***; ***Yue et al., 2009***; ***Tabas-Madrid et al., 2014***). Also as expected, the two negative control transcripts, which are not responsive to *Cerox1* transcript levels and have no predicted MREs for miR-488–3p, failed to bind miR-488–3p.

Considered together, these findings indicate that: (i) *Cerox1* can post-transcriptionally regulate OXPHOS enzymatic activity as a miRNA decoy, and (ii) of 12 miR-488–3p:*Nduf* transcript interactions that were investigated, all 12 are substantiated either by responsiveness to miR-488–3p through miRNA overexpression or inhibition (***Figure 6A,B***), or by direct interaction with a biotinylated miR-488–3p mimic (***Figure 6G***). Consequently, our data demonstrates that miR-488–3p directly regulates the transcript levels of *Cerox1* and at least 12 nuclear encoded mitochondrial complex I subunit genes (31% of all) and indirectly modulates complex I activity (***Figure 6F***) in N2A cells.

Cerox1 is an evolutionarily conserved regulator of mitochondrial complex I activity

Fewer than 20% of lncRNAs are conserved across mammalian evolution (***Necsulea et al., 2014***) and even for these functional conservation has rarely been investigated. In our final set of experiments we demonstrated that *CEROX1*, the orthologous human transcript, is functionally equivalent to mouse *Cerox1* in regulating mitochondrial complex I activity. Similar to mouse *Cerox1,* human *CEROX1* is highly expressed in brain tissue, is otherwise ubiquitously expressed (***Figure 1—figure supplement 1B***), and is enriched in the cytoplasm of human embryonic kidney (HEK293T) cells (***Figure 7A***). *CEROX1* is expressed in human tissues at unusually high levels: it occurs among the top 0.3% of all expressed lncRNAs (***Figure 7B***) and its average expression is higher than 87.5% of all protein coding genes (***GTEx Consortium, 2017***). Its expression is highest within brain tissues, particularly within the basal ganglia and cortex (***Figure 7C***).

Importantly, mitochondrial complexes' I and III activities increased significantly following *CEROX1* overexpression in HEK293T cells (***Figure 7D***). *CEROX1* overexpression had a greater effect on complex I activity than the mouse orthologous sequence and also increased the activity of

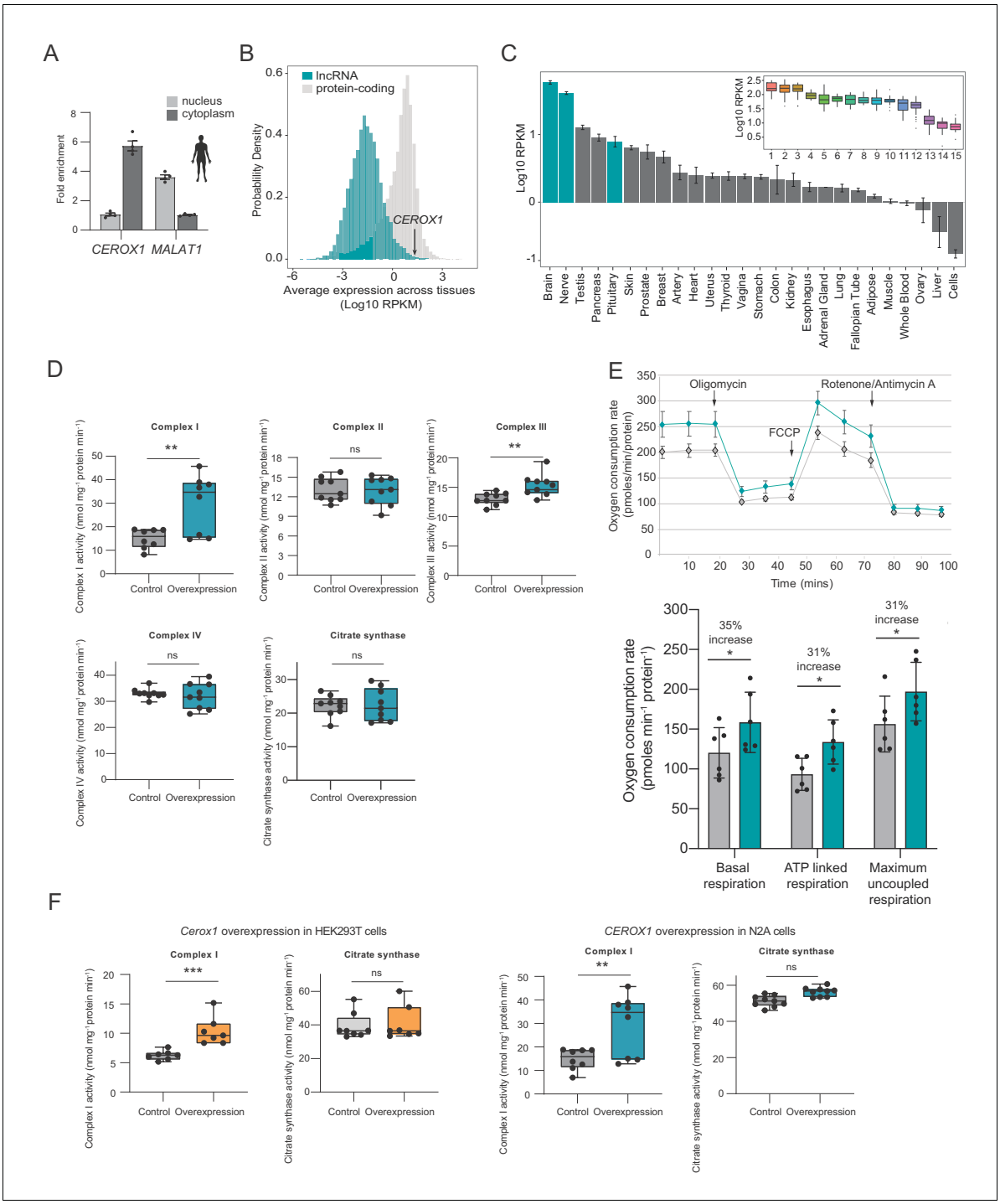


Figure 7. Human *CEROX1* modulates complex I activity in mouse cells. (**A**) *CEROX1* is enriched in the cytoplasm. Error bars s.e.m. $n = 4$ biological replicates. (**B**) Relative levels of lncRNA (blue) and protein-coding gene (grey) expression across individuals and tissues in human. The black arrow indicates the expression level of *CEROX1* in the set of 5161 lncRNAs. RPKM: reads per kilobase per million reads. (**C**) Average expression levels of *CEROX1* in human tissues. Blue bars highlight neurological tissues used to build the inset graph. The inset graphic represents the comparison of gene

Figure 7 continued on next page

Figure 7 continued

expression variation among individuals for neurological tissues: 1–Putamen, 2-Caudate nucleus, 3-Nucleus accumbens, 4–Cortex, 5-Substantia nigra, 6–Amygdala, 7–Hippocampus, 8-Spinal cord, 9-Anterior cingulate cortex, 10-Frontal cortex, 11–Hypothalamus, 12-Tibial nerve, 13–Cerebellum, 14–Pituitary gland, 15-Cerebellar hemisphere. (**D**) OXPHOS enzyme activities in human HEK293T cells after 72 hr of *CEROX1* overexpression. Overexpression of *CEROX1* results in significant increases in the activities of complexes I (31% increase) and III (18% increase), with no significant change in other enzyme activities. $n = 8$ biological replicates. 2-sided *t*-test: $p<0.01$, ns = not significant. (**E**) Oxygen consumption, as measured on a Seahorse XF^{e}24 Analyzer, by HEK293T cells overexpressing *CEROX1*. Top: normalised real time oxygen consumption rate in basal conditions and after sequential injections of oligomycin, FCCP and rotenone/antimycin A. Bottom: changes in basal, ATP-linked and maximum uncoupled respiration respectively. Error bars s.e.m. $n = 6$ biological replicates. (**F**) Reciprocal overexpression of mouse *Cerox1* in human HEK293T cells or human *CEROX1* in mouse N2A cells results in elevated complex I activity. $n = 8$ biological replicates. 2-sided *t*-test: ***$p<0.001$, **$p<0.01$, ns = not significant.

DOI: https://doi.org/10.7554/eLife.45051.023

The following source data is available for figure 7:

Source data 1. HEK293T CEROX1 overexpression specific enzyme assays - ***Figure 7D***.

DOI: https://doi.org/10.7554/eLife.45051.024

Source data 2. HEK293T CEROX1 overexpression seahorse bioanalyzer - ***Figure 7E***.

DOI: https://doi.org/10.7554/eLife.45051.025

Source data 3. Reciprocal overexpression, complex I and citrate synthase assay - ***Figure 7F***.

DOI: https://doi.org/10.7554/eLife.45051.026

complex III, rather than complex IV activity, in these cells. In addition to these observed increases in enzyme activity, basal respiration increased by 35%, ATP-linked respiration increase by 31% and maximum uncoupled respiration increased by 31% ($p=0.02$, $p=0.04$, $p=0.01$ respectively, 2-tailed Student's *t*-test; ***Figure 7E***). The latter distinction could reflect the differences in miRNA pools between mouse and human cell lines and/or the presence of different MREs in the lncRNA and human OXPHOS transcripts.

Strikingly, either reciprocal expression of mouse *Cerox1* in human HEK293T cells or human *CEROX1* in mouse N2A cells, recapitulates the previously observed increase in complex I activity (***Figure 7F***). This effect of mouse *Cerox1* overexpression in mouse N2A cells is greater than for human *CEROX1* overexpression in these cells. The role of both *Cerox1* and *CEROX1* in modulating the activity of mitochondrial complex I has thus been conserved over 90 million years since the last common ancestor of mouse and human.

