Investigating molecular mechanisms of *Dali*, an intergenic chromatin-associated lincRNA regulating genes locally and neural differentiation genome-wide

by

Vladislava Chalei

Department of Physiology, Anatomy and Genetics

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2014

To Alexei

Word count

Chapters	**Words**
1	7,703
2	11,273
3	3,838
4	6,735
5	6,993
6	4,020
7	6,503
Total	47,065

Table of contents

List of figures

List of tables

Acknowledgements

I would like to express my sincere gratitude to my supervisors Dr Keith Vance and Prof Chris Ponting for their continuous support, encouragement and unwavering belief throughout my studies. It has been a privilege and pleasure to work with them both. I am especially grateful to Dr Keith Vance for being there for me every step of the way, for his contagious enthusiasm for research and for his patience. I would also like to acknowledge the help and advice of my collaborators in this project, in particular Dr Steve Sansom, Dr Lesheng Kong, and Sheena Lee.

I am also grateful to all the past and present members of the DARCGENs group for making this a stimulating and enjoyable atmosphere to work in, as well as to the wider Ponting group for creating a dynamic and diverse research environment to do science and learn in.

My deepest gratitude also goes to all the amazing friends I have been so lucky to have met in Oxford. I thank them for their belief and support, and for making my life outside work so happy over these three years.

I would also like to thank my former mentors and colleagues, both in Minsk and Edinburgh, from whom I learnt so much professionally and personally.

I am deeply grateful for the financial support over the years from the Darwin Trust of Edinburgh. I am particularly grateful to Sir Kenneth Murray, for the generosity of his spirit in supporting students and researchers and for giving me this tremendous opportunity to come to the UK.

I would like to express my deep appreciation to Huw Thomas. Over the past four years, he has been a source of continuous and patient support and encouragement. His belief in me far exceeds my own, and I am deeply grateful for that. This journey would be very different without him.

I am also thankful to my family for their love and support, even if from a distance.

Last but not least, my gratitude to Dr Alexei Lagodich can hardly be expressed in words. He has been my mentor, my friend, and my biggest inspiration ever since I started doing research. I have been enormously privileged to have benefited from his presence in my life and my career. And for that I am lost for words of gratitude.

My DPhil work has been supported by the Darwin Trust of Edinburg, MRC, and EPSRC.

Declaration

The work presented in this thesis has been conducted by the author Vladislava Chalei in the Department of Physiology, Anatomy and Genetics at the University of Oxford. Except where acknowledgement is made, all the work reported in this thesis is my own.

Specifically, microarray chip processing was performed by Sheena Lee (OXION microarray facility, DPAG, University of Oxford). Next-generation sequencing data generation was performed by the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (University of Oxford). Sanger sequencing was performed by Source Bioscience (Department of Biochemistry, University of Oxford). Mass spectrometry was performed at the Central Proteomics Facility (Sir William Dunn School of Pathology, University of Oxford). Computational analysis of transient *Dali* and *Pou3f3* knockdown microarray data and CHART-Seq data (peak calling, genome segment analysis, chromatin mark enrichment analysis, motif discovery analysis) were performed by Dr Stephen Sansom (former at CGAT, MRC FGU, University of Oxford, currently at The Kennedy Institute of Rheumatology, University of Oxford). Mouse work (mouse embryonic and adult brain dissection) was performed by Dr Juan Montiel (MRC FGU, University of Oxford).

Brief descriptions of the methods used by the referenced authors are given in the Chapter 2 Materials and Methods purely for ease of understanding.

I hereby declare that this thesis has not been submitted either in the same or different form, to this or any other university for a degree.

Vladislava Chalei

Investigating molecular mechanisms of *Dali*, an intergenic chromatin-associated lincRNA regulating genes locally and neural differentiation genome-wide

A thesis submitted to the degree of Doctor of Philosophy

Vladislava Chalei, St Cross College, Trinity Term 2014

Abstract

Recently, long non-coding RNAs (lncRNAs) emerged as important regulators of many cellular functions. Many nuclear lncRNAs regulate the expression of geomically proximal or overlapping protein coding genes. Less clear is whether intergenic lncRNAs can regulate transcription by modulating chromatin at genomically distant loci in an RNA-dependent manner. This thesis investigated molecular functions of *Dali*, an intergenic central nervous system expressed lncRNA conserved in therian mammals. *Dali* is transcribed from a locus 50 kb downstream of the *Pou3f3* transcription factor gene and performs both genomically local and distal RNA-dependent roles. Its depletion disrupts the differentiation of neuroblastoma cells. Locally, *Dali* regulates transcription of the *Pou3f3* locus. Distally, it preferentially binds near to and regulates active promoters across the genome, including by physically associating with the POU3F3 transcription factor. *Dali* also interacts with the DNMT1 DNA methyltransferase in mouse and human and regulates CpG island-associated promoters by modulating their DNA methylation levels *in trans*. This work is the first todemonstrate that a lncRNA can regulate the DNA methylation of CpG island-associated promoters *in trans* and one of the first large scale studies to identify direct transcriptional targets of a lncRNA genome-wide. It also provides a more detailed molecular dissection of the extended *Pou3f3* locus and a framework for the prioritisation and comprehensive functional characterisation of nuclear lncRNAs.

List of abbreviations

%	per cent
°C	degree Celcius
1/2sbsRNAs	half-STAU1-binding site RNAs
116HG	SNORD116 host gene
1d-eRNAs	unidirectional enhancer RNAs
ADP	adenine diphosphate
Airn	antisense to Igf2r RNA non-coding
AldoC	aldolase C
Alu element	Arthrobacter luteus element
ANRIL	antisense non-coding RNA in the INK4 locus
AR	androgen receptor
AREs	androgen response elements
Atm	ataxia-telangiectasia mutated
ATP	adenine triphosphate
ATPase	adenosine triphosphatase
B2 SINE RNA	B2 short interspersed element RNA
BAC	Bacterial artificial chromosome
BACE1-AS	BACE1 antisense RNA
BAF	BRG-associated factor
BAF53a	BRG-associated factor 53a
Borg3	binder of Rho GTPases 3
bp	base pair
BRG1	Brahma-related gene 1

Brn1	Brain-1
C/EBPα	CCAAT-enhancer-binding protein α
cAMP	cyclic adenosine monophosphate
Cbs	cystathionine-beta-synthase
CBX7	chromobox homolog 7
CCND1	cyclin D1
CD3	CD3 antigen
Cdc25b/25c/7	cell division cycle 25B/25C/7
Cdc42ep5	CDC42 effector protein (Rho GTPase binding) 5
Cdc42	Cell division control protein 42
Cdca2/3/7	Cell division cycle associated 2/3/7
CDKN2A/CDKN2B	cyclin dependent kinase 2a/2b
cDNA	complementary DNA
CEBPA	CCAAT/enhancer-binding protein alpha
Cend1	cell cycle exit and neuronal differentiation 1
ceRNA	Competitive endogenous RNA
CGI	CpG island
CHART-Seq	capture hybridization analysis of RNA targets sequencing
CHD	Chromo Helicase DNA-binding
Chek1	Checkpoint kinase 1
ChIP	Chromatin immunoprecipitation
ChIRP	Chromatin Isolation by RNA purification
circRNAs	Circular RNAs
Cited1	CREB-binding protein/p300-interacting transactivator with Asp/Glu-rich C-terminal domain

cm^2	Centimetre squared
CMV	cytomegalovirus
CNS	Central nervous system
COBRA	Combined bisulfite restriction analysis
C-oligos	capture oligos
CoREST	co-repressor for element-1-silencing transcription factor
CpG	*cytosine-phosphate-guanin*
CRE	cAMP response element
CREB	*cAMP response element-binding* protein
CREST	calcium-responsive transactivator
Ct	threshold cycle
Ctbp1	C-terminal-binding protein 1
Ctbp1-as	C-terminal-binding protein 1 antisense RNA
CTCF	CCCTC-binding factor
Dali	DNMT1-associated long intergenic RNA
DCP	dorsal cortical plate
DG	Dentate gyrus
DHFR	Dihydrofolate reductase
Dkk1	Dickkopf Homolog 1
Dlgap5	Disks large-associated protein 5
Dlx	The Distal-less family
Dlx1os1	*distal-less* homeobox 1 overlaping transcript
DLX2	*distal-less* homeobox 2
Dlx5/6	*distal-less* homeobox 5/6
Dlx6os1	*distal-less* homeobox 6 opposite strand RNA

DNA	Deoxyribonucleic acid
DNase I	Deoxyribonuclease I
DNMT1/3A/3B	DNA methyltransferase 1/3A/3B
DOT1L	DOT1-like
DPP	Decapentaplegic
Dscam	Down Syndrome Cell Adhesion Molecule
DVZ	dorsal ventricular zone
E13.5/15.5/18.5	Embryonic day 13.5/15.5/18.5
E2	estradiol
E2f2	E2F transcription factor 2
ecCEBPA	*extra-coding* CEBPA RNA
EF1a	elongation factor 1-alpha
Efna1	Ephrin-A1
elncRNA	enhancer-associated long non-coding RNA
EPCAM	epithelial cell adhesion molecule
ER	estrogen receptor
eRNAs	enhancer RNAs
ES cells	embryonic stem cells
Evf-2	Embryonic ventral forebrain-2
EZH2	enhancer of zeste homolog 2
Fam5b	family with sequence similarity 5, member B
Fbn1	fibrillin 1
FDR	False discovery rate
Fendrr	Fetal-lethal noncoding developmental regulatory RNA
Fkbp14	FK506 binding protein 14