Discussion

Cerox1 is the first evolutionarily conserved lncRNA to our knowledge that has been demonstrated experimentally to regulate mitochondrial energy metabolism. Its principal location in N2A cells is in the cytoplasm (***Figure 1F***) where it post-transcriptionally regulates the levels of mitochondrial OXPHOS subunit transcripts and proteins by decoying for miRNAs (***Figure 5—figure supplement 1***), most particularly miR-488–3p. This microRNA shares with *Cerox1* an early eutherian origin and elevated expression in brain samples (***Landgraf et al., 2007***), and it previously was shown to alter mitochondrial dynamics in cancer cells (***Yang et al., 2017***). Changes in *Cerox1* abundance in vitro alter mitochondrial OXPHOS subunit transcript levels and, more importantly, elicit larger changes in their protein subunits levels, leading to unexpectedly large changes in mitochondrial complex I catalytic activity. The observed changes in catalytic activity are in line with the degree of change seen in diseases exhibiting mitochondrial dysfunction (***Schapira et al., 1990***; ***Ritov et al., 2005***; ***Andreazza et al., 2010***). Overexpression of *Cerox1* in N2A cells increases oxidative metabolism, halves cellular oxidative stress and enhances protection against the complex I inhibitor rotenone. The effect of *Cerox1* on complex I subunit transcript levels can be explained by their sharing MREs with *Cerox1*, and subsequent competition for miRNA binding, most notably for miR-488–3p, which buffers the OXPHOS transcripts against miRNA-mediated repression.

Multiple RNA transcripts have been experimentally shown to compete with mRNAs for binding to miRNAs, thereby freeing the protein coding mRNA from miRNA-mediated repression (***Cesana et al., 2011***; ***Karreth et al., 2011***; ***Sumazin et al., 2011***; ***Tay et al., 2011***; ***Karreth et al., 2015***; ***Tan et al., 2014***). It has been experimentally demonstrated that this miRNA:RNA regulatory crosstalk can initiate rapid co-ordinate modulation of transcripts whose proteins participate within

the same complex or process (**Tan et al., 2014**). Physiological relevance of this crosstalk mechanism remains incompletely understood. Furthermore, mathematical modelling (**Ala et al., 2013**; **Figliuzzi et al., 2013**; **Martirosyan et al., 2016**) and experimental investigation (**Bosson et al., 2014**; **Denzler et al., 2014**) of the dynamics and mechanism of endogenous transcript competition for miRNA binding have resulted in contrasting conclusions. Current mathematical models do not take full account of miRNA properties, such as the repressive effect not being predictable from its cellular abundance (**Mullokandov et al., 2012**), intracellular localisation such as at the rough ER (**Stalder et al., 2013**), loading on the RNA-induced silencing complex (RISC) (**Flores et al., 2014**), or AGO2's phosphorylation status within the RISC (**Golden et al., 2017**). The conclusions of experiments have also assumed that all miRNA target transcripts that contain the same number and affinity of miRNA binding sites are equivalent, that steady-state measurements are relevant to repression dynamics, and that observations for one miRNA in one experimental system are equally applicable to all others (**Smillie et al., 2018**).

Considered together, our lines of experimental evidence indicate that miRNA-mediated target competition by *Cerox1* substantially perturbs a post-transcriptional gene regulatory network that includes at least 12 complex I subunit transcripts. This is consistent with the expression level of miR-488–3p (**Kozomara and Griffiths-Jones, 2014**) and the high in vivo expression of both *Cerox1* and OXPHOS transcripts (**Schwanhäusser et al., 2011**; **Vogel et al., 2010**). Human *CEROX1* levels, for example, exceed those of all complex I subunit transcripts (those in **Figure 6a**) in both newly-formed and myelinating oligodendrocytes (**Forrest et al., 2014**). *Cerox1* could maintain OXPHOS homeostasis in cells with sustained high metabolic activity and high energy requirements. Such cells occur in the central nervous system, in which *Cerox1* levels are high (**Sokoloff, 1977**), and in haematopoiesis where depletion of *Cerox1* results in B cell depletion or myeloid enrichment (**Delás et al., 2019**).

Our experiments demonstrate that post-transcriptional regulation of a subset of complex I subunits by *Cerox1* leads to elevated oxygen consumption. How consumption increases when there is a higher abundance of only a subset of OXPHOS transcripts remains unclear. However, this phenomenon has been observed previously in mouse dopaminergic neurons (**Alvarez-Fischer et al., 2011**) and primary mouse embryonic fibroblasts and pinnal tissues (**Shyh-Chang et al., 2013**). Our observation of increased enzymatic activity may relate to the formation, by the complexes of the respiratory chain, of higher order supercomplexes (**Schägger and Pfeiffer, 2000**; **Genova and Lenaz, 2014**). Alternatively, the observed increases in OXPHOS activity may reflect some subunits of the complex I holo-enzyme (including NDUFS3 and NDUFA2) being present as a monomer pool and therefore being available for direct exchange without integration into assembly intermediates (**Dieteren et al., 2012**). This monomer pool facilitates the rapid swapping out of oxidatively damaged complex I subunits (**Dieteren et al., 2012**). It is thus possible that *Cerox1*-mediated expansion of the monomer pool thereby improves complex I catalysis efficiency (**Figure 8**). However, we note that overexpression of *Cerox1* results in the differential expression of 286 genes, most (83%) of which are upregulated. These genes' transcripts will be both the targets of miR-488–3p decoying by *Cerox1* (**Figure 6**) and those whose upregulation is secondary to *Cerox1* (and miR-488–3p) mediated effects, for example relating to the observed changes in cellular metabolism and proliferation (**Figure 2—figure supplement 1I,J,K**).

More efficient ETC enzymatic activity might be relevant to mitochondrial dysfunction, a feature of many disorders that often manifests as decreases in the catalytic activities of particular mitochondrial complexes. A decrease in catalytic activity can result in elevated ROS production, leading to oxidative damage of lipids, DNA, and proteins, with OXPHOS complexes themselves being particularly susceptible to such damage (**Musatov and Robinson, 2012**). Parkinson's and Alzheimer's diseases both feature pathophysiology associated with oxidative damage resulting from increased ROS production and a cellular energy deficit associated with decreased complex I and IV activities (a reduction of 30% and 40%, respectively) (**Schapira et al., 1990**; **Keeney et al., 2006**; **Canevari et al., 1999**). A deficiency in complexes II, III and to a lesser extent complex IV, has also been described in Huntington disease (**Mochel and Haller, 2011**). Currently no effective treatments exist that help to restore mitochondrial function despite demonstration that a 20% increase in complex I activity protects mouse midbrain dopaminergic neurons against MPP+, a complex I inhibitor and a chemical model of Parkinson's disease (**Alvarez-Fischer et al., 2011**). We note that the highest expression of *CEROX1* occurs primarily in the basal ganglia (**Figure 7C** inset) regions of which are specifically vulnerable to the progressive neurological disorders Parkinson's (substantia nigra pars compacta) and

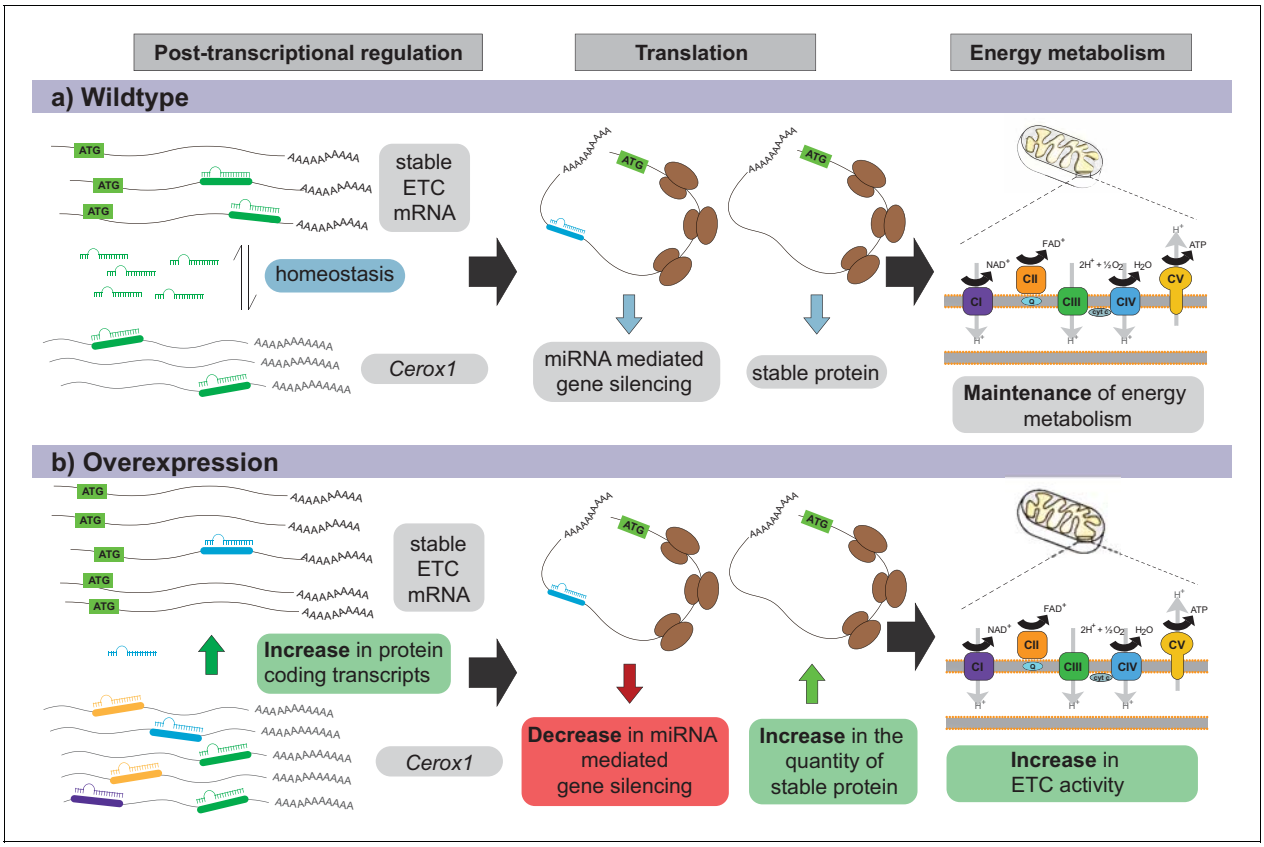


Figure 8. Proposed model for *Cerox1* as a post-transcriptional regulator of mitochondrial protein production and energy metabolism. In this model, *Cerox1* (**A**) post-transcriptionally maintains energy metabolism homeostasis through buffering the stable ETC transcripts against miRNA-mediated gene silencing. Overexpression of *Cerox1* (**B**) leads to a depletion of the pool of miRNAs that bind ETC transcripts, and therefore a decrease in miRNA mediated gene silencing of the ETC protein-coding transcripts. This has two subsequent effects: (1) a further accumulation of ETC protein coding transcripts, and (2) an increase in the overall translation of ETC subunit proteins owing to decreased miRNA binding to ETC transcripts. More rapid replenishment by undamaged subunits in mitochondrial complex I, leads to increased efficiency of complex I activity and hence an increase in overall oxygen consumption of the ETC.

DOI: https://doi.org/10.7554/eLife.45051.027

Huntington's diseases (striatum: caudate and putamen). The specific energy demands of these neurons may make them particularly susceptible to damage due to an energy deficit. For instance, the dopaminergic neurons of the substantia nigra, which are especially sensitive to degeneration in Parkinson's disease, have unusually large axonal arbours that require tight regulation of cellular energy to maintain (***Bolam and Pissadaki, 2012***). In addition, the medium spiny neurons of the striatum, which preferentially degenerate in Huntington disease, exhibit a high degree of axonal collateralization – a morphological trait which implies high cellular energy consumption for its maintenance (***Parent and Parent, 2006***) – therefore causing these cells to be vulnerable to decreased cellular ATP production. *CEROX1*'s ability to increase mitochondrial complex I activity might be recapitulated pharmacologically to restore mitochondrial function, as an exemplar of therapeutic upregulation of gene expression (***Wahlestedt, 2013***).

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Gene (*Mus musculus*)	*Cerox1*	NA	AK079380; 2810468N07Rik; ENSMUST00000163493	
Gene (*Homo sapiens*)	*CEROX1*	NA	BC098409; RP11-161M6.2; ENST00000562570	
Cell line (*Mus musculus*)	N2A (mouse neuroblastoma cells)	European Collection of Authenicated Cell Cultures (ECACC)	RRID: CVCL_0470; ECACC: 89121404	
Cell line (*Homo sapiens*)	HEK293T (human embryonic kidney cells)	European Collection of Authenicated Cell Cultures (ECACC)	RRID: CVCL_0063; ECACC 12022001	
Transfected construct (*Mus musculus* and *Homo sapiens*)	pCAG-GFP	https://www.addgene.org/89684/	RRID:Addgene_89684	This backbone was modified for overexpression of *Cerox1*, *Cerox1* MRE mutants and *CEROX1*
Transfected construct (*Mus musculus*)	BLOCK-it U6 shRNA expression construct	Invitrogen	K494500	
Transfected construct (*Mus musculus*)	BLOCK-iT Pol II miR RNAi expression vector	Invitrogen	K493600	
Antibody	anti-NDUFS1, rabbit monoclonal	Abcam	RRID: AB_2687932; ab169540	Overnight, four degrees (1:30,000)
	anti-NDUFS3, mouse monoclonal	Abcam	RRID:AB_10861972; ab110246	Overnight, four degrees, 0.15 mg/ml
	anti-alpha tubulin, mouse monoclonal	Abcam	RRID:AB_2241126; ab7291	Overnight, four degrees (1:30,000)
	goat anti-rabbit HRP, goat polyclonal	Invitrogen	RRID:AB_2536530; G-21234	Room temperature, 1 hr (1:30,000)
	goat anti-mouse HRP, goat polyclonal	Dako	RRID:AB_2617137; P0447	Room temperature, 1 hr (1:3,000)
Recombinant DNA reagent	Fugene6	Promega	E2691	
Commercial assay or kit	Amplex Red Hydrogen Peroxide/Peroxidase Assay Kit	Thermofisher	A22188	
	OxyBlot protein oxidation detetion kit	Merck Millipor	S7150	
	QuantiGene ViewRNA miRNA ISH cell assay kit	Affymetrix	QVCM0001	
Chemical compound, drug	Rotenone	Sigma Aldrich	R8875-5G	
	Antimycin A	Sigma Aldrich	A8674-25MG	
	Sodium Azide	Sigma Aldrich	S8032-25G	
	Oligomycin	Sigma Aldrich	75351–5 MG	
	NADH (reduced)	Sigma Aldrich	N8129-1G	
	coenzyme Q	Sigma Aldrich	C7956-2mg	
	Fatty acid free albumin	Sigma Aldrich	A8806-1G	

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
	Sodium succinate	Sigma Aldrich	S5047-100G	
	Dichloropheno lindophenol (DCPIP)	Sigma Aldrich	33125–5 G-R	
	Decylubiquinone	Sigma Aldrich	D7911-10mg	
	cytochrome c	Sigma Aldrich	C7752	
	Carbonyl cyanide 4(trifluoromethoxy) phenylhydrazone	Sigma Aldrich	C2920-10mg	
	miRNA inhibitors	Ambion	4464084	
	miRCURY LNA biotinylated miRNAs	Exiqon	339178	
Software, algorithm	TargetScan v7.0	**Agarwal et al., 2015**	DOI: 10.7554/eLife.05005	

Gene expression profiling

The lncRNA transcripts were assessed for coding potential using the coding potential calculator (**Kong et al., 2007**), PhyloCSF (**Lin et al., 2011**) and by mining proteomics and small open reading frame resources for evidence of translation (**Wilhelm et al., 2014**; **Kim et al., 2014a**; **Bazzini et al., 2014**). A lack of protein-coding potential for human *CEROX1* (also known as RP11-161M6.2, LMF1-3) is supported by a variety of computational and proteomic data summarised in LNCipedia (**Volders et al., 2015**). Expression data from somite-staged mouse embryos were acquired from ArrayExpress (E-ERAD-401 - Strand-specific RNA-seq of somite-staged second generation genotypically wild-type embryos of mixed G0 lineage from the Mouse Genetics Project/DMDD). Genome wide associations were performed on the UK Biobank data as described in **Canela-Xandri et al. (2018)** using data from up to 452,264 individuals.

Total RNA from twenty normal human tissues (adipose, bladder, brain, cervix, colon, oesophagus, heart, kidney, liver, lung, ovary, placenta, prostate, skeletal muscle, small intestine, spleen, testes, thymus, thyroid and trachea) were obtained from FirstChoice Human Total RNA Survey Panel (Invitrogen). Total RNA from twelve mouse tissues (bladder, brain, colon, heart, kidney, liver, pancreas, skeletal muscle, small intestine, stomach and testis) were obtained from Mouse Tissue Total RNA Panel (Amsbio). RNA from cell lines was extracted using the RNeasy mini kit (Qiagen) according to the manufacturer's instructions, using the optional on column DNase digest. cDNA synthesis for all samples was performed on 1 µg of total RNA using a QuantiTect Reverse Transcription kit (Qiagen) according to the manufacturer's instructions. RNA was extracted from samples used for the detection of miRNAs using the miRNeasy mini kit (Qiagen) according to the manufacturer's instructions (with on column DNase digest). All RNA samples were quantified using the 260/280 nm absorbance ratio, and RNA quality assessed using a Tapestation (Agilent). RNA samples with an RNA integrity number (RIN) >8.5 were reverse transcribed. 1 µg of total RNA from the miRNA samples were reverse transcribed using the NCode VILO miRNA cDNA synthesis kit. Expression levels were determined by real-time quantitative PCR, using SYBR Green Master Mix (Applied Biosystems) and standard cycling parameters (95˚C 10 min; 40 cycles 95˚C 15 s, 60˚C 1 min) followed by a melt curve using a StepOne thermal cycler (Applied Biosystems). All amplification reactions were performed in triplicate using gene specific primers. Multiple reference genes were assessed for lack of variability using geNorm (**Vandesompele et al., 2002**). Human expression data were normalised to *TUBA1A* and *POLR2A*, whilst mouse expression data were normalised to *Tbp* and *Polr2a*. Oligonucleotide sequences are provided in **Supplementary file 4**.

Tissue culture and flow cytometry

Mouse Neuro-2a neuroblastoma cells (N2A; RRID: CVCL_0470; ECACC 89121404) and human embryonic kidney (HEK293T; RRID: CVCL_0063; ECACC 12022001) were sourced from the European authenticated cell culture collection. HEK293T cells were confirmed by STR profiling and cell lines were tested monthly for mycoplasma contamination. Cells were grown at 37˚C in a humidified

incubator supplemented with 5% CO_2. Both cell lines were grown in Dulbecco's modified Eagle medium containing penicillin/streptomycin (100 U/ml, 100 ug/ml respectively) and 10% fetal calf serum. Cells were seeded at the following densities: six well dish, 0.3×10^6; 48 well dish, 0.2×10^4; T75 flask 2.1×10^6. We had three reasons for the choice of HEK293T cells for this experiment. First, they are of neural crest ectodermal origin (**Lin et al., 2014**) and therefore have a number of neural cell line characteristics in that they express neuronal markers (**Shaw et al., 2002**). Second, they are in use as a cell culture model for neurodegenerative diseases such as Parkinson's disease (**Falkenburger et al., 2016**; **Schlachetzki et al., 2013**). Third, HEK293T cells are a widely used cell line to interrogate human mitochondrial biochemistry, and exhibit complex I dependent respiration (**Kim et al., 2014b**). Mouse embryonic stem cells and dicer knock-out embryonic stem cells were maintained as described previously (**Nesterova et al., 2008**). Cells were counted using standard haemocytometry. For flow cytometry the cells were harvested by trypsinization, washed twice with PBS and fixed in 70% ethanol (filtered, −20˚C). The cell suspension was incubated at 4˚C for 10 min and the cells pelleted, treated with 40 µg/ml RNase A and propidium iodide (40 µg/ml) for 30 min at room temperature. Cells were analysed using a FACSCalibur (BD-Biosciences) flow cytometer.