FoxC1	Forkhead box C1
Fpgs	folylpolyglutamate synthase
g	gram
G9a	Protein G9a
GABA	gamma-aminobutyric acid
Gad67	Glutamic Acid Decarboxylase 67
Gapdh	Glyceraldehyde 3-phosphate dehydrogenase
Gas5	Growth arrest-specific 5
GC	guanine-cytosine
gDNA	genomic DNA
GO	Gene Ontology
GR	glucocorticoid receptor
GREs	glucocorticoid receptor response elements
Grin1	glutamate receptor, ionotropic, N-methyl D-aspartate 1
GTPase	guanosine triphosphatase
h	hour
H3K27ac	histone H3 Lysine 27 acetylation
H3K27me3	histone H3 Lysine 27 tri-methylation
H3K36me	histone H3 Lysine 36 methylation
H3K4	histone H3 Lysine 4
H3K4me1	histone H3 Lysine 4 mono-methylation
H3K4me3	histone H3 Lysine 4 tri-methylation
H3K9	histone H3 Lysine 9
HAR1A	human accelerated region 1A
HBA2	haemoglobin alpha 2

hGH	human growth hormone
Hmgb2	high-mobility group protein B2
HOTAIR	HOX antisense intergenic RNA
HOTAIRM1	HOX antisense intergenic RNA myeloid 1
HOTTIP	HOXA transcript at the distal tip
HOXA/C/D	Homeotic gene cluster A/C/D
Hoxa1/6/7/13	Homeobox A1/6/7/13
HS	hyper-sensitivity
HS1	Hyper sensitive region 1
IFN-γ	interferon gamma
Igf2r	insulin-like growth factor 2 receptor
Igfbp5	insulin-like growth factor binding protein 5
IgG	immunoglobulin G
JNK	c-Jun N-terminal kinase
Jpx	*Jpx* transcript, Xist activator
kb	kilobase
Kcnq1	Potassium (K) voltage-gated channel, KQT-like subfamily, member 1
Kcnq1ot1	KCNQ1 overlapping transcript 1
KDM1A	lysine (K)-specific demethylase 1A
Kif2c/11	Kinesin family member 2c/11
lacZ	beta-galactosidase
LCP	lateral cortical plate
linc-Brn1a/1b	long intergenic non-coding RNA-Brn1- a/b
linc-Hoxa1	long intergenic non-coding-Hoxa1

lincRNA-p21	long intergenic non-coding RNA-p21
lincRNAs	long intergenic non-coding RNAs
lncRNA	long non-coding RNAs
lncRNA-N1/N3	long non-coding RNA N1/N3
loxP	locus of crossover in P1
LPP	lateral pre-late
LSD1	lysine-specific demethylase 1
LUC7L	Luc7 homolog (S. cerevisiae)-like
LVZ	lateral ventricular zone
M	molar
Malat1	metastasis associated lung adenocarcinoma transcript 1
MAP	microtubule-associated *protein*
MAPKKK	MAP kinase kinase kinase
Mb	megabase
MeCP1/2	methyl-CpG binding protein 1/2
Meis1	myeloid ecotropic viral integration site 1 homolog
Meis2	Myeloid Ecotropic Viral Integration Site 1 Homolog 2
mg	milligram
Miat	myocardial infarction associated transcript
min	minute
miRNAs	micro RNAs
Mitf	Microphthalmia-associated transcription factor
ml	milliliter
mL	millilux
MLE	mailless

MLL	Mixed lineage leukemia
mM	millimolar
mm	millimeter
mRNAs	messenger RNAs
MSH2	mutS homolog 2
MSL	male specific lethal
MSL2	male specific lethal 2
My	million years
MyoD1	myogenic differentiation 1
nBAF	neuronal BRG-associated factor complex
ncRNA	non-coding RNA
ncRNA-a	ncRNA-activating
Neat1	noncoding nuclear-enriched abundant transcript 1
Neat2	noncoding nuclear-enriched abundant transcript 2
NeST	nettoie Salmonella pas Theiler's
NF-YA	nuclear transcription factor Y subunit A
ng	nanogram
NK-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
nm	nanometer
NMIs	non-methylated islands
n-Myc	V-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog
NO	nitrogen oxide
Nos1	nitrogen oxide synthase 1
npBAF	neuronal precursor BRG-associated factor complex

NSC	neuronal stem cell
Nrcam	neuronal cell adhesion molecule
NRIP1	nuclear receptor-interacting protein 1
NSC	neural stem cells
nt	nucleotide
OB	olfactory bulb
Oct8	octamer-binding protein 8
ORF	open reading frame
ORC3	origin recognition complex subunit 3
p53	tumor protein p53
P56	postnatal day 5
p66	transcriptional repressor p66
PANDA	P21 associated *ncRNA* DNA damage activated
PARP1	Poly (ADP-ribose) polymerase 1
Paupar	Pax6 upstream antisense non-coding RNA
Pax3/6	paired box 3/6
Pc2	Polycomb 2 homolog
Pcgem1	prostate-specific transcript (non-protein coding)
PCR	polymerase chain reaction
pg	picogram
piRNAs	PIWI-interacting RNA
Pitx2	Paired-like homeodomain transcription factor 2
pmol	picomolar
POU	Pit-Oct-Unc family of transcription factors
Pou3f3	POU Class 3 Homeobox 3

PRC1	Polycomb-group repressive complex 1
PRC2	Polycomb-group repressive complex 2
Prdm8	PR domain containing 8
Prncr1	prostate cancer associated non-coding RNA 1
P-TEFb	positive transcription elongation factor,
PUF protein	Pumilio/FBF homology proteins
PWS	Prader-Willi syndrom
PYGO2	pygopus family PHD finger 2
qPCR	quantitative polymerase chain reaction
R	correlation coefficient
RA	retinoic acid
Rac	Ras-related C3 botulinum toxin substrate protein
RAP	RNA antisense purification
RBBP5	Retinoblastoma-binding protein 5
RBP	RNA-binding protein
rDNA	ribosomal DNA
RepA	Repeat A
RepC	Repeat C
REST	RE1-Silencing Transcription factor,
RIP	RNA immunoprecipitation
Rit2	Ras-like without CAAX 2
RMST	Rhabdomyosarcoma 2 associated transcript
RNA	ribonucleic acid
RNA PolII	RNA polymerase II
RNase H	ribonuclease H

rox2	RNA on the X 2
rpm	rotations per minute
rRNA	Ribosomal RNA
RT-qPCR	reverse transcription quantitative polymerase chain reaction
sec	second
SET1	Su(var)3-9, Enhancer-of-zeste, Trithorax domain protein 1
shRNA	short hairpin RNA
SIN3A	SIN3 transcription regulator homolog A
Six3	SIX homeobox 3
Six3os1	Six3 overlapping transcript 1
Smad2	Mothers against decapentaplegic homolog 2
SNORD116	small nucleolar RNA SNORD116
snoRNAs	small nucleolar RNAs
snRNAs	small nuclear RNAs
snRNP	small nuclear ribonucleic particles
SOX2	SRY-box 2
Sp1	specificity protein 1
Sparc	secreted protein acidic and rich in cysteine
ssDNA	single-strand RNA
ssEF1a	EF1a ssRNA
STAT3	Signal transducer and activator of transcription 3
SUZ12	Suppressor Of Zeste 12 Protein Homolog
SVZ	subventricular zone
SWI/SNF	SWItch/Sucrose NonFermentable
TALE	transcription activator-like effector

TALEN	transcription activator-like effector nuclease
TALETF	transcription activator-like effector transcription factor
Tbp	TATA-binding protein
TERC	elomerase RNA component
TF	transcription factor
TFF1	Trefoil factor
TFIIB	Transcription factor IIB
Tgfb2/3	transforming growth factor, beta 2/3
THRIL	TNF and HNRNPL related immunoregulatory long non-coding RNA
TINCR	terminal differentiation-induced *ncRNA*
Tlr4	Toll-like receptor 4
TLS	Translocated in Sarcoma
TNFα	Tumor necrosis factor alpha
tRNAs	Transport RNAs
Trp53cor1	tumor protein p53 pathway corepressor 1
Tsix	transcript, XIST antisense RNA
TSS	transcription start site
Tug1	taurine up-regulated gene 1
U	Unit
UBE3a	Ubiquitin-protein ligase E3A
Uchl1-as1	UCHL1 antisense RNA 1
UCRs	ultraconserved regions
UHRF1	Ubiquitin-like, containing PHD and RING finger domains, 1
UTR	untranslated region

UV-RIP	ultraviolet cross-linked RNA immunoprecipitation
V	Volt
Vax2os	Vax2 opposite strand transcript
VZ	ventricular zone
WDR5	WD repeat-containing protein 5
Wnt	wingless and integrated (int-1)
x g	gravitational acceleration
XACT	X-active coating transcript
Xist	X-inactive specific transcript
YY1	Yin Yang 1
μg	microgram
μl	microliter
μM	micromolar

Chapter 1 General Introduction

1.1 Introduction to lncRNAs

The existence of non-coding RNA in the cell has been recognized for a long time. The author of the "central dogma" of molecular biology, Francis Crick, recognized that there may be "special" and "unknown" mechanisms of information transfer between nucleic acids (Crick, 1970). However, the diversity and abundance of the non-coding portion of the transcriptome which has come to light recently were not anticipated. To date, at least 85% of the human genome has been shown to be transcribed at some point, and the idea of pervasive transcription is now firmly established (ENCODE, 2012; Hangauer et al., 2013). This contrasts greatly with the approximately 1-2% of the genome that encode protein exons (Djebali et al., 2012). Furthermore, while neither protein coding gene number nor genome size correlate with the organismal complexity of different taxonomic groups (even after the correction for the diversity of protein isoforms produced by the alternative splicing and post-translational processing events), the proportion of genomes being transcribed appears to be a much better predictor of complexity (Taft et al., 2007). This implies that non-coding transcriptome may perform important regulatory roles that have contributed to the evolution of developmental and organismal complexity.

In mammals, the number of identified non-coding RNA (ncRNA) loci already exceeds the number of protein-coding genes. Apart from different groups of small ncRNAs (such as miRNAs, snoRNAs, snRNAs, piRNAs and tRNAs) (Kowalczyk et al., 2012), a large portion of the genome is transcribed into RNAs that are longer than 200 nt and are devoid

of protein-coding potential. These RNAs are referred to as long non-coding RNAs, or lncRNAs (Derrien et al., 2012; Guttman et al., 2013).

LncRNA can be produced from a variety of genomic loci. Those that are transcribed fully outside of the annotated protein-coding genes are referred to as intergenic lncRNAs, or lincRNAs. LncRNAs can also be produced from intronic regions (intronic lncRNAs). Sense lncRNAs at least partially overlap protein-coding genes on the sense strand, while antisense lncRNAs overlap such genes on the opposite, or antisense, strand. Additionally, some lncRNAs are transcribed divergently from neighbouring protein-coding genes with which they share transcription initiation region (Ma et al., 2013).

1.2 LncRNAs as regulators of mammalian gene expression

The recent explosion of lncRNA studies has implicated them in almost every aspect of gene expression regulation (reviewed in (Fatica and Bozzoni, 2014)). Physicochemical properties of RNAs make them uniquely suited for performing versatile regulatory roles. First and most obviously, RNA molecules can base-pair with nucleic acids, both other RNAs as well as DNA. Therefore, in evolution RNAs can be efficiently co-opted to encode sequence-specific recognition of targets either in the genome or the transcriptome. By comparison, to encode similar functions in a sequence-specific protein (such as transcription activator-like effector (TALE) or Pumilio and FBF homology (PUF) proteins) requires two orders of magnitude more genomic space (Filipovska and Rackham, 2012). Importantly, as RNAs can interact with nucleic acids at any stage of their metabolism, they can conceivably regulate all and every stage of gene expression, from transcription to

processing to translation and degradation of mRNAs. Second, long RNA molecules, such as lncRNAs, can adopt complex three-dimensional structures that can serve as interaction interfaces for other molecules, such as small ligands or RNA-binding proteins (Stoltenburg et al., 2007). Also, an RNA molecule can be rapidly produced, becomes functional essentially instantaneously, and can be equally rapidly degraded. This makes it uniquely suited for highly dynamic regulatory contexts. RNA structure is also dynamic and can be altered by ligand binding, allowing for the integration of multiple regulatory inputs (Serganov and Patel, 2007). Importantly, unconstrained by protein-coding potential, RNAs are more tolerant of mutations and, therefore, potentially provide a platform for rapid evolutionary innovation.