Fluorescent in situ hybridization of miRNA and lncRNA, cellular fractionation and RNA turnover

Branched chain DNA probes to *Cerox1*, *Malat1* and mmu-miR-488–3p were sourced from Affymetrix. The protocol was preformed according to the manufacturer's instructions using the QuantiGene ViewRNA miRNA ISH cell assay kit (QVCM0001) for adherent cells. The following parameters were optimised: cells were fixed in 4% formaldehyde for 45 min and a 1:2000 dilution of the protease was optimal for the N2A cells. Cells were imaged using an Andor Dragonfly confocal inverted microscope, and images acquired using an Andor Zyla 4.2 plus camera.

Cells were fractionated into nuclear and cytoplasmic fractions in order to determine the predominant cellular localization of lncRNA transcripts. Briefly, approximately 2.8×10^6 cells were collected by trypsinization, washed three times in PBS and pelleted at 1000 g for 5 min at 4˚C. The cell pellet was resuspended in 5 volumes of lysis buffer (10 mM Tris-HCl, pH 7.5, 3 mM $MgCl_2$, 10 mM NaCl, 5 mM EGTA, 0.05% NP40, and protease inhibitors [Roche, complete mini]) and incubated on ice for 15 min. Lysed cells were then centrifuged at 2000 g for 10 min at 4˚C, and the supernatant collected as the cytoplasmic fraction. Nuclei were washed three times in nuclei wash buffer (10 mM HEPES, pH 6.8, 300 mM sucrose, 3 mM $MgCl_2$, 25 mM NaCl, 1 mM EGTA), and pelleted by centrifugation at 400 g, 1 min at 4˚C. Nuclei were extracted by resuspension of the nuclei pellet in 200 µl of nuclei wash buffer containing 0.5% Triton X-100 and 700 units/ml of DNase I and incubated on ice for 30 mins. Nucleoplasm fractions were collected by centrifugation at 17 000 g for 20 min at 4˚C. RNA was extracted as described above, and RNA samples with RIN values > 9.0 used to determine transcript localisation.

To determine the stability of the lncRNA transcripts, cells were cultured to ~50% confluency and then transcription was inhibited by the addition of 10 µg/ml actinomycin D (Sigma) in DMSO. Control cells were treated with equivalent volumes of DMSO. Transcriptional inhibition of the N2A cells was conducted for 16 hr with samples harvested at 0 hr, 30 mins, 1 hr, 2 hr, 4 hr, 8 hr and 16 hr. RNA samples for fractionation and turnover experiments were collected in Trizol (Invitrogen) and RNA purified and DNAse treated using the RNeasy mini kit (Qiagen). Reverse transcription for cellular localisation and turnover experiments was performed as described earlier.

Constructs and biotinylated miRNAs

The 5' and 3' ends of *Cerox1* and *CEROX1* were identified by 5' and 3' RACE using the GeneRacer Kit (Invitrogen) according to manufacturer's instructions. As an overexpression/transfection control the pCAG-EGFP backbone was used (RRID:Addgene_89684). The EGFP was removed from this backbone, and all full length lncRNAs were cloned into the pCAG backbone. For cloning into the pCAGs vector, PCR primers modified to contain the cloning sites BglII and XhoI sites were used to amplify the full length mouse *Cerox1*, whilst human *CEROX1* and the mouse 5x MRE mutant were synthesized by Biomatik (Cambridge, Ontario), and also contained BglII and XhoI sites at the 5' and 3' ends respectively. All other MRE mutants were produced using overlapping PCR site directed mutagenesis to mutate 3 bases of the miRNA seed region. All purified products were ligated into

the prepared backbone and then transformed by heat shock into chemically competent DH5α, and plated on selective media. All constructs were confirmed by sequencing. Short hairpin RNAs specific to the transcripts were designed using a combination of the RNAi design tool (Invitrogen) and the siRNA selection program from the Whitehead Institute (**Yuan et al., 2004**). Twelve pairs of shRNA oligos to the target genes and β-galactosidase control oligos were annealed to create double-stranded oligos and cloned into the BLOCK-iT U6 vector (Invitrogen), according to the manufacturer's instructions. miRNA expression constructs were generated and cloned into the BLOCK-iT Pol II miR RNAi expression vector (Invitrogen) according to the manufacturer's instructions. miRNA inhibitors were sourced from Ambion and used according to manufacturer's instructions.

Transfection efficiency was initially assessed by FACS, and the optimised transfection protocol used for all further assays (6:1 transfection reagent to DNA ratio). One day prior to transfection cells were either seeded in six well dishes (0.3×10^6 cells/well), or in T75 flasks (2.1×10^6 cells/flask). Twenty-four hours later cells in six well dishes were transfected with 1 µg of shRNA, miRNA or overexpression construct and their respective control constructs using FuGENE 6 (Promega) according to the manufacturer's guidelines. Cells in T75 flasks were transfected with 8 µg of experimental or control constructs. Transfected cells were grown for 48-72 hrs under standard conditions, and then harvested for either gene expression studies or biochemical characterisation. Efficacy of the overexpression and silencing constructs was determined by real-time quantitative PCR.

Transcripts for the luciferase destabilisation assays were cloned into the pmirGLO miRNA target expression vector (Promega) and assayed using the dual-luciferase reporter assay system (Promega). miRCURY LNA biotinylated miRNAs (mmu-miR-488–3p and mmu-negative control 4) were purchased from Exiqon, and direct mRNA-miRNA interactions were detected using a modified version of **Orom and Lund (2007)** and enrichment of targets was detected by qPCR.

Computational techniques

MREs were predicted using TargetScan v7.0 (**Agarwal et al., 2015**; RRID: SCR_010845) in either the 3'UTR (longest annotated UTR, ENSEMBL build 70; RRID: SCR_002344) or the full length transcript of protein coding genes, and across the entire transcript for lncRNAs. The average expression across 46 human tissues and individuals according to the Pilot one data from the GTEx Consortium (**Lonsdale et al., 2013**; RRID:SCR_013042; 98) was computed for both protein-coding genes and intergenic lncRNAs from the Ensembl release 75 annotation (**Flicek et al., 2014**). We used the normalised number of CAGE tags across 399 mouse cells and tissues from the FANTOM5 Consortium (http://fantom.gsc.riken.jp; **Kawaji et al., 2014**) as an approximation of expression levels for protein-coding genes and intergenic lncRNAs from the Ensembl release 75 annotation. If multiple promoters were associated with a gene, we selected the promoter with the highest average tag number. Conserved sequence blocks in the lncRNA sequences were identified using LALIGN (**Goujon et al., 2010**).

Microarray analysis

Microarray analysis was performed on 16 samples (four overexpression/four overexpression controls; four knock-down/four knock-down controls), and hybridizations were performed by the OXION array facility (University of Oxford). Data were analysed using the web-based Bioconductor interface, CARMAweb (**Rainer et al., 2006**). Differentially expressed genes (Bonferroni corrected p-value<0.05) were identified between mouse lncRNA overexpression and control cells using Limma from the Bioconductor package between the experimental samples and the respective controls. Microarray data are accessible through ArrayExpress, accession E-MATB-6792.

Extraction of metabolites and liquid chromatography – mass spectrometry (LC-MS)

Metabolites were extracted from 6-well plates by washing individual wells with ice-cold PBS and addition of cold extraction buffer (50% methanol, 30% acetonitrile, 20% water solution at −20˚C or lower). Extracts were clarified and stored at −80˚C until required. LC-MS was carried out using a 100 mm x 4.6 mm ZIC-pHILIC column (Merck-Millipore) using a Thermo Ultimate 3000 HPLC inline with a Q Exactive mass spectrometer. A 32 min gradient was developed over the column from 10% buffer A (20 mM ammonium carbonate), 90% buffer B (acetonitrile) to 95% buffer A, 5% buffer B. 10 µl of

metabolite extract was applied to the column equilibrated in 5% buffer A, 95% buffer B. Q Exactive data were acquired with polarity switching and standard ESI source and spectrometer settings were applied (typical scan range 75–1050). Metabolites were identified based upon m/z values and retention time matching to standards. Quantitation of metabolites was carried out using AssayR (***Wills et al., 2017***). Data were normalised using levels of 9 essential amino acids (histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, valine), and errors propagated, in order to account for cell count differences.

Western blots

Total protein was quantified using a BCA protein assay kit (Pierce). 10 µg of protein was loaded per well, and samples were separated on 12% SDS-PAGE gels in Tris-glycine running buffer (25 mM Tris, 192 mM glycine, 0.1% SDS). Proteins were then electroblotted onto PVDF membrane (40V, 3 hr) in transfer buffer (25 mM Tris-HCl, 192 mM glycine, 20% methanol), the membrane blocked in TBS-T (50 mM Tris-HCl, 150 mM NaCl, 0.1% Tween 20) with 5% non–fat milk powder for 1 hr. The membrane was incubated with primary antibodies overnight at 4˚C with the following dilutions: anti-NDUFS1 (RRID: AB_2687932; rabbit monoclonal, ab169540, 1:30,000), anti-NDUFS3 (RRID:AB_10861972; mouse monoclonal, 0.15 mg/ml, ab110246), or anti-alpha tubulin loading control (RRID: AB_2241126; mouse monoclonal, ab7291, 1:30,000). Following incubation with the primary antibodies, blots were washed 3 × 5 min, and 2 × 15 mins in TBS-T and incubated with the appropriate secondary antibody for 1 hr at room temperature: goat anti-rabbit HRP (RRID:AB_2536530; Invitrogen G-21234) 1:30,000; goat anti-mouse HRP (RRID:AB_2617137; Dako P0447) 1:3000. After secondary antibody incubations, blots were washed and proteins of interested detected using ECL prime chemiluminescent detection reagent (GE Healthcare) and the blots imaged using an ImageQuant LAS 4000 (GE Healthcare). Signals were normalised to the loading control using ImageJ (***Schneider et al., 2012***).

Oxidative phosphorylation enzyme assays and oxygen consumption

Cell lysates were prepared 48 hr post-transfection, by harvesting cells by trypsinisation, washing three times in ice cold phosphate buffered saline followed by centrifugation to pellet the cells (two mins, 1000 g). Cell pellets were resuspended to homogeneity in KME buffer (100 mM KCl, 50 mM MOPS, 0.5 mM EGTA, pH 7.4) and protein concentrations were determined using a BCA protein assay detection kit (Pierce). Cell lysates were flash frozen in liquid nitrogen, and freeze-thawed three times prior to assay. 300–500 µg of cell lysate was added per assay, and assays were normalised to the total amount of protein added.

All assays were performed using a Shimadzu UV-1800 spectrophotometer, absorbance readings were taken every second and all samples were measured in duplicate. Activity of complex I (CI, NADH:ubiquinone oxidoreductase) was determined by measuring the oxidation of NADH to NAD^+ at 340 nm at 30˚C in an assay mixture containing 25 mM potassium phosphate buffer (pH 7.2), 5 mM $MgCl_2$, 2.5 mg/ml fatty acid free albumin, 0.13 mM NADH, 65 µM coenzyme Q and 2 µg/ml antimycin A. The decrease in absorbance was measured for 3 mins, after which rotenone was added to a final concentration of 10 µM and the absorbance measured for a further 2 mins. The specific complex I rate was calculated as the rotenone-sensitive rate minus the rotenone-insensitive rate. Complex II (CII, succinate dehydrogenase) activity was determined by measuring the oxidation of DCPIP at 600 nm at 30˚C. Lysates were added to an assay mixture containing 25 mM potassium phosphate buffer (pH 7.2) and 2 mM sodium succinate and incubated at 30˚C for 10 mins, after which the following components were added, 2 µg/ml antimycin A, 2 µg/ml rotenone, 50 µM DCPIP and the decrease in absorbance was measured for 2 mins. Complex III (CIII, Ubiquinol:cytochrome *c* oxidoreductase) activity was determined by measuring the oxidation of decylubiquinol, with cytochrome *c* as the electron acceptor at 550 nm. The assay cuvettes contained 25 mM potassium phosphate buffer (pH 7.2), 3 mM sodium azide, 10 mM rotenone and 50 µM oxidized cytochrome *c*. Decylubiquinol was synthesized by acidifying decylubiquinone (10 mM) with HCl (6M) and reducing the quinine with sodium borohydride. After the addition of 35 µM decylubiquinol, the increase in absorbance was measured for 2 mins. Activity of Complex IV (CIV, cytochrome *c* oxidase) was measured by monitoring the oxidation of cytochrome *c* at 550 nm, 30˚C for 3 min. A 0.83 mM solution of reduced cytochrome *c* was prepared by dissolving 100 mg of cytochrome *c* in 10 ml of potassium

phosphate buffer, and adding sodium ascorbate to a final concentration of 5 mM. The resulting solution was added into SnakeSkin dialysis tubing (7 kDa molecular weight cutoff, Thermo Scientific) and dialyzed against potassium phosphate buffer, with three changes, at 4˚C for 24 hr. The redox state of the cytochrome *c* was assessed by evaluating the absorbance spectra from 500 to 600 nm. The assay buffer contained 25 mM potassium phosphate buffer (pH 7.0) and 50 µM reduced cytochrome *c*. The decrease in absorbance at 550 nm was recorded for 3 mins. As a control the enzymatic activity of the tricarboxylic acid cycle enzyme, citrate synthase (CS) was assayed at 412 nm at 30˚C in a buffer containing 100 mM Tris-HCl (pH 8.0), 100 µM DTNB (5,5-dithiobis[2-nitrobenzoic acid]), 50 µM acetyl coenzyme A, 0.1% (w/v) Triton X-100 and 250 µM oxaloacetate. The increase in absorbance was monitored for 2 mins.

The following extinction coefficients were applied: complex I (CI), ε = 6220 M^{-1} cm^{-1}, CII, ε = 21,000 M^{-1} cm^{-1}; CIII, ε = 19,100 M^{-1} cm^{-1}; CIV, ε = 21,840 M^{-1} cm^{-1} (the difference between reduced and oxidised cytochrome *c* at 550 nm); CS, ε = 13,600 mM^{-1} cm^{-1}.

Cellular oxygen consumption rate was determined using the Seahorse XF^e24 Analyzer (Agilent). Cells were plated on poly-L-lysine coated XF^e24 microplates at 50,000 cells per well and incubated at 37˚C and 5% CO_2 for 24 hr. Cells were transfected with an overexpression (pCAG-*Cerox1* or pCAG-*CEROX1*) or the *Cerox1* shRNA silencing construct and assayed 24–36 hr later. Cells were washed three times with Seahorse Assay media (Agilent), supplemented with 10 mM glucose and 2 mM pyruvate. After 30 min of pre-incubation in a non-CO_2 37˚C incubator, cells were entered into the analyser for oxygen consumption rate measurements. After basal respiration measurements, 1 µM of oligomycin was injected to inhibit ATP synthase, then 0.2 µM of carbonyl cyanide 4-(trifluoromethoxy)phenylhydrazone (FCCP) was injected to uncouple respiration. Finally 1 µM each of rotenone and antimycin A were injected to inhibit complex I and complex III respectively. Data was normalised to total cellular protein content using the sulforhodamine B assay (***Skehan et al., 1990***). Basal respiration, ATP linked respiration and maximum uncoupled respiration were calculated from the normalised data using the Agilent Seahorse XF calculation guidelines. Briefly, basal respiration = measurement 3 (last rate measurement before oligomycin injection) – measurement 10 (minimum respiration after injection of rotenone/antimycin A); ATP-linked respiration = measurement 3 – measurement 4 (minimum rate measurement after oligomycin injection); maximum uncoupled respiration = measurement 7 (maximum rate after FCCP injection) – measurement 10.

Oxidative stress measurements

Hydrogen peroxide production was assessed as a marker of reactive oxygen species generation using the fluorescent indicator Amplex Red (10 µM, Invitrogen) in combination with horseradish peroxidise (0.1 units ml^{-1}). Total amount of H_2O_2 produced was normalised to mg of protein added. Protein carbonylation was assessed using the OxyBlot protein oxidation detection kit (Merck Millipore), and differential carbonylation was assessed by densitometry. The cell stress assay was performed on cells seeded in 48 well plates, and assayed 12 hr later by the addition of (final concentration): rotenone (5 µM), malonate (40 µM), antimycin A (500 µM), oligomycin (500 µM), sodium azide (3 mM), NaCl (300 mM), $CaCl_2$ (5.4 mM) for 1 hr. Cells were heat shocked at 42˚C and UV irradiated using a Stratlinker UV Crosslinker for 10 min (2.4 J cm^{-2}). Cell viability was assessed by the addition of Alamar Blue (Invitrogen) according to the manufacturer's instructions.

Acknowledgements

This work was funded by the Wellcome Trust (106956/Z/15/Z; CPP, TMS, OBR, SRG), a European Research Council Advanced Grant (DARCGENs; CPP, ACM, TMS, KR), the Medical Research Council (CPP, WH), a Marie Curie Intra-European Career Development Award (ACM), the University of Oxford (ACM), the Royal Society (ACM), the Clarendon Fund (JYT), the Natural Sciences Engineering Research Council of Canada (JYT), and Diabetes UK (LCH, 11/0004175). RNC was funded by a Wellcome Trust Investigator award (WT100981/z/13/z) to NMM. Microarray hybridizations were performed by the OXION array facility. FISH confocal images were acquired by Matt Pearson (IGMM advanced imaging support team). We would like to thank Ava Khamseh (IGMM) for normalisation of the metabolomics data and Andrew Dodd for critical reading of the manuscript.

Additional information

Competing interests

Chris P Ponting: Reviewing editor, *eLife*. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Wellcome	WT106956/Z/15/Z	Tamara M Sirey Oscar Bedoya-Reina Sebastian Rogatti-Granados Chris P Ponting
European Research Council	249869	Tamara M Sirey Kenny Roberts Ana Claudia Marques Chris P Ponting
Medical Research Council		Wilfried Haerty Chris P Ponting
Oxford University	The Clarendon Fund	Jennifer Y Tan
Natural Sciences and Engineering Research Council of Canada		Jennifer Y Tan
Diabetes UK	11/0004175	Lisa C Heather
Wellcome	WT100981/z/13/z	Roderick N Carter Nicholas M Morton
European Commission	Marie Curie Intra-European Career Development Award	Ana Claudia Marques
Oxford University		Ana Claudia Marques
Royal Society		Ana Claudia Marques

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Tamara M Sirey, Conceptualization, Formal analysis, Supervision, Validation, Investigation, Methodology, Writing—original draft, Writing—review and editing; Kenny Roberts, Investigation, Methodology; Wilfried Haerty, Data curation, Formal analysis; Oscar Bedoya-Reina, Formal analysis; Sebastian Rogatti-Granados, Validation, Investigation; Jennifer Y Tan, Nick Li, Investigation; Lisa C Heather, Jimi Wills, Formal analysis, Methodology; Roderick N Carter, Sarah Cooper, Nicholas M Morton, Resources, Methodology; Andrew J Finch, Resources, Formal analysis, Methodology; Ana Claudia Marques, Conceptualization, Formal analysis; Chris P Ponting, Conceptualization, Resources, Supervision, Funding acquisition, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Tamara M Sirey https://orcid.org/0000-0001-5606-2858
Kenny Roberts http://orcid.org/0000-0001-6155-0821
Wilfried Haerty https://orcid.org/0000-0003-0111-191X
Sebastian Rogatti-Granados https://orcid.org/0000-0002-8438-7999
Lisa C Heather https://orcid.org/0000-0002-7246-1338
Andrew J Finch http://orcid.org/0000-0002-8065-4623
Jimi Wills https://orcid.org/0000-0003-1669-007X
Nicholas M Morton https://orcid.org/0000-0001-8218-8462