The unique chemical properties of RNAs allow them to interact with both proteins (such as transcriptional regulators and chromatin modifiers) and DNA. LncRNAs in particular can interact with both classes of molecules simultaneously, potentially serving as molecular and functional bridges between the genome and its regulatory machinery (Batista and Chang, 2013; Guttman and Rinn, 2012). Furthermore, these transcripts can also bring together proteins or protein complexes which otherwise would not interact via protein-protein interactions. This property of lncRNAs can potentially greatly expand the functional repertoire of protein ensembles (Guttman et al., 2011; Ng et al., 2012; Rinn et al., 2007). If that was not enough, the diversity of lncRNA isoforms (generated by alternative transcriptional start and termination sites, alternative splicing, post-transcriptional processing etc) can further increase the combinatorial complexity of regulatory interactions.

Large scale transcriptomic surveys in various vertebrate species have shown that lncRNAs tend to have more tissue-restricted expression profiles compared to protein-coding genes (Cabili et al., 2011; Derrien et al., 2012; Mercer et al., 2008; Pauli et al., 2012), indicating that their transcription may be spatially tightly regulated. Also, many lncRNAs have been found to be differentially expressed during cell differentiation, suggesting temporal regulation of transcription during development (Guttman et al., 2011; Ng et al., 2012; Ramos et al., 2013). Within the cell, lncRNAs can localise to both the nucleus and the cytoplasm (Derrien et al., 2012; van Heesch et al., 2014). Together with the previously discussed chemical properties of lncRNAs, this suggests that they can regulate both transcriptional and post-transcriptional stages of gene expression in specific temporal and spatial contexts. Despite the great variety of regulatory roles assigned to lncRNAs to date, it is likely that many more functions are yet to emerge.

1.2.1 Nuclear lncRNAs and transcriptional regulation

A growing number of nuclear localised lncRNAs have been found to regulate transcription and chromatin organisation (Vance and Ponting, 2014a). Some of these act near to their site of synthesis to regulate gene expression locally on the same chromosome in an allele specific manner (*cis*-acting), while others act *in trans* to regulate genes located on multiple chromosomes, and on either allele. For some lncRNA loci it was demonstrated that the act of transcription itself, and not the resulting RNA product, confers function (a transcription-dependent RNA-independent mode of action), while for others the RNA molecule itself was shown to be important (an RNA-dependent mode of action). Of note, some nuclear lncRNAs regulate gene expression indirectly, for example, by sequestering transcription

factors away from their targets (Hung et al., 2011; Sun et al., 2013) or by allosterically modulating the activity of effector proteins (Wang et al., 2008). The prevalence of *cis* versus *trans*, as well as transcription- versus RNA-dependent mechanisms among lncRNAs, are not presently known.

The extent to which lncRNAs are functional is intensely debated. On one hand, many lncRNAs display features consistent with functionality, such as evolutionary conservation. Multiple studies have found lncRNAs to be evolving under modest evolutionary constraint (Ponjavic et al., 2009). Furthermore, expression of many lncRNAs is actively regulated, both transcriptionally (Hangauer et al., 2013) and post-transcriptionally (Clark et al., 2012). Genome-wide transcriptomic studies in various vertebrates found lncRNAs to have more tissue-restricted expression profiles than protein-coding genes (Cabili et al., 2011; Derrien et al., 2012; Mercer et al., 2008; Pauli et al., 2012). Also, many lncRNAs are differentially expressed during cell differentiation, indicating temporal regulation of transcription during development (Guttman et al., 2011; Ng et al., 2012; Ramos et al., 2013).

However, lncRNAs as a group are expressed at lower levels than protein-coding genes. Moreover, only 19% of human organ-specific lncRNAs have been conserved over > 90 My of evolution (Necsulea et al., 2014) and, if conservation is a good indicator, are likely to be biologically important. Also, a large proportion of lncRNAs are mono-exonic. Mono-exonic lncRNAs are less conserved evolutionarily compared to multi-exonic ones and are often expressed at very low levels and not in all individuals (W. Haerty, personal communication). Therefore, these lncRNAs are less likely to harbour an important function. These features can be associated with either transcriptional “noise” spuriously

initiated at low frequency from random regions in the genome or genomic contamination of sequencing libraries.

In summary, it is likely that not all lncRNAs will prove functional, although thousands of these transcripts discovered to date mean that there are potentially large numbers of important novel cellular regulators of this class.

1.2.1.1 Transcription-dependent and RNA-dependent modes of action

Evidence has accumulated that the traversal of the transcriptional machinery across a regulatory DNA element or the change of overall chromatin structure of a locus can affect gene expression nearby, either positively or negatively (reviewed in (Kornienko et al., 2013)). Transcription-mediated silencing of the expression of one transcript by another transcriptional event is referred to as "transcriptional interference". The mechanistic details behind this phenomenon remain largely unclear, but most often transcriptional interference occurs when the two transcripts originate from overlapping promoters, although in mouse overlapping transcription has been shown to control the choice of polyadenylation sites of two imprinted genes (McIsaac et al., 2013; Wood et al., 2008).

Several studies suggest that DNA methylation may be involved in transcriptional interference-mediated silencing. In a patient with inherited α-thalassemia, a pathologic deletion of the 3′ end of the *LUC7L* gene resulted in the read-through transcription and repression of the downstream *HBA2* (Tufarelli et al., 2003). In mouse models of this deletion, *HBA2* silencing was shown to be caused primarily by the acquisition of DNA methylation at its promoter. Importantly, DNA methylation was not consequential to the

cessation of transcription from the promoter, because when *HBA2* transcription was disrupted by removing its TATA-box no DNA methylation acquisition was observed. The contributory roles of the aberrant read-through RNA product or the *LUC7L* sequence were also ruled-out because the same repressive effect on *HBA2* was observed when the region was replaced with an unrelated protein-coding gene (Tufarelli et al., 2003). Similarly, aberrant transcription from the *EPCAM* locus carrying a 3′ deletion causes DNA methylation and transcriptional silencing of the downstream *MSH2* gene in a subset of patients with Lynch syndrome (Ligtenberg et al., 2009). In both of these cases, the mechanistic details of how DNA methylation is established are not yet clear. However, all evidence thus far points to a *cis*-acting mechanism, because the observed silencing effects are allele-specific.

Of note, the repression of imprinted genes by lncRNAs is also often associated with the acquisition of DNA methylation at their promoters (Santoro and Barlow, 2011). However, at least in the case of *Igf2r*, DNA methylation appears to reinforce the repressed state, but not to initiate it (Santoro et al., 2013). Instead, the available evidence indicates that *Igf2r* is silenced by the imprinted lncRNA *Airn* without introducing chromatin changes (Koerner et al., 2009; Santoro and Barlow, 2011). *Airn* is an inefficiently spliced 118 kb long lncRNA transcribed from the paternal copy of the chromosome. It transcriptionally overlaps and represses the *Igf2r* promoter. By using a polyadenylation signal, the *Airn* transcript was shortened either before or after the *Igf2r* promoter to show that only the transcript traversing the promoter was capable of silencing it (Latos et al., 2012). Although *Igf2r* silencing is eventually accompanied by DNA methylation, repressive histone marks and chromatin compaction at its promoter, *Airn* was shown to induce initial silencing in the

context of open chromatin by interfering with the occupancy of functional RNA PolII (Latos et al., 2012). Additional support to the transcriptional interference model of *Igf2r* silencing is provided by experiments that demonstrated that *Airn* transcription is continuously required for *Igf2r* repression and that repression is relaxed when the *Igf2r* promoter is strongly induced (Santoro et al., 2013). LncRNAs of approximately 0.8–7.3 kb in length transcribed from the *DHFR* promoter have also been shown to inhibit RNA PolII pre-initiation complex assembly by forming stable triplex structures with the region. Additionally, these transcripts were shown to interact with the general transcription factor TFIIB causing dissociation of the pre-initiation complex (Blume et al., 2003; Martianov et al., 2007).

It has been suggested that the alternative promoter of the *Fpgs* gene may also be silenced by lncRNA transcription in the absence of chromatin changes. However, this system has not been experimentally dissected to the same extent as *Igf2r* (Racanelli et al., 2008), and other examples of this mechanism in mammals are currently lacking. The mechanistic details of how transcription from an interfering promoter can induce transcriptional silencing of the sensitive promoter is not well understood, but stalling of the interfering transcriptional complex was suggested to occlude binding sites of essential transcription factors (Palmer et al., 2009; Palmer et al., 2011).

Transcriptional interference may also suppress distal enhancer function, which could help explain how lncRNAs such as *Airn* and *Kcnq1ot1* that overlap only one gene can repress whole clusters of genes in a tissue-specific manner (Pauler et al., 2012). A two-step model was proposed whereby lncRNA transcription initiates repression of non-overlapping genes by enhancer interference. The repressed state is then maintained by histone modifying

enzymes to which some of these lncRNAs bind (Korostowski et al., 2012; Mager et al., 2003; Nagano et al., 2008; Terranova et al., 2008).

Positive regulation by lncRNA transcription has also been demonstrated. For example, many (if not all) enhancers produce lncRNAs referred to as enhancer RNAs (eRNAs) whose levels positively correlate with the expression of their neighbouring genes (Kim et al., 2010; Li et al., 2013b; Orom et al., 2010). In fact, approximately 20% of lncRNAs map to enhancer regions (Bonasio and Shiekhattar, 2014). Moreover, a large proportion of intergenic lncRNAs were found to originate from loci bearing enhancer-like chromatin marks, suggesting that these may represent novel enhancers (Marques et al., 2013). It was suggested that enhancer transcription may establish a chromatin state that is permissive to enhancer function. Likewise, lncRNA transcription may result in promoter activation by either facilitating the opening of chromatin or disrupting repressor protein binding. For example, in the mouse β-globin locus intergenic lncRNA transcription was shown to establish a stable transcriptionally-permissive and hyper-accessible chromatin state required for β-globin expression (Ashe et al., 1997; Gribnau et al., 2000). Similarly, antisense intergenic transcription through the V, D and J genes is observed prior to, and is required for, the recombination process that results in the assembly of functional B cell immunoglobulins and T cell receptors. Presumably, in this case lncRNA transcription facilitates the opening of chromatin for greater access to the recombination machinery (Abarrategui and Krangel, 2007; Bolland et al., 2007; Giallourakis et al., 2010).

The activation of the *HOXA* cluster in both mouse and human provides an example of positive regulation mediated by lncRNA transcription. Here, transcription is thought to prevent access of repressive protein complexes to chromatin, as lncRNA expression

positively correlated with the loss of Polycomb chromatin occupancy that precedes transcriptional activation (Sessa et al., 2007). LncRNAs have also been observed at promoters of other Polycomb-regulated genes in mouse, where they have been suggested to either facilitate or inhibit Polycomb binding (Davidovich et al., 2013; Hekimoglu-Balkan et al., 2012). However, it is presently unclear if promoter-associated lncRNA transcription is a by-product or else is required for the maintenance of open chromatin state in these cases (Kowalczyk et al., 2012; Wang et al., 2011a).

In summary, due to wide spread lncRNA transcription across mammalian genomes, including in the vicinity of genes and regulatory elements, the act of transcription itself may have profound implications for gene expression regulation and thus merits further detailed investigation.

1.2.1.2 *Cis*-acting lncRNAs

A transcription-dependent RNA-independent mode of lncRNA function *a priori* implicates a *cis*-acting mechanism. However, lncRNA loci that function in an RNA-dependent manner also can do so *in cis*. *Cis*-acting lncRNA regulatory mechanisms have been described in detail for a number of enhancer associated nuclear lncRNAs, as well as lncRNAs involved in the processes of genomic imprinting and X chromosome inactivation (Melo et al., 2013; Monnier et al., 2013; Mousavi et al., 2013; Santoro et al., 2013; Tian et al., 2010; Vallot et al., 2013).

The mechanism by which *cis*-acting lncRNA molecules regulate local gene expression relies on their interaction with transcriptional regulators and/or chromatin modifiers. For

example, the vertebrate-conserved *HOTTIP* lncRNA, which is transcribed upstream of *HOXA13* at the 5′ end of the *HOXA* cluster, regulates several of the downstream 5′-*HOXA* genes in human. This regulation relies on both *HOTTIP*'s interaction with the activating H3K4 methyltransferase complex MLL1 (mixed-lineage leukemia) and looping of the chromatin fiber to bring the *HOTTIP* locus into close proximity to some of its target promoters (Wang et al., 2011b). The same mechanism was proposed for the mouse-specific lncRNA *Mira* that also originates from the *HOXA* locus and positively regulates its neighbouring *HOXA6* and *HOXA7* genes. The *HOTAIRM1* lncRNA, which is transcribed in myeloid lineages from the 3′ end of the *HOXA* cluster, positively regulates 3′-*HOXA* genes *in cis* via a hitherto unknown mechanism (Zhang et al., 2009), while *linc-Hoxa1* represses *Hoxa1 in cis* in mouse ES cells by recruiting the transcriptional co-factor PURB (Maamar et al., 2013).

Overall, despite a recent dramatic increase in our understanding of local *cis*-regulatory roles of lncRNAs, many important questions remain unanswered. In particular, it is unclear how these transcripts are retained at their site of synthesis where they are presumed to accumulate sometimes to relatively high concentrations to perform their functions.

1.2.1.3 *Trans*-acting lncRNAs

Although it is now apparent that lncRNAs can be involved extensively in regulating genes in the vicinity of their sites of synthesis, the ability of lncRNAs to exert widespread effects on gene expression *in trans* is poorly understood. This is in large part because direct transcriptional targets for only very few of these transcripts have thus far been identified (Chu et al., 2011; Ng et al., 2013; Simon et al., 2011; Vance et al., 2014). Moreover, it is

not clear whether these transcripts act directly, or within ribonucleoprotein complexes, how they modify their target genes' regulatory landscape and/or chromatin status, or even how they are targeted to specific genomic regions.

One of the first examples of a *trans*-acting lncRNA was provided by *HOTAIR* (Rinn et al., 2007). *HOTAIR* is an approximately 2.2 kb long lncRNA produced from the *HOXC* cluster on chromosome 12 which acts *in trans* to repress the *HOXD* cluster located on chromosome 2. *HOTAIR* is thought to recruit two repressive complexes, PRC2 (polycomb repressive complex 2) and KDM1A-CoREST-REST, to specific target genes. The knockdown of *HOTAIR* resulted in the decreased activity of the two protein complexes in the locus and simultaneous transcriptional up-regulation of *HOXD* genes. Mouse *Hotair* knockout models produced by a several kb long genomic deletion are viable, but exhibit developmental malformations and homeotic transformations in their skeletal system (Maamar et al., 2013). Consistently with *Hotair*'s role in *HoxD* repression, the knockout model also displayed de-repression of some of the *HoxD* genes (Li et al., 2013a). However, *Hotair* is transcribed from a locus positioned within a dense cluster of tightly developmentally regulated genes. A genomic deletion in such a locus is likely to produce topological re-arrangements as well as affect DNA functional elements, which may perturb the gene expression in the cluster. Therefore, more work is needed to conclusively show the role of *Hotair* in the observed phenotypes.

Apart from transcription factors and chromatin modifiers, *trans*-acting lncRNAs can associate with transcription elongation factors or the RNA polymerases themselves to regulate transcription. For example, an approximately 330 bp long 7SK RNA was found to perform a scaffolding function in the 7SK snRNP complex, which represses positive

transcription elongation factor P-TEFb, leading to the inhibition of transcriptional elongation of many genes genome-wide (Nguyen et al., 2001; Yang et al., 2001). The mouse B2 SINE RNA, on the other hand, stably associates with RNA PolII to repress its activity during heat shock response (Espinoza et al., 2004), while *Alu* RNA performs an analogous role in human (Mariner et al., 2008).

1.2.2 Cytoplasmic lncRNAs and post-transcriptional regulation

According to some estimates, the majority of human lncRNAs are enriched in the cytoplasm (van Heesch et al., 2014). Non-surprisingly, they have been ascribed a multitude of important regulatory roles there (Batista and Chang, 2013). The mechanism of function of many cytoplasmic lncRNAs relies on the sequence complementarity with either antisense transcripts from the same locus, or completely independent transcripts encoded elsewhere in the genome. For example, cytoplasmic lncRNAs, such as ubiquitin carboxy-terminal hydrolase L1 antisense RNA 1 (*Uchl1-as1*) and tumour protein p53 pathway co-repressor 1 (*Trp53cor1*; otherwise known as *lincRNA-p21*), regulate the translation of their complementary targets positively and negatively, respectively (Carrieri et al., 2012; Yoon et al., 2013). Similarly, cytoplasmic lncRNA can also both positively and negatively regulate mRNA stability. For example, β-site APP-cleaving enzyme 1-antisense (*BACE1-AS*) and tissue differentiation-inducing non-protein-coding RNA (*TINCR*) lncRNAs stabilise, while half-STAU1-binding site RNAs (1/2sbsRNAs) de-stabilise, their respective target mRNAs (Faghihi et al., 2008; Gong and Maquat, 2011; Kretz et al., 2013; Wang et al., 2013a).

Furthermore, some cytoplasmic lncRNAs have been proposed to act as miRNA "sponges" by binding and sequestering miRNAs away from the target mRNAs they normally repress. LncRNAs that work in this way are referred to as competitive endogenous RNAs (ceRNA) (Salmena et al., 2011). Examples of lncRNA acting as ceRNAs include *HULC*, *PTCSC3*, *lincMD1*, and *linc-RoR*. The lncRNA *HULC* contains binding sites for miR-372. When over-expressed in the Hep3B hepatocellular carcinoma cell line, *HULC* reduced expression and activity of miR-372. Importantly, miR-372 was down-regulated in all hepatocellular carcinoma patient biopsies compared to matched adjacent healthy tissue, consistent with the observation that *HULC* is one of the most significantly up-regulated transcripts in this cancer (Wang et al., 2010). Drastically down-regulated in thyroid cancers, lncRNA *PTCSC3* significantly decreased the expression of the oncogenic miR-574-5p when re-expressed in three thyroid cancer cell lines, leading to growth inhibition, cell-cycle arrest and increased apoptosis (Fan et al., 2013). *lincMD1* is a muscle-specific lncRNA which controls muscle differentiation in human and mouse. It does so by sequestering miR-133 and miR-135, and effectively regulating the expression of *MAML1* and *MEF2C* transcription factor genes. Notably, when *lincMD1* was over-expressed, increased levels of *MAML1* and *MEF2C* could be reduced in a dose-dependent manner by increasing the levels of miR-133 and miR-135, confirming the direct role of these miRNAs in the observed regulation (Cesana et al., 2011). *linc-RoR* is a highly expressed lncRNA in human ES cells which is down-regulated upon differentiation. It shares miRNA recognition elements with mRNAs of *OCT4*, *SOX2* and *NANOG* transcription factor genes essential for self-renewal. *linc-RoR* was shown to sequester miR-145 and protect its target *OCT4*, *SOX2* and *NANOG* mRNAs transcripts from miR-145-mediated repression. Importantly, this regulatory effect was abolished when the two miR-145-binding sites in

linc-RoR were mutated, supporting the hypothesis that the cross-talk between the lncRNA and mRNAs is miR-145 dependent (Wang et al., 2013b). Examples of this ceRNA mechanism have also been described as acting in pluripotency (Wang et al., 2013b), cell differentiation (Cesana et al., 2011) and tumorigenesis (Karreth et al., 2011; Sumazin et al., 2011). However, it is presently unclear how these often very lowly expressed and unstable transcripts regulate miRNAs and, by extension, their mRNA targets. Some authors have even called the ceRNA model into question on the grounds of molecular stoichiometry (Denzler et al., 2014).

A special class of circular RNAs (circRNAs) that contain multiple miRNA target sites have also been proposed to act as miRNAs sponges in neuronal cells (Hansen et al., 2013; Hansen et al., 2011; Memczak et al., 2013). In contrast to linear ceRNAs, circRNAs appear to be much more stable molecules. As the exploration of cytoplasmic lncRNAs continues, it is almost certain to reveal many more additional levels of gene expression regulation at which these transcripts are involved. It may also be expected that cytoplasmic lncRNAs can function in other cytoplasmic processes, such as enzymatic reactions, subcellular localization and transport and many others.

1.3 Nuclear localised lncRNAs in chromatin mechanisms

As evident from the examples discussed above, nuclear lncRNAs tend to act within chromatin to regulate gene expression and chromatin organization. However, in most cases the nature of their interaction with chromatin remains a mystery. It remains unclear how

some lncRNAs are retained and accumulate at their site of synthesis, while others are delivered to sometimes thousands of specific locations genome-wide.