Ana Claudia Marques http://orcid.org/0000-0001-5174-8092
Chris P Ponting https://orcid.org/0000-0003-0202-7816

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.45051.042
Author response https://doi.org/10.7554/eLife.45051.043

Additional files

Supplementary files

- Supplementary file 1. Association of CEROX1 single nucleotide polymorphism on anthropomorphic traits. Data was accessed through http://geneatlas.roslin.ed.ac.uk/.
DOI: https://doi.org/10.7554/eLife.45051.028
- Supplementary file 2. Differentially expressed genes after overexpression of mouse *Cerox1*.
DOI: https://doi.org/10.7554/eLife.45051.029
- Supplementary file 3. MRE predictions for Nduf subunit transcripts.
DOI: https://doi.org/10.7554/eLife.45051.030
- Supplementary file 4. shRNA oligos and PCR primers.
DOI: https://doi.org/10.7554/eLife.45051.031
- Transparent reporting form
DOI: https://doi.org/10.7554/eLife.45051.032

Data availability

Microarray data are available through ArrayExpress, accession code E-MATB-6792.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Tamara M Sirey, Chris P Ponting	2019	Investigation of potential gene regulatory or post-transcriptional regulatory function of a long noncoding RNA (Cerox1)	https://www.ebi.ac.uk/arrayexpress/experiments/E-MATB-6792/	ArrayExpress, E-MATB-6792

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
GTEx Consortium	2013	The Genotype-Tissue Expression (GTEx) project	https://gtexportal.org/home/datasets	GTEx Portal, phs000424.v7.p2
Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, Ninomiya-Fukuda N, Ishikawa-Kato S, Nagao-Sato S, Noma S, Hayashizaki Y, Forrest ARR	2014	Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing	http://fantom.gsc.riken.jp/data/	FANTOM Consortium, FANTOM5
Canela-Xandri O, Rawlik K, Tenesa A	2018	An atlas of genetic associations in the UK Biobank.	http://geneatlas.roslin.ed.ac.uk/downloads/	GeneATLAS, downloads

References

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**:e05005. DOI: https://doi.org/10.7554/eLife.05005

Ala U, Karreth FA, Bosia C, Pagnani A, Taulli R, Léopold V, Tay Y, Provero P, Zecchina R, Pandolfi PP. 2013. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive

molecular environments. *PNAS* **110**:7154–7159. DOI: https://doi.org/10.1073/pnas.1222509110, PMID: 23536298

Alvarez-Fischer D, Fuchs J, Castagner F, Stettler O, Massiani-Beaudoin O, Moya KL, Bouillot C, Oertel WH, Lombès A, Faigle W, Joshi RL, Hartmann A, Prochiantz A. 2011. Engrailed protects mouse midbrain dopaminergic neurons against mitochondrial complex I insults. *Nature Neuroscience* **14**:1260–1266. DOI: https://doi.org/10.1038/nn.2916, PMID: 21892157

Andreazza AC, Shao L, Wang JF, Young LT. 2010. Mitochondrial complex I activity and oxidative damage to mitochondrial proteins in the prefrontal cortex of patients with bipolar disorder. *Archives of General Psychiatry* **67**:360–368. DOI: https://doi.org/10.1001/archgenpsychiatry.2010.22, PMID: 20368511

Balsa E, Marco R, Perales-Clemente E, Szklarczyk R, Calvo E, Landázuri MO, Enríquez JA. 2012. NDUFA4 is a subunit of complex IV of the mammalian electron transport chain. *Cell Metabolism* **16**:378–386. DOI: https://doi.org/10.1016/j.cmet.2012.07.015, PMID: 22902835

Bassett AR, Azzam G, Wheatley L, Tibbit C, Rajakumar T, McGowan S, Stanger N, Ewels PA, Taylor S, Ponting CP, Liu JL, Sauka-Spengler T, Fulga TA. 2014. Understanding functional miRNA-target interactions in vivo by site-specific genome engineering. *Nature Communications* **5**:4640. DOI: https://doi.org/10.1038/ncomms5640, PMID: 25135198

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, Giraldez AJ. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* **33**:981–993. DOI: https://doi.org/10.1002/embj.201488411, PMID: 24705786

Bergmann JH, Li J, Eckersley-Maslin MA, Rigo F, Freier SM, Spector DL. 2015. Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome Research* **25**:1336–1346. DOI: https://doi.org/10.1101/gr.189027.114, PMID: 26048247

Bolam JP, Pissadaki EK. 2012. Living on the edge with too many mouths to feed: why dopamine neurons die. *Movement Disorders* **27**:1478–1483. DOI: https://doi.org/10.1002/mds.25135, PMID: 23008164

Bosson AD, Zamudio JR, Sharp PA. 2014. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Molecular Cell* **56**:347–359. DOI: https://doi.org/10.1016/j.molcel.2014.09.018, PMID: 25449132

Canela-Xandri O, Rawlik K, Tenesa A. 2018. An atlas of genetic associations in UK biobank. *bioRxiv*. DOI: https://doi.org/10.1101/176834

Canevari L, Clark JB, Bates TE. 1999. beta-Amyloid fragment 25-35 selectively decreases complex IV activity in isolated mitochondria. *FEBS Letters* **457**:131–134. DOI: https://doi.org/10.1016/S0014-5793(99)01028-5, PMID: 10486579

Cao Y, Zhu J, Jia P, Zhao Z. 2017. scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells. *Genes* **8**:368. DOI: https://doi.org/10.3390/genes8120368, PMID: 29206167

Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Research* **10**:1617–1630. DOI: https://doi.org/10.1101/gr.145100, PMID: 11042159

Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, Bozzoni I. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**:358–369. DOI: https://doi.org/10.1016/j.cell.2011.09.028, PMID: 22000014

Delás MJ, Jackson BT, Kovacevic T, Vangelisti S, Munera Maravilla E, Wild SA, Stork EM, Erard N, Knott SRV, Hannon GJ. 2019. lncRNA *Spehd* regulates hematopoietic stem and progenitor cells and is required for multilineage differentiation. *Cell Reports* **27**:719–729. DOI: https://doi.org/10.1016/j.celrep.2019.03.080, PMID: 30995471

Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M. 2014. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Molecular Cell* **54**:766–776. DOI: https://doi.org/10.1016/j.molcel.2014.03.045, PMID: 24793693

Dieteren CE, Koopman WJ, Swarts HG, Peters JG, Maczuga P, van Gemst JJ, Masereeuw R, Smeitink JA, Nijtmans LG, Willems PH. 2012. Subunit-specific incorporation efficiency and kinetics in mitochondrial complex I homeostasis. *Journal of Biological Chemistry* **287**:41851–41860. DOI: https://doi.org/10.1074/jbc.M112.391151, PMID: 23038253

Dörrbaum AR, Kochen L, Langer JD, Schuman EM. 2018. Local and global influences on protein turnover in neurons and glia. *eLife* **7**:e34202. DOI: https://doi.org/10.7554/eLife.34202, PMID: 29914620

Dumortier O, Hinault C, Van Obberghen E. 2013. MicroRNAs and metabolism crosstalk in energy homeostasis. *Cell Metabolism* **18**:312–324. DOI: https://doi.org/10.1016/j.cmet.2013.06.004, PMID: 23850315

Eskelinen EL. 2008. To be or not to be? examples of incorrect identification of autophagic compartments in conventional transmission electron microscopy of mammalian cells. *Autophagy* **4**:257–260. DOI: https://doi.org/10.4161/auto.5179, PMID: 17986849

Falkenburger BH, Saridaki T, Dinter E. 2016. Cellular models for parkinson's disease. *Journal of Neurochemistry* **139**:121–130. DOI: https://doi.org/10.1111/jnc.13618, PMID: 27091001

Figliuzzi M, Marinari E, De Martino A. 2013. MicroRNAs as a selective channel of communication between competing RNAs: a steady-state theory. *Biophysical Journal* **104**:1203–1213. DOI: https://doi.org/10.1016/j.bpj.2013.01.012, PMID: 23473503

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kähäri AK, Keenan S, Kulesha E, et al.

2014. Ensembl 2014. *Nucleic Acids Research* **42**:D749–D755. DOI: https://doi.org/10.1093/nar/gkt1196, PMID: 24316576

Flores O, Kennedy EM, Skalsky RL, Cullen BR. 2014. Differential RISC association of endogenous human microRNAs predicts their inhibitory potential. *Nucleic Acids Research* **42**:4629–4639. DOI: https://doi.org/10.1093/nar/gkt1393, PMID: 24464996

Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**:462–470. DOI: https://doi.org/10.1038/nature13182, PMID: 24670764

Friedel CC, Dölken L, Ruzsics Z, Koszinowski UH, Zimmer R. 2009. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Research* **37**:e115. DOI: https://doi.org/10.1093/nar/gkp542, PMID: 19561200

Genova ML, Lenaz G. 2014. Functional role of mitochondrial respiratory supercomplexes. *Biochimica Et Biophysica Acta (BBA) - Bioenergetics* **1837**:427–443. DOI: https://doi.org/10.1016/j.bbabio.2013.11.002

Golden RJ, Chen B, Li T, Braun J, Manjunath H, Chen X, Wu J, Schmid V, Chang TC, Kopp F, Ramirez-Martinez A, Tagliabracci VS, Chen ZJ, Xie Y, Mendell JT. 2017. An argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* **542**:197–202. DOI: https://doi.org/10.1038/nature21025, PMID: 28114302

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research* **38**:W695–W699. DOI: https://doi.org/10.1093/nar/gkq313, PMID: 20439314

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**:204–213. DOI: https://doi.org/10.1038/nature24277, PMID: 29022597