1.3.1 LncRNA targeting and chromatin interactions

Understanding how lncRNAs are targeted to the genome is crucial if we are to understand how these molecules work to regulate what appears to be every aspect of gene expression. It is difficult to envisage how some transcripts, that have been shown to be present at very low abundance and are unstable, locate their targets in the nuclear space particularly because these are greatly outnumbered by non-specific sequences. Ideas that have been proposed include a "proximity transfer" model in which lncRNAs are guided to genomically distal but spatially proximal loci by the folding of the chromatin fibre itself, as well as a diffusion model where lncRNAs are targeted as a component of ribonucleoprotein particles whose interaction with the genome are expected to resemble that of transcription factors (Engreitz et al., 2013; Jeon and Lee, 2011). In the diffusion model, the sequence-specificity of the ribonucleoprotein targeting can be provided by either the lncRNA or protein components. Indeed, a growing number of lncRNAs have been found to act in conjunction with interacting transcription factor proteins (Hung et al., 2011; Kino et al., 2010; Ng et al., 2013; Rapicavoli et al., 2013; Sun et al., 2013; Vance et al., 2014; Yang et al., 2013). Conceptually, a lncRNA can recognise its target sequence by base-pairing with its cognate DNA (or nascent RNA) according to either canonical Watson-Crick or non-canonical Hoogsten rules (Schmitz et al., 2010). Alternatively, it has been proposed that the three-dimensional folding of lncRNAs may create modular DNA recognition domains akin to those in DNA-binding proteins (Chu et al., 2011). The

"proximity transfer" and diffusion models may not need to be mutually exclusive, and some lncRNAs may navigate within nuclear space using both modes.

In summary, although the details of how lncRNAs are targeted are largely unknown, it appears that their interactions with the genome are complex and are mediated by the interplay between the spatial organization of the genome, lncRNA-protein interactions, and potentially sequence recognition properties of the lncRNAs themselves.

1.3.2 Mechanisms of nuclear lncRNA function

LncRNAs have recently emerged as an important layer in the regulation of various nuclear processes. We now know that lncRNAs can regulate gene expression by altering chromatin state both locally and distally, interacting with and modulating various aspects of effector protein activity, and mediating large scale chromatin folding and sub-nuclear compartmentalization. Due to the chemical versatility of RNA discussed previously, lncRNAs as a class are now being placed at the centre of many appealing models that are invoked to explain aspects of specific yet versatile and dynamic regulation of nuclear processes. In many instances, especially when lncRNAs are involved in the regulation of transcription or chromatin state locally, it is challenging to unambiguously demonstrate if the regulation is conferred by the mature RNA molecule itself, or by the act of transcription and associated enzymatic activities (Kornienko et al., 2013). Below, I only highlight selected studies that support direct regulatory roles for lncRNA transcripts in the nucleus.

Overall, most of the currently available examples suggest that lncRNAs function largely by interacting with specific proteins which in turn can either act on chromatin or the

enzymatic machinery of gene expression, or are themselves directly involved in gene expression regulation (Batista and Chang, 2013; Lee, 2012; Rinn and Chang, 2012). Potential roles of nuclear lncRNAs within ribonucleoprotein complexes include the genomic targeting of the associated proteins, the inhibition of binding of these proteins to the genome, and the direct modulation of their biochemical activity. Chromatin-associated lncRNAs appear to also perform structural roles in long-distance chromatin interactions, as well as other regulatory aspects of nuclear sub-structure.

Among the proteins recruited to chromatin by lncRNAs are both chromatin modifiers and transcriptional regulators. The most extensively studied example of transcriptional regulation at the level of chromatin mediated by a lncRNA is provided by *Xist*. *Xist* is a key player in X chromosome inactivation that occurs in female placental mammals as a mechanism of sex chromosome dosage compensation. X chromosome inactivation results in permanent transcriptional silencing of nearly all genes on one of the copies of the X chromosome. *Xist* is specifically transcribed from and coats the inactive X chromosome (Clemson et al., 1996). Mechanistically, chromatin-associated *Xist* is thought to directly interact with and recruit the Ezh2 catalytic subunit of the Polycomb repressive complex 2 (PRC2) via its Repeat A (RepA) motif (Zhao et al., 2008). Initially, such interaction occurs at the "nucleation centre" to which *Xist* is tethered by the YY1 protein via its Repeat C (RepC) motif (Jeon and Lee, 2011). *Xist* and PRC2 subsequently spread along the whole X chromosome where PRC2 catalyses the deposition of the repressive H3K27me3 mark, leading to transcriptional silencing (Pinter et al., 2012). It has also been suggested that *Xist* may play a role in altering the spatial organization of the inactive X chromosome to facilitate the repositioning of active genes into transcriptionally repressive compartments

(Engreitz et al., 2013). Of note, when expressed from an inducible transgene inserted into one of the copies of chromosome 21 in tri-somic Down syndrome cells, *Xist* was shown to specifically coat that copy of the chromosome and to induce chromosome-wide chromatin modifications and gene silencing (Jiang et al., 2013). Interestingly, *XACT*, a lncRNA recently identified in human cells appears to associate with the active copy of the X chromosome, although its role and mechanisms are still to be determined (Vallot et al., 2013).

LncRNAs can also target chromatin modifiers to alter chromatin state at loci *in trans*. *HOTAIR* was among the first reported examples of this mode of action and has been extensively studied since (Rinn et al., 2007). It is produced from a locus within the human *HOXC* cluster on chromosome 12 and is another lncRNA shown to interact with Ezh2 (Kaneko et al., 2010; Rinn et al., 2007; Tsai et al., 2010; Zhao et al., 2010). Despite originating from within the *HOXC* cluster, *HOTAIR* regulates gene expression from the *HOXD* cluster on chromosome 2. siRNA-mediated depletion of *HOTAIR* results in the reduction of PRC2 occupancy and H3K27me3 as well as transcriptional activation of some of the *HOXD* genes (Rinn et al., 2007). Subsequent studies showed that *HOTAIR* binds to over 800 loci genome-wide, where its occupancy is independent of Ezh2 and has been proposed to be dependent on the DNA recognition properties of the *HOTAIR* transcript itself (Chu et al., 2011). Furthermore, over-expression of *HOTAIR* is sufficient to cause mis-localization of Ezh2 to several ectopic loci, supporting the hypothesis that this lncRNA plays a role in recruiting Ezh2 to chromatin (Gupta et al., 2010).

Human *HOTAIR* also provides one of the first examples of modular lncRNAs. These transcripts can simultaneously interact with different proteins via their distinct structural

RNA domains, thereby bringing different biochemical functionalities together into one complex. This property which may be widespread among lncRNAs may provide a mechanism for dramatically increasing the number of functional protein complexes, by bringing together subunits that would not otherwise be able to interact via protein-protein interactions. In the *HOTAIR* example, the lncRNA interacts with PRC2 via its 5′ domain, while its 3′ domain binds to the H3K4 demethylase LSD1. This results in both repressive complexes the being simultaneously recruited to a subset of loci (Tsai et al., 2010).

Fendrr is a *trans*-acting lncRNA involved in regulating developmental gene expression programs by recruiting chromatin modifiers to chromatin via direct interactions with DNA. *Fendrr* was shown to be important for heart and body wall development, with its mouse knockout being embryonic lethal (Grote et al., 2013). *Fendrr* appears to interact with both PRC2 and the histone methyl-transferase complex MLL. Depletion of *Fendrr* is associated with reduced binding of PRC2 at a number of loci tested. Furthermore, *Fendrr* can directly interact with the DNA of some of these loci *in vitro*, supporting the notion of a RNA-DNA interaction-mediated mechanism of chromatin targeting.

A mechanism whereby a *trans*-acting lncRNA inhibits, rather than facilitates, chromatin binding of their interacting protein is illustrated by transcription factor-associated lncRNAs such as *Gas5*, *PANDA* and *Lethe*. *Gas5* is up-regulated in growth-arrested cells where it binds to the glucocorticoid receptor (GR) and inhibits its binding to GR response elements (GREs), thereby disrupting the induction of GR-mediated transcriptional program (Kino et al., 2010). Likewise, *PANDA* is induced by p53 and has been reported to bind the NF-YA transcription factor (Hung et al., 2011). Depletion of *PANDA* was associated with an increased occupancy of NF-YA and concomitant induction of its pro-apoptotic target

genes. Similarly, *Lethe* binds to the RelA subunit of NK-κB to prevent the whole complex from binding to DNA in mouse embryonic fibroblasts (Rapicavoli et al., 2013).

LncRNAs originating from the *CCND1* promoter region have been shown to allosterically regulate the activity of the regulatory TLS protein. Binding of lncRNAs causes a conformational change resulting in the activation of TLS. Active TLS inhibits histone acetyltransferase activity in the locus, leading to transcriptional down-regulation of the *CCND1* gene and to a repressive chromatin state (Wang et al., 2008).

LncRNAs have also been implicated in the establishment of long-range chromatin interactions. A number of lncRNA loci have been shown to mediate enhancer-promoter looping in an RNA- and Mediator-dependent manner (Lai et al., 2013; Orom et al., 2010). Similar RNA-dependent enhancer-promoter looping has been described at a number of estrogen receptor-regulated loci in MCF-7 human breast cancer cells (Li et al., 2013b).

Additionally, lncRNAs are involved in other aspects of nuclear organization, such as the establishment and function of sub-nuclear bodies. Independent investigations have concluded that *Neat1* is essential for the maintenance of nuclear paraspeckles (Chen and Carmichael, 2009; Clemson et al., 1996; Sunwoo et al., 2009). Furthermore, the expression of *Neat1* from an ectopic locus or tethering of the transcript to a specific region is sufficient to induce co-transcriptional paraspeckle assembly at such loci (Mao et al., 2011; Shevtsov and Dundr, 2011). Expression of *Neat1* is developmentally regulated and correlates with paraspeckle formation (Chen and Carmichael, 2009; Sunwoo et al., 2009), highlighting the role of lncRNAs in nuclear compartmentalisation.

Given the tiny proportion of lncRNAs that have been functionally investigated to date, the examples discussed above likely represent the tip of the iceberg of nuclear mechanisms mediated by lncRNAs.

1.4 LncRNAs in neuronal development and disease

Given the diverse regulatory repertoire of lncRNAs and the enrichment of a large proportion of these transcripts in central nervous system (CNS) transcriptomes, it has been argued that in many species lncRNAs may contribute to the complexity of the nervous system by providing an additional layer of control over the protein cellular machinery (Mehler and Mattick, 2007). Below, I consider examples of lncRNAs performing important roles in neuronal development and neurological function.

It has been noted that lncRNAs loci preferentially occur in the vicinity of either signaling genes or genes encoding for brain-expressed transcription factors with which they are often co-expressed (Ponjavic et al., 2009). Furthermore, genome-wide expression data and *in situ* hybridization studies in the brain suggest that lncRNAs are transcribed in specific cell types and at specific developmental time-points. Additionally, these transcripts are often more region-restricted compared to mRNAs (Mercer et al., 2008; Ramos et al., 2013). Interestingly, many brain-specific lncRNAs appear to be primate- or human-specific and show signatures of positive selection and accelerated evolution, suggesting that they may play important roles in brain development and evolution (Lindblad-Toh et al., 2011; Pollard et al., 2006a; Xu et al., 2010). For example, highly accelerated region 1 (*HAR1A*) lncRNA has evolved rapidly since the divergence of the human lineage and is co-expressed

with reelin, a crucial gene for brain development. Although the function of *HAR1A* is unknown at present, it has been suggested that, similar to reelin, it may be involved in regional forebrain organization (Pollard et al., 2006a; Pollard et al., 2006b; Qureshi and Mehler, 2012). On the other hand, other brain-expressed lncRNAs are highly conserved between birds to mammals and exhibit similar spatial and temporal expression patterns, suggesting that they might perform conserved roles in brain development and function (Chodroff et al., 2010; Dorus et al., 2004).

Several genome-wide transcriptomic studies reported sets of lncRNAs that are up-regulated upon neuronal differentiation, indicating that these transcripts may play roles in this process (Hjelm et al., 2013; Ramos et al., 2013; Talkowski et al., 2012). One study in adult brain also documented the induction of lncRNAs by electrical activity in neurons (Barry et al., 2014). Based on the frequent genomic co-location and spatiotemporal co-expression of lncRNAs with genes encoding key brain developmental regulators, it has been proposed that lncRNAs may be involved in modulating the expression of these genes (Amaral et al., 2009; Ponjavic et al., 2009). Indeed, some lncRNAs originating from the ultraconserved regions (UCRs) associated with brain-specific genes have been demonstrated to regulate these genes by recruiting regulatory proteins to the locus (Bond and Fox, 2009; Feng et al., 2006). For example, *Evf-2* (also known as *Dlx6os1*) originates from a UCR located in the *Dlx-5/6* locus. The *Dlx* genes encode homeodomain transcription factors that are expressed in the developing ventral forebrain and are involved in forebrain and craniofacial development. *Evf-2* regulates *Dlx5*, *Dlx6*, as well as *Gad67* genes via both *cis*- and *trans*-acting mechanisms (Feng et al., 2006). *In cis*, *Dlx6* is regulated by *Evf-2* transcription occurring antisense to it. *In trans*, *Evf-2* RNA recruits two

transcription factors, an activating homeobox protein DLX2 and a repressive methyl-CpG-binding protein MECP2, to modulate the transcription of *Dlx5* and *Gad67* (Bond and Fox, 2009; Feng et al., 2006). In the murine *Evf-2* loss-of-function model, reduced numbers of hippocampal GABAergic interneurons were observed in the early embryo. Although adult mice had normal numbers of these neurons, they displayed synaptic inhibition defects, indicating an importance role for *Evf-2* in neuronal function *in vivo* (Bond and Fox, 2009; Feng et al., 2006).

Evf-2 is one of a class of UCR-associated lncRNAs that constitute a new type of transcriptional regulators (Feng et al., 2006). For example, *Dlx1os1* which is expressed antisense to *Dlx1* also regulates this gene *in cis*. Loss-of-function models of *Dlxos1* exhibit mild skeletal and neurological abnormalities in mice reminiscent of those observed in *Dlx1* gain-of-function models (Kraus et al., 2013).

Specially designed screening studies have identified multiple lncRNAs with roles in neuronal differentiation (Ng et al., 2012). For example, human *lncRNA-N1* and *lncRNA-N3*, which bind both the SUZ12 subunit of PRC2 and the transcriptional repressor complex REST, appear to promote neurogenesis by reinforcing the REST-induced repression of a subset of glial genes following the recruitment of PRC2 to their loci (Ng et al., 2012).

Several lncRNAs have been found to be specifically transcribed in the retina. For example, both the depletion and over-expression of a nuclear intergenic lncRNA *Miat* (also known as *Gomafu*) in the retina of neonatal mice produced defects in the specification of retinal cell types (Rapicavoli et al., 2010). Similar defects of retinal cell specification were observed upon the knockdown and over-expression of *Six3os1*, a lncRNA expressed from

a distal promoter of the homeodomain transcription factor *Six3* (Rapicavoli et al., 2011). Unlike many antisense lncRNAs, *Six3os1* does not regulate *Six3* expression. Instead, it modulates the activity of the SIX3 protein in the developing retina by both binding its co-regulators and also recruiting histone modification enzymes to its targets (Rapicavoli et al., 2011). Two other nuclear lncRNAs, *Tug1* and *Vax2os*, have been implicated in the photoreceptor progenitor cell differentiation via unknown molecular mechanisms (Meola et al., 2012; Young et al., 2005).

LncRNAs have also been implicated in neurological and neurodevelopmental disorders. An example of one such disorder is Prader-Willi syndrome (PWS). PWS patients suffer from intellectual disability, sleep disorders, and psychosis. The disease is caused by a genomic deletion of the 15q11-13 region on the paternally inherited chromosome. This region is maternally imprinted, which means that only the paternal copy is normally expressed. Mapping studies in Prader-Willi syndrome (PWS) patients identified a region containing a nuclear-localised lncRNA (*116HG*) locus and a snRNA *SNORD116* as the smallest deleted region associated with the disease (Sahoo et al., 2008). Functional studies revealed that *116HG* aggregates in the so called RNA "clouds" in the nuclei of mouse brain cells, and these structures appear to be dynamically regulated during brain development (Powell et al., 2013). Also, *116HG* transcript likely interacts with the *UBE3a* locus which is important for neuronal development and is situated immediately upstream of its own. As *116HG* was also found to interact with the RBBP5 subunit of MLL, an activating H3K4 methylation complex, it was proposed that *HG116* may positively regulate *UBE3a* expression via chromatin-based mechanisms (Powell et al., 2013).

1.5 LncRNAs in the *Pou3f3* locus

Genomic co-occurrence and spatiotemporal co-expression of intergenic lncRNA and transcription factor genes is most pronounced during the development of the mouse CNS (Ponjavic et al., 2009). The genomic locus of a transcription factor gene *Pou3f3*, (also known as *Brn1* or *Oct8*), which encodes a class III POU family transcription factor provides an example of such a locus which is particularly relevant to this thesis.

Pou3f3 is a single exon gene whose protein binds to DNA in a sequence-specific manner. *Pou3f3* contributes to both neuronal and kidney development by regulating the proliferation and differentiation of progenitor cells (Nakai et al., 2003). Mouse mutants with homozygous loss of *Pou3f3* die of renal failure within 36 hours *post partum* (Nakai et al., 2003), with severe defects of the hippocampus and forebrain among others (McEvilly et al., 2002). In the developing neocortex, *Pou3f3* is expressed in late neuronal precursors and in migrating neurons and, together with its closely related paralog *Pou3f2*, is required in ventricular zone progenitors for deep-to-upper layer fate transition, sustained neurogenesis and cell migration (Dominguez et al., 2013).

The wider *Pou3f3* locus contains a number of lncRNAs on either side of the protein-coding gene. Recently, a capture sequencing (Capture-Seq) study found that the regions flanking *Pou3f3* are nearly pervasively transcribed in neuronal lineages giving rise to sometimes very long precursor transcripts spanning tens of kilobases of sequence (Ramos et al., 2013). The lncRNA locus I chose to investigate in this thesis, *Dali*, is also encoded in a wider *Pou3f3* locus. However, it is located approximately 50 kb downstream of *Pou3f3* in

the sense orientation and does not overlap previously described lncRNA loci or regulatory elements.

Sauvageau et al. recently generated mouse knockout models for two of the intergenic lncRNAs from this locus, *linc-Brn1a* and *linc-Brn1b* (Sauvageau et al., 2013). *Linc-Brn1b* is an approximately 3 kb long transcript produced from a 6.8 kb locus residing approximately 10 kb downstream of *Pou3f3*. The *linc-Brn1b* transcript localises predominantly to the nucleus, with some of it present in the cytoplasm. *In vivo*, it is expressed by neuronal progenitors in both the dorsal and ventral telencephalon by E13.5. By E15.5, *linc-Brn1b* is expressed by progenitors of both ventricular (VZ) and sub-ventricular (SVZ) zones of the developing cerebral cortex, but by E18.5 its expression becomes restricted to the upper cortical layers. At post-natal day 5, *linc-Brn1b* is expressed in the primary somatosensory and visual areas of the cortex. In the somatosensory cortex, *linc–Brn1b* is detected in the barrel structures of the posteromedial barrel subfield where cortical projection neurons receiving afferent signals from the thalamus reside (Aubert et al., 2007). Overall, the spatiotemporal expression profile of *linc-Brn1b* suggests its possible role in the development of specific types of projection neurons in certain areas of the cortex.

Genomic deletion of the *linc-Brn1b* locus resulted in significant (approximately 50%) expression down-regulation of the upstream *Pou3f3* gene and concomitant up-regulation of *linc-Brn1a* which shares a bi-directional promoter with *Pou3f3*. Genome-wide, a significant down-regulation of cellular growth and proliferation genes and up-regulation of neuronal maturation genes were observed, supporting a role for *linc-Brn1b* in telencephalic neural progenitor differentiation and neurogenesis. *Linc-Brn1b*$^{-/-}$ mice exhibited impaired

proliferation of a specific sub-population of cortical progenitors in the SVZ of the developing cortex. The numbers of mature upper layer cortical neurons which are predominantly derived from this progenitor population were also reduced. *Linc-Brn1b*$^{-/-}$ mice also exhibited abnormalities of cortical lamination and barrel cortex organization, as well as a general growth defect (Sauvageau et al., 2013).

Overall, it appears that *linc–Brn1b* is required for proper generation of specific populations of cortical neurons. However, the observed abnormalities may derive from either the loss of the *linc-Brn1b* RNA transcript or, alternatively, from the deletion of functional DNA elements (Bassett et al., 2014b). In fact, another lncRNA from the *Pou3f3* locus, *linc-Brn1a*, for which a knockout model was also generated by genomic deletion, may serve as an example of the latter. *Linc-Brn1a* is encoded immediately upstream of *Pou3f3* and shares a bi-directional promoter with it. Similar to *linc-Brn1b*, it is expressed in neuronal progenitors. The regulation of *linc-Brn1a* appears to occur in the opposite way to the *Pou3f3* gene, which is down-regulated upon the *linc-Brn1b* deletion, while *linc-Brn1a* is up-regulated (Sauvageau et al., 2013). Therefore, deletion of the approximately 45 kb long *linc-Brn1a* locus immediately upstream of an important developmental transcription factor is lukely to disrupt its expression due to the removal of promoter and other regulatory elements involved in its expression. Consequently, the defects observed in knockout mice cannot be attributed solely to the removal of the non-coding locus. Below, I discuss the implications of this for lncRNA functional studies.