Guerrero-Castillo S, Baertling F, Kownatzki D, Wessels HJ, Arnold S, Brandt U, Nijtmans L. 2017. The assembly pathway of mitochondrial respiratory chain complex I. *Cell Metabolism* **25**:128–139. DOI: https://doi.org/10.1016/j.cmet.2016.09.002, PMID: 27720676

Helwak A, Kudla G, Dudnakova T, Tollervey D. 2013. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**:654–665. DOI: https://doi.org/10.1016/j.cell.2013.03.043, PMID: 23622248

Isakova A, Quake S. 2018. A mouse tissue atlas of small non-coding RNA. *bioRxiv*. DOI: https://doi.org/10.1101/430561

Karreth FA, Tay Y, Perna D, Ala U, Tan SM, Rust AG, DeNicola G, Webster KA, Weiss D, Perez-Mancera PA, Krauthammer M, Halaban R, Provero P, Adams DJ, Tuveson DA, Pandolfi PP. 2011. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* **147**:382–395. DOI: https://doi.org/10.1016/j.cell.2011.09.032, PMID: 22000016

Karreth FA, Reschke M, Ruocco A, Ng C, Chapuy B, Léopold V, Sjoberg M, Keane TM, Verma A, Ala U, Tay Y, Wu D, Seitzer N, Velasco-Herrera MC, Bothmer A, Fung J, Langellotto F, Rodig SJ, Elemento O, Shipp MA, et al. 2015. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell* **161**:319–332. DOI: https://doi.org/10.1016/j.cell.2015.02.043, PMID: 25843629

Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, Ninomiya-Fukuda N, Ishikawa-Kato S, Nagao-Sato S, Noma S, Hayashizaki Y, Forrest AR, Carninci P, FANTOM Consortium. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Research* **24**:708–717. DOI: https://doi.org/10.1101/gr.156232.113, PMID: 24676093

Keeney PM, Xie J, Capaldi RA, Bennett JP. 2006. Parkinson's disease brain mitochondrial complex I has oxidatively damaged subunits and is functionally impaired and misassembled. *Journal of Neuroscience* **26**:5256–5264. DOI: https://doi.org/10.1523/JNEUROSCI.0984-06.2006, PMID: 16687518

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, et al. 2014a. A draft map of the human proteome. *Nature* **509**:575–581. DOI: https://doi.org/10.1038/nature13302, PMID: 24870542

Kim C, Patel P, Gouvin LM, Brown ML, Khalil A, Henchey EM, Heuck AP, Yadava N. 2014b. Comparative analysis of the mitochondrial physiology of pancreatic β cells. *Bioenergetics : Open Access* **3**:110. DOI: https://doi.org/10.4172/2167-7662.1000110, PMID: 25309834

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* **35**:W345–W349. DOI: https://doi.org/10.1093/nar/gkm391, PMID: 17631615

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42**:D68–D73. DOI: https://doi.org/10.1093/nar/gkt1181, PMID: 24275495

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller RU, et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**:1401–1414. DOI: https://doi.org/10.1016/j.cell.2007.04.040, PMID: 17604727

Lazarou M, Thorburn DR, Ryan MT, Mckenzie M. 1793. Assembly of mitochondrial complex I and defects in disease. *Biochimica Et Biophysica Acta* **2009**:78–88. DOI: https://doi.org/10.1016/j.bbamcr.2008.04.015

Leucci E, Vendramin R, Spinazzi M, Laurette P, Fiers M, Wouters J, Radaelli E, Eyckerman S, Leonelli C, Vanderheyden K, Rogiers A, Hermans E, Baatsen P, Aerts S, Amant F, Van Aelst S, van den Oord J, de

Strooper B, Davidson I, Lafontaine DL, et al. 2016. Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**:518–522. DOI: https://doi.org/10.1038/nature17161, PMID: 27008969

Li P, Jiao J, Gao G, Prabhakar BS. 2012. Control of mitochondrial activity by miRNAs. *Journal of Cellular Biochemistry* **113**:1104–1110. DOI: https://doi.org/10.1002/jcb.24004, PMID: 22135235

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**:i275–i282. DOI: https://doi.org/10.1093/bioinformatics/btr209, PMID: 21685081

Lin YC, Boone M, Meuris L, Lemmens I, Van Roy N, Soete A, Reumers J, Moisse M, Plaisance S, Drmanac R, Chen J, Speleman F, Lambrechts D, Van de Peer Y, Tavernier J, Callewaert N. 2014. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature Communications* **5**:4767. DOI: https://doi.org/10.1038/ncomms5767, PMID: 25182477

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**:580–585. DOI: https://doi.org/10.1038/ng.2653, PMID: 23715323

Martirosyan A, Figliuzzi M, Marinari E, De Martino A. 2016. Probing the limits to MicroRNA-Mediated control of gene expression. *PLOS Computational Biology* **12**:e1004715. DOI: https://doi.org/10.1371/journal.pcbi.1004715, PMID: 26812364

Mathieson T, Franken H, Kosinski J, Kurzawa N, Zinn N, Sweetman G, Poeckel D, Ratnu VS, Schramm M, Becher I, Steidel M, Noh KM, Bergamini G, Beck M, Bantscheff M, Savitski MM. 2018. Systematic analysis of protein turnover in primary cells. *Nature Communications* **9**:689. DOI: https://doi.org/10.1038/s41467-018-03106-1, PMID: 29449567

Matsumoto S, Uchiumi T, Saito T, Yagi M, Takazaki S, Kanki T, Kang D. 2012. Localization of mRNAs encoding human mitochondrial oxidative phosphorylation proteins. *Mitochondrion* **12**:391–398. DOI: https://doi.org/10.1016/j.mito.2012.02.004, PMID: 22406259

Mazière P, Enright AJ. 2007. Prediction of microRNA targets. *Drug Discovery Today* **12**:452–458. DOI: https://doi.org/10.1016/j.drudis.2007.04.002, PMID: 17532529

Melin BS, Barnholtz-Sloan JS, Wrensch MR, Johansen C, Il'yasova D, Kinnersley B, Ostrom QT, Labreche K, Chen Y, Armstrong G, Liu Y, Eckel-Passow JE, Decker PA, Labussière M, Idbaih A, Hoang-Xuan K, Di Stefano AL, Mokhtari K, Delattre JY, Broderick P, et al. 2017. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nature Genetics* **49**:789–794. DOI: https://doi.org/10.1038/ng.3823, PMID: 28346443

Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF. 2010. Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neuroscience* **11**:14. DOI: https://doi.org/10.1186/1471-2202-11-14, PMID: 20137068

Michaud M, Maréchal-Drouard L, Duchêne AM. 2014. Targeting of cytosolic mRNA to mitochondria: naked RNA can bind to the mitochondrial surface. *Biochimie* **100**:159–166. DOI: https://doi.org/10.1016/j.biochi.2013.11.007, PMID: 24252184

Mochel F, Haller RG. 2011. Energy deficit in Huntington disease: why it matters. *Journal of Clinical Investigation* **121**:493–499. DOI: https://doi.org/10.1172/JCI45691, PMID: 21285522

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, et al. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**:267–273. DOI: https://doi.org/10.1038/ng1180, PMID: 12808457

Mullokandov G, Baccarini A, Ruzo A, Jayaprakash AD, Tung N, Israelow B, Evans MJ, Sachidanandam R, Brown BD. 2012. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nature Methods* **9**:840–846. DOI: https://doi.org/10.1038/nmeth.2078, PMID: 22751203

Murphy MP. 2009. How mitochondria produce reactive oxygen species. *Biochemical Journal* **417**:1–13. DOI: https://doi.org/10.1042/BJ20081386, PMID: 19061483

Musatov A, Robinson NC. 2012. Susceptibility of mitochondrial electron-transport complexes to oxidative damage. Focus on cytochrome c oxidase. *Free Radical Research* **46**:1313–1326. DOI: https://doi.org/10.3109/10715762.2012.717273, PMID: 22856385

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**:635–640. DOI: https://doi.org/10.1038/nature12943, PMID: 24463510

Nesterova TB, Popova BC, Cobb BS, Norton S, Senner CE, Tang YA, Spruce T, Rodriguez TA, Sado T, Merkenschlager M, Brockdorff N. 2008. Dicer regulates Xist promoter methylation in ES cells indirectly through transcriptional control of Dnmt3a. *Epigenetics & Chromatin* **1**:2. DOI: https://doi.org/10.1186/1756-8935-1-2, PMID: 19014663

Orom UA, Lund AH. 2007. Isolation of microRNA targets using biotinylated synthetic microRNAs. *Methods* **43**:162–165. DOI: https://doi.org/10.1016/j.ymeth.2007.04.007, PMID: 17889804

Parent M, Parent A. 2006. Relationship between axonal collateralization and neuronal degeneration in basal ganglia. *Journal of Neural Transmission. Supplementum* **70**:85–88. PMID: 17017513

Perales-Clemente E, Fernández-Vizarra E, Acín-Pérez R, Movilla N, Bayona-Bafaluy MP, Moreno-Loshuertos R, Pérez-Martos A, Fernández-Silva P, Enríquez JA. 2010. Five entry points of the mitochondrially encoded subunits in mammalian complex I assembly. *Molecular and Cellular Biology* **30**:3038–3047. DOI: https://doi.org/10.1128/MCB.00025-10, PMID: 20385768

Pitkanen S, Robinson BH. 1996. Mitochondrial complex I deficiency leads to increased production of superoxide radicals and induction of superoxide dismutase. *Journal of Clinical Investigation* **98**:345–351. DOI: https://doi.org/10.1172/JCI118798, PMID: 8755643

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research* **17**:556–565. DOI: https://doi.org/10.1101/gr.6036807, PMID: 17387145

Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, Trajanoski Z. 2006. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Research* **34**:W498–W503. DOI: https://doi.org/10.1093/nar/gkl038, PMID: 16845058

Rich P. 2003. Chemiosmotic coupling: The cost of living. *Nature* **421**:583. DOI: https://doi.org/10.1038/421583a, PMID: 12571574