1.6 Experimental approaches to studying nuclear lncRNAs

A growing number of lncRNA loci have been shown to regulate gene expression locally as well as distally. Moreover, they have been demonstrated to do so via both transcription-dependent and RNA-dependent modes of action. Consequently, when undertaking a functional study of a lncRNA, the role of its locus needs to be clearly separated from that of its RNA product. This feature distinguishes lncRNAs from protein-coding loci and requires special care to be exercised when designing and interpreting the results of loss- or gain-of-function experiments.

1.6.1 Loss- and gain-of-function models

Despite recent progress, the field of lncRNAs is still in its early days. As such, the experimental approaches used to study their functions have been either adapted from those used to investigate protein-coding genes or developed anew only recently.

Loss- and gain-of-function studies have been extremely fruitful for uncovering functions of protein-coding genes and regulatory DNA elements. For lncRNAs, if designed and interpreted with care, these approaches have the potential to not only provide clues as to their function but also to dissect the mechanism of action of these loci. The choice of a strategy or a combination of strategies to use ultimately depends on the pre-existing knowledge about the locus, including its chromatin status and the subcellular localisation of its transcript.

LncRNAs that are enriched in the cytoplasm (van Heesch et al., 2014) may be expected to work *in trans* and, therefore, post-transcriptional knockdown strategies (such as shRNA-

mediated knockdown) and (over-)expression from an exogenous construct can represent legitimate loss- and gain-of-function strategies, respectively.

Nuclear lncRNAs may work either *in cis* or *in trans*, or both. Furthermore, both modes of action, in principle, can be accomplished via either RNA-dependent or transcription-dependent RNA-independent mechanisms. In the latter case, the RNA transcript itself is a non-functional by-product of a functional act of transcription. In these cases, the above-mentioned strategies that are relatively straightforward technically can be used first in an exploratory fashion. For nuclear lncRNAs suspected to work *in trans*, RNase H-mediated knockdown can also be effective (Ideue et al., 2009). If no effect on either local or distal genes can be detected, this may indicate a transcription-dependent but RNA-independent role for the locus, which needs to be investigated using more challenging and time-consuming genomic engineering approaches. For example, as a loss-of-function strategy, one can insert transcriptional terminators into the genomic locus of interest to abrogate transcription through it. However, care must be taken to ensure that the underlying locus is preserved as intact as possible, including the integrity and spacing between known or predicted functional DNA elements. Also, one should be mindful of the new regulatory elements that may be inserted, including promoters driving selection genes, *loxP* sites and others, which may also alter local chromatin landscape and gene expression (Rassoulzadegan et al., 2002; Steshina et al., 2006). Fortunately, newly developed custom sequence-specific nuclease technologies provide researchers with more precise tools to address potential confounding issues arising from genomic engineering approaches (reviewed in (Kim and Kim, 2014)). Additionally, the same basic technology has been recently adopted to provide tools to engineer not only loss- but also gain-of-function

models. For example, custom transcription factors can be used to drive transcription from the endogenous locus of interest to study its *cis*-acting roles (Gao et al., 2013; Mercer et al., 2013).

To investigate lncRNA function, full-length lncRNA or, parts of its sequence or control region can be deleted from the genome. Alternatively, promoter inversion can be used to disrupt transcription from the locus. In all cases, the organization of the genomic neighbourhood and the nature of transcription in the locus and nearby (uni- or bi-directional) should be carefully considered (Wu and Sharp, 2013). However, even a carefully thought through deletion removes not only the transcript but also the underlying DNA elements and may also disturb the relationship between the remaining functional elements in its vicinity. Even if these elements are unimportant for lncRNA function, they may be so for the neighbouring genes.

For the reasons discussed above, any genomic deletion experiment needs to be complemented by additional studies such as rescue experiments. For example, lethal (although contrasting) phenotypes were observed in both the genomic deletion (Sauvageau et al., 2013) and transcriptional termination (Grote et al., 2013) experiments on the lncRNA *Fendrr*. Importantly, the lethal phenotype observed in the transcriptional termination experiments was rescued using a single transgenic copy of the *Fendrr* locus, indicating that the RNA product rather than its DNA locus are important for *Fendrr* function (Grote et al., 2013).

Of note, lncRNAs can and have been shown to act via different mechanisms in different contexts. As an example, *Airn* is known to silence its neighbouring *Igf2r* gene in embryonic tissues in a transcription-dependent manner (Latos et al., 2012), while acting on

the more distally located imprinted genes in extraembryonic tissues by recruiting the histone methyltransferase G9a (Nagano et al., 2008). Therefore, to uncover the full extent of even an individual lncRNA function, one must employ multiple approaches in a wide variety of expression contexts.

1.7 Aims of the thesis project

LncRNAs have been implicated in a wide range of cellular functions, including multiple stages of gene expression both in the nucleus and the cytoplasm. However, despite the ever-growing lncRNA catalogs and the abundance of computational predictions, functional information about lncRNAs remains scarce. Therefore, to determine what proportion of lncRNAs is actually functional as well the range of functions and molecular mechanisms they are involved in, it is important to study individual lncRNAs in depth. The aim of this thesis was to provide detailed functional characterisation of a single nuclear intergenic lncRNA *Dali* in mouse neuronal cells in culture.

The specific aims were:

1. To identify likely functional lncRNA candidates from a subset of evolutionarily constrained CNS-specific intergenic lncRNAs encoded in the vicinity of developmental transcription factor genes.
2. To characterise the genomic locus of the selected candidate, *Dali*, and the properties of its transcript.
3. To provide insights into the role of *Dali* in neural differentiation.

4. To determine if *Dali* is involved in regulating its neighbouring genes.
5. To determine if *Dali* regulates other targets *in trans*.
6. To determine the genome-wide binding profile of *Dali* in N2A cells.
7. To identify genes directly regulated by *Dali*.
8. To provide insights into the genomic targeting of *Dali* and to identify its protein binding partners.
9. To provide initial insights into the molecular mechanism whereby *Dali* regulates its target genes.

An over-arching aim of the project was to establish an experimental framework for prioritising lncRNA candidates and for performing a functional study of a nuclear-localised lncRNA.

Chapter 2 Materials and methods

2.1 Materials

2.1.1 Antibodies

Material	Source
BRG1 (ab4081), a rabbit polyclonal antibody against a synthetic peptide conjugated to KLH derived from within residues 1400 - 1500 of human BRG1	Abcam
CTCF (ab70303), a rabbit polyclonal antibody against a synthetic peptide derived from within residues 650 - 700 of human CTCF	Abcam
DNMT1 (ab87656), a rabbit polyclonal antibody against a recombinant fragment derived from within residues 1037-1386 of mouse DNMT1.	Abcam
FLAG, a mouse monoclonal anti-FLAG M2 beads	Sigma-Aldrich
IgG, a normal rabbit IgG antibody	Millipore
ORC3L (MBS421369), goat polyclonal antibody against peptide with sequence QKTDHVARLTWGGC from the C terminus of the human ORC3L	MyBiosource
P66beta (ab76924),), a rabbit polyclonal antibody against a synthetic peptide derived from within residues 1 - 50 of human P66beta	Abcam
POU3F3 (C-17) (sc-6028-R), a rabbit polyclonal antibody against an epitope mapping at the C-terminus of human POU3F3	Santa Cruz
SIN3A (39865), a rabbit polyclonal antibody against a peptide derived from the N-terminus of human SIN3A.	Active Motif

2.1.2 Bacteriology

One Shot TOP10 Chemically Competent *Escherichia coli*	Life Technologies

2.1.3 Cell lines

N2A cells are mouse albino neuroblastoma cells	Sigma-Aldrich
SH-SY5Y are human neuroblastoma cells	Gift from Wade-Martins lab, DPAG
E14 are mouse embryonic stem cells	Gift from Brockdorff lab, Department of Biochemistry

2.1.4 Chemicals and Reagents

4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES) pH7.5 (H3537)	Sigma
4-20% Tris-glycine polyacrylamide gel (456-1093)	BioRad
5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) (V3941)	Promega
Actinomycin D	
Agarose (BP160-500)	Fisher Scientific
Bovine Serum Albumin, 100×, included with restriction enzymes	New England Biolabs
Calf intestinal phosphatase (CIP), (M0290S)	New England Biolabs
Carbenicillin, disodium salt (10177-012)	Life Technologies
Chloroform (C-2437)	Sigma
cOmplete, Mini; Protease Inhibitor Cocktail (11836153001)	Roche
Deoxynucleoside Triphosphate Set (dNTPs), PCR Grade, 10 mM each (1196904001)	Roche
Digitonin, water-soluble (407565000)	Acros Organics
DL-dithiothreitol (DTT) (R0862)	ThermoScientific
Dulbecco's Modified Eagle Medium (DMEM), liquid (high glucose, GlutaMAX™ Supplement, pyruvate) (31966-021)	Life Technologies
Dulbecco's Modified Eagle Medium (DMEM), liquid, high glucose, no glutamine, (EC10) (11960-044)	Life Technologies
Dulbecco's Modified Eagle Medium : Nutrient Mixture F-12 (DMEM / F-12) (31330-038)	Life technologies
ESGRO®Leukemia Inhibitory Factor (ESG1107)	Millipore
Ethanol	Sigma
Ethylendiaminetetraacetic acid (EDTA) (EDS)	Sigma-Aldrich
Ethylene glycol tetraacetic acid (EGTA) (4100-OP)	Calbiochem
Fetal Bovine Serum (FBS)	Invitrogen
Fetal Calf Serum (for ES cells)	BioSera
Formaldehyde, solution, 36.5-38% in water (F8775)	Sigma
FuGENE® 6 Transfection Agent (E2691)	Promega
Gelatin, powder, for cell culture (G9391)	Sigma
GelRed Nucleic Acid Stain (41003)	Biotium
Glasgow Minimum Essential Medium (GMEM), liquid (with sodium bicarbonate, without L-glutamine) (G5154)	Sigma
Glycerol	Sigma
Glycine, powder, for molecular biology (50046)	Sigma
Hygromycin, solution, 0.4 mg/ml (10687-010)	Life Technologies
Igepal, for molecular biology (I8896)	Sigma
Inorganic pyrophosphatase from yeast (EF0221)	Thermo Scientific
Isopropanol	Sigma
Laemmli sample buffer (S3401)	Sigma
Lennox L (LB) Agar (22700-041)	Life Technologies

Lennox L (LB) Broth Base (12780-029)	Life Technologies
L-glutamine (25030-024)	Life Technologies
Lithium chloride (LiCl), solution, 8M (L7026)	Sigma
Low melting temperature agarose	Bioline
Low melting temperature agarose (A9414)	Sigma
Magnesium chloride ($MgCl_2$), solution, 1M (M1028)	Sigma
MyOne streptavidin C1 Dynabeads (65001)	Life Technologies
Non-Essential Amino Acids (11140-035)	Life Technologies
Nonidet P40 (NP-40) (chemically indistinguishable from Igepal, for molecular biology (I8896), no longer available	Sigma
Nucleoside triphosphates, set of four, 100 mM solutions (772451PK)	New England Biolabs
OPTI-MEM medium (31985-062)	Life Technologies
Penicillin/streptomycin (15140-122)	Life Technologies
Phenol, solution (P4457)	Sigma
Phenol:chloroform:isoamyl alcohol (P2069)	Sigma
Phosphate buffered Saline (PBS), no calcium, no magnesium (14190-094)	Life Technologies
Potassium chloride (KCl), solution, 1M, for molecular biology (60142)	Sigma
Protein A/G magnetic beads (88802)	Pierce
Protein G Dynabeads (10003D)	Life Technologies
Proteinase K, solutions (03115887001)	Roche
Retinoic acid (R2625)	Sigma
RNAlater (AM7021)	Life Technologies
RNase-free water (non-DEPC treated) (AM9932)	Life Technologies
RNaseOUT (10777-019)	Life Technologies
RNasIn RNase inhibitor (N2511)	Promega
Salmon sperm DNA, sheared (10 mg/ml) (AM9680)	Life Technologies
Sodium bicarbonate ($NaHCO_3$), powder (S5761)	Sigma
Sodium Deoxycholate (D6750)	Sigma
Sodium dodecyl sulphate (SDS), powder (71727)	Sigma-Aldrich/Fluka
Sodium dodecyl sulphate (SDS), solution, 20%, RNase-free (AM9820)	Life Technologies
Sodium Pyruvate (11360-070)	Life Technologies
ß-mercaptoethanol (31350-010)	Life Technologies
Streptavidin coated Dynabeads M-280 (11205D)	Life Technologies
Sucrose, solution, ~20%, for molecular biology (49163)	Sigma
Super Optimal broth with Catabolic repression (SOC) medium, included with chemically competent cells	Life Technologies
SYBR® Green PCR Mastermix (4309155)	Life Technologies
T4 DNA ligation buffer (M0202L)	New England Biolabs
T4 RNA ligase buffer (EL0021)	Thermo Scientific
Tris(hydroxymethyl)aminomethane hydrochloride (Tris HCl), pH 7.5 (15567-027)	Life Technologies

Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl) pH 8.0 (AM98556)	Life Technologies
Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl), pH 6.5 (20-160)	EMD Millipore
Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl), pH 8.5 (93283)	Sigma-Aldrich/Fluka
Tris-Acetate-EDTA (TAE) buffer (40 mM Tris-Acetate, 1 mM EDTA, pH 8.0) (TAE50X01)	MP Biomedicals
Tris-borate-EDTA (TBE) buffer (T4415)	Sigma
Triton X-100, for molecular biology (T8787)	Sigma,
TRIzol Reagent (15596-026)	Life Technologies
tRNA (AM7119)	Life Technologies
Trypsin, 0.05% (1X), with EDTA 4Na, liquid (25300-054)	Life Technologies
Xfect TM Transfection Reagent for Mouse Embryonic Stem Cells (631320)	Clontech

2.1.5 Enzymes

AmpliTaq® 360 DNA Polymerase (4398891)	Life Technologies
ApaI (R0114S)	New England Biolabs
BfaI (R0568S)	New England Biolabs
BglII (R0144S)	New England Biolabs
BsaI-HF (R3535L)	New England Biolabs
BstUI (R0518S)	New England Biolabs
Calf Intestinal Alkaline Phosphatase (M0290S)	New England Biolabs
DNA Polymerase I Large (Klenow) fragment (M0210S)	New England Biolabs
EpiMark Hot Start *Taq* Polymerase (M0490L)	New England Biolabs
Esp3I (*BsmBI*) (FERER0451)	Thermo Scientific
Herculase II fusion polymerase (600675)	Agilent Technologies
HpyCH4IV (R0619S)	New England Biolabs
Hsp92II (R7161)	Promega
MspI (R0106S)	New England Biolabs
PlasmidSafe ATP-dependent DNase (E3101K)	Epicentre
PmeI (R0560S)	New England Biolabs
RNaseA (DNase-free) (EN0531)	Thermo Scientific
RNaseH (M0297S)	New England Biolabs
T3 RNA polymerase (EP0101)	Thermo Scientific
T4 DNA ligase (M0202S)	New England Biolabs
T4 polynucleotide kinase (M0201S)	New England Biolabs
T4 RNA ligase (EL0021)	Thermo Scientific
T7 DNA Ligase (L602L)	Enzymatics
T7 RNA polymerase (M0251S)	New England Biolabs
TaqI (ER0671)	Thermo Scientific
Tobacco acid pyrophosphatase (TAP) (component of GeneRacer™ Core kit)	Life Technologies

2.1.6 Equipment

Bioruptor sonicator	Diagenode
Centrifuge for tissue culture (Sorvall TC6)	Dupont
Class II Microbiological safety cabinet	Holten
CO_2 Incubator (Heto-Holten CellHouse 170)	Heto Holten
Digital microscope (EVOS®)	Advanced Microscopy Group
Dissecting microscope (MZFLIII)	Leica
Gel imaging system (AlphaImager MultiImage Light Cabinet)	Alpha Innotech
Heat block	Techne
Hemocytometer	Assistent
Heraeus Multifuge 3S-R Cenrifuge	Thermoelectron Corporation
Horizontal gel electrophoresis system	Biorad
Hybridiser	Techne
Microcentrifuge	Eppendorf
Mini vortex (Vortex-Genie 2)	Scientific Industries
NanoDrop ND-1000 Spectrophotometer	NanoDrop
Orbital incubator	Stuart
PCR Machine	Applied Biosystems
Pipetboy	Integra Biosciences
Pipettes	Gilson
Plate reader (FLUOostar OPTIMA)	BMG Labtech
Power bloc	Biorad
Rollers	IBI Scientific
StepOne™ Real-Time PCR system	Applied Biosystems
Syringe needles, gauge 27 (Z192384 Aldrich)	Sigma-Aldrich
Transilluminator	Vilber Lourmet
Tube rotator	Glas-Col, LLC
Vertical gel electrophoresis system	Biorad
Vibrating microtome (VT1000S)	Leica
Water bath	Jencons-PLS

2.1.7 Kits

5′ EndTag Nucleic Acid Labelling system (MB-9001)	Vector laboratories
Affymetrix Hybridization, Wash, and Stain Kit (900720)	Affymetrix
Affymetrix WT Terminal Labelling and Controls Kit (901525)	Affymetrix
Ambion WT Expression Kit (4411973)	Affymetrix
BCA Protein Assay Kit (23225)	Pierce
Click-iT® Nascent RNA Capture Kit (C-10365)	Invitrogen
Dynabeads kilobaseBINDER Kit (60101)	Life Technologies
EZ DNA Methylation-Gold kit (D5005)	Zymo Research
GeneJET Gel Extraction kit (K0692)	Thermo Scientific

GeneRacer™ Core kit (450168)	Life Technologies
Magna RIP™ RNA-Binding Protein Immunoprecipitation Kit (17-700)	Millipore
Omniscript RT kit (205110)	QIAGEN
Platinum Taq DNA Polymerase High-Fidelity PCR kit (11304011)	Life Technologies
QIAGEN Mini RNeasy kit (74106)	QIAGEN
QIAGEN QIAquick PCR purification kit (28104)	QIAGEN
QIAprep Spin Miniprep kit (27104)	QIAGEN
QIAquick HiSpeed Plasmid Midi Kit (12643)	QIAGEN
QIAQuick PCR purification columns (28104)	QIAGEN
QuantiTect Reverse Transcription Kit (205313)	QIAGEN
SuperScript™ III RT Module (450167)	Life Technologies
TOPO TA Cloning Kit for Sequencing (K4575-01)	Life Technologies
TURBO™ DNase I kit (AM2238)	Life Technologies

2.1.8 Miscellaneous

Affymetrix GeneChip Mouse Gene 1.0 ST Array (901169)	Affymetrix
Mouse Total RNA Master Panel (636644)	Clontech
RP23-92N4	C.H.O.R.I.
Total RNA from human fetal brain (636526)	Clontech

2.1.9 Molecular weight ladders

1 kb DNA ladder (N3232L)	New England Biolabs
100 bp DNA ladder (N3231L)	New England Biolabs
Full Range Molecular Weight Marker Rainbow (RPN800E)	GE Healthcare

2.1.10 Plasmids

N-terminal pFLAG-CMV-6a	Sigma-Aldrich
pBS-U6-GFP	Gift from Dr Shirin Bonni (University of Calgary)
pcDNA3	Life Technologies
pCR4-TOPO (K4575-02)	Life Technologies
Pou3f3 (NM_008900) Mouse cDNA clone in pCMV Entry vector	Origene
pTK-Hyg (6153-1)	Clontech
TALE Toolbox kit (TALEN Kit #1000000019)	Addgene

2.1.11 Plastics

14-ml BD Falcon polypropylene round-bottom tubes	BD Falcon
Cell culture cluster 6 well/12 well/24 well/96 well, flat bottom, with lid	Corning
Cell culture dishes, 10 cm/15 cm	BD Falcon
Cell culture flasks 75 cm^2 polystyrene, angled neck, sterile	Corning
Centrifuge tube 15 ml, CentriStar Cap, RNase/DNase-free	Corning
Conical tube 50 ml, polypropylene	BD Falcon
Cryogenic vial, Polypropylene, Self Standing, Silicone Washer	Corning
Disposable cell scrapers	Fisherbrand
Filter tips (sterile)	Axygen
MicroAmp® Optical 96-Well Reaction Plate with Barcode	Applied Biosystems
Microcentrifuge Safe-Lock tubes 2 ml	Eppendorf
Microcentrifuge tubes 0.5 ml/1.7 ml	Axygen
Petri dish, 90 mm, bacteriological	Fisher Scientific
Pipette tips	TipOne
Serological pipettes 5 ml/10 ml/ 25 ml	Sigma
Thin-walled strips of 8 x 0.2 ml microtubes and domed caps for PCR	Fisher Scientific

2.1.12 Primers

CHART-seq C-oligos	ATDBio
GeneRacer™ RNA Oligo (component of GeneRacer™ Core kit)	Life Technologies
GeneRacer™ RNA Oligo Oligo(dT) primer (component of SuperScript™ III RT Module)	Life Technologies
PCR, qPCR and sequencing primers	Sigma
RT primer mix (component of QuantiTec QuantiTect Reverse Transcription Kit)	Life Technologies

2.1.13 Software

AlphaEase FC	Alpha Innotech
BiQ Analyzer HT	
MethPrime	
ND-1000 V3.6.0	NanoDrop
StepOne Software v2.1	Applied Biosystems
TAL Effector Nucleotide Targeter 2.0	
Triplexator	

2.2. Methods

2.2.1 Cell