Ritov VB, Menshikova EV, He J, Ferrell RE, Goodpaster BH, Kelley DE. 2005. Deficiency of subsarcolemmal mitochondria in obesity and type 2 diabetes. *Diabetes* **54**:8–14. DOI: https://doi.org/10.2337/diabetes.54.1.8, PMID: 15616005

Scarpulla RC, Vega RB, Kelly DP. 2012. Transcriptional integration of mitochondrial biogenesis. *Trends in Endocrinology & Metabolism* **23**:459–466. DOI: https://doi.org/10.1016/j.tem.2012.06.006, PMID: 22817841

Schägger H, Pfeiffer K. 2000. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *The EMBO Journal* **19**:1777–1783. DOI: https://doi.org/10.1093/emboj/19.8.1777, PMID: 10775262

Schapira AH, Cooper JM, Dexter D, Clark JB, Jenner P, Marsden CD. 1990. Mitochondrial complex I deficiency in Parkinson's disease. *Journal of Neurochemistry* **54**:823–827. DOI: https://doi.org/10.1111/j.1471-4159.1990.tb02325.x, PMID: 2154550

Schlachetzki JC, Saliba SW, Oliveira AC. 2013. Studying neurodegenerative diseases in culture models. *Revista Brasileira de Psiquiatria* **35 Suppl 2**:S92–S100. DOI: https://doi.org/10.1590/1516-4446-2013-1159, PMID: 24271231

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**:671–675. DOI: https://doi.org/10.1038/nmeth.2089, PMID: 22930834

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**:337–342. DOI: https://doi.org/10.1038/nature10098, PMID: 21593866

Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, Ko MS, Ms K. 2009. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Research* **16**:45–58. DOI: https://doi.org/10.1093/dnares/dsn030, PMID: 19001483

Shaw G, Morse S, Ararat M, Graham FL. 2002. Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* **16**:869–871. DOI: https://doi.org/10.1096/fj.01-0995fje, PMID: 11967234

Shyh-Chang N, Zhu H, Yvanka de Soysa T, Shinoda G, Seligson MT, Tsanov KM, Nguyen L, Asara JM, Cantley LC, Daley GQ. 2013. Lin28 enhances tissue repair by reprogramming cellular metabolism. *Cell* **155**:778–792. DOI: https://doi.org/10.1016/j.cell.2013.09.059, PMID: 24209617

Sirey TM, Ponting CP. 2016. Insights into the post-transcriptional regulation of the mitochondrial electron transport chain. *Biochemical Society Transactions* **44**:1491–1498. DOI: https://doi.org/10.1042/BST20160100, PMID: 27911731

Skehan P, Storeng R, Scudiero D, Monks A, McMahon J, Vistica D, Warren JT, Bokesch H, Kenney S, Boyd MR. 1990. New colorimetric cytotoxicity assay for anticancer-drug screening. *JNCI Journal of the National Cancer Institute* **82**:1107–1112. DOI: https://doi.org/10.1093/jnci/82.13.1107, PMID: 2359136

Smillie CL, Sirey T, Ponting CP. 2018. Complexities of post-transcriptional regulation and the modeling of ceRNA crosstalk. *Critical Reviews in Biochemistry and Molecular Biology* **53**:231–245. DOI: https://doi.org/10.1080/10409238.2018.1447542, PMID: 29569941

Sokoloff L. 1977. Relation between physiological function and energy metabolism in the central nervous system. *Journal of Neurochemistry* **29**:13–26. DOI: https://doi.org/10.1111/j.1471-4159.1977.tb03919.x, PMID: 407330

Stalder L, Heusermann W, Sokol L, Trojer D, Wirz J, Hean J, Fritzsche A, Aeschimann F, Pfanzagl V, Basselet P, Weiler J, Hintersteiner M, Morrissey DV, Meisner-Kober NC. 2013. The rough endoplasmatic reticulum is a central nucleation site of siRNA-mediated RNA silencing. *The EMBO Journal* **32**:1115–1127. DOI: https://doi.org/10.1038/emboj.2013.52, PMID: 23511973

Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J, Califano A. 2011. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* **147**:370–381. DOI: https://doi.org/10.1016/j.cell.2011.09.041, PMID: 22000015

Tabas-Madrid D, Muniategui A, Sánchez-Caballero I, Martínez-Herrera DJ, Sorzano CO, Rubio A, Pascual-Montano A. 2014. Improving miRNA-mRNA interaction predictions. *BMC Genomics* **15 Suppl 10**:S2. DOI: https://doi.org/10.1186/1471-2164-15-S10-S2, PMID: 25559987

Tan JY, Vance KW, Varela MA, Sirey T, Watson LM, Curtis HJ, Marinello M, Alves S, Steinkraus B, Cooper S, Nesterova T, Brockdorff N, Fulga T, Brice A, Sittler A, Oliver PL, Wood MJ, Ponting CP, Marques AC. 2014. Cross-talking noncoding RNAs contribute to cell-specific neurodegeneration in SCA7. *Nature Structural & Molecular Biology* **21**:955–961. DOI: https://doi.org/10.1038/nsmb.2902, PMID: 25306109

Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, Wakamatsu A, Isogai T, Suzuki Y, Akimitsu N. 2012. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research* **22**:947–956. DOI: https://doi.org/10.1101/gr.130559.111, PMID: 22369889

Tat TT, Maroney PA, Chamnongpol S, Coller J, Nilsen TW. 2016. Cotranslational microRNA mediated messenger RNA destabilization. *eLife* **5**:e12880. DOI: https://doi.org/10.7554/eLife.12880, PMID: 27058298

Tay Y, Kats L, Salmena L, Weiss D, Tan SM, Ala U, Karreth F, Poliseno L, Provero P, Di Cunto F, Lieberman J, Rigoutsos I, Pandolfi PP. 2011. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* **147**:344–357. DOI: https://doi.org/10.1016/j.cell.2011.09.029, PMID: 22000013

van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, de Bruijn E, Hao W, MacInnes AW, Cuppen E, Simonis M. 2014. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biology* **15**:R6. DOI: https://doi.org/10.1186/gb-2014-15-1-r6, PMID: 24393600

van Waveren C, Moraes CT. 2008. Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. *BMC Genomics* **9**:18. DOI: https://doi.org/10.1186/1471-2164-9-18, PMID: 18194548

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology* **3**:research0034.1. DOI: https://doi.org/10.1186/gb-2002-3-7-research0034, PMID: 12184808

Vendramin R, Marine JC, Leucci E. 2017. Non-coding RNAs: the dark side of nuclear-mitochondrial communication. *The EMBO Journal* **36**:1123–1133. DOI: https://doi.org/10.15252/embj.201695546, PMID: 28314780

Vinothkumar KR, Zhu J, Hirst J. 2014. Architecture of mammalian respiratory complex I. *Nature* **515**:80–84. DOI: https://doi.org/10.1038/nature13686, PMID: 25209663

Vogel RO, Dieteren CE, van den Heuvel LP, Willems PH, Smeitink JA, Koopman WJ, Nijtmans LG. 2007. Identification of mitochondrial complex I assembly intermediates by tracing tagged NDUFS3 demonstrates the entry point of mitochondrial subunits. *Journal of Biological Chemistry* **282**:7582–7590. DOI: https://doi.org/10.1074/jbc.M609410200, PMID: 17209039

Vogel C, Abreu RS, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO, RdS A, Le S-Y SBA. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology* **6**:400. DOI: https://doi.org/10.1038/msb.2010.59, PMID: 20739923

Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P. 2015. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research* **43**:D174–D180. DOI: https://doi.org/10.1093/nar/gku1060, PMID: 25378313

Wahlestedt C. 2013. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nature Reviews Drug Discovery* **12**:433–446. DOI: https://doi.org/10.1038/nrd4018, PMID: 23722346

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**:582–587. DOI: https://doi.org/10.1038/nature13319, PMID: 24870543

Wills J, Edwards-Hicks J, Finch AJ. 2017. AssayR: A Simple Mass Spectrometry Software Tool for Targeted Metabolic and Stable Isotope Tracer Analyses. *Analytical Chemistry* **89**:9616–9619. DOI: https://doi.org/10.1021/acs.analchem.7b02401, PMID: 28850215

Wu Y, Whiteus C, Xu CS, Hayworth KJ, Weinberg RJ, Hess HF, De Camilli P. 2017. Contacts between the endoplasmic reticulum and other membranes in neurons. *PNAS* **114**:E4859–E4867. DOI: https://doi.org/10.1073/pnas.1701078114, PMID: 28559323

Yang Z, Feng Z, Gu J, Li X, Dong Q, Liu K, Li Y, OuYang L. 2017. microRNA-488 inhibits chemoresistance of ovarian cancer cells by targeting Six1 and mitochondrial function. *Oncotarget* **8**:80981–80993. DOI: https://doi.org/10.18632/oncotarget.20941, PMID: 29113360

Yuan B, Latek R, Hossbach M, Tuschl T, Lewitter F. 2004. siRNA Selection Server: an automated siRNA oligonucleotide prediction server. *Nucleic Acids Research* **32**:W130–W134. DOI: https://doi.org/10.1093/nar/gkh366, PMID: 15215365

Yue D, Liu H, Huang Y. 2009. Survey of Computational Algorithms for MicroRNA Target Prediction. *Current genomics* **10**:478–492. DOI: https://doi.org/10.2174/138920209789208219, PMID: 20436875

Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, Zhang C, Daneman R, Maniatis T, Barres BA, Wu JQ. 2014. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *Journal of Neuroscience* **34**:11929–11947. DOI: https://doi.org/10.1523/JNEUROSCI.1860-14.2014, PMID: 25186741

Zhou F, Li X, Wang W, Zhu P, Zhou J, He W, Ding M, Xiong F, Zheng X, Li Z, Ni Y, Mu X, Wen L, Cheng T, Lan Y, Yuan W, Tang F, Liu B. 2016. Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* **533**:487–492. DOI: https://doi.org/10.1038/nature17997, PMID: 27225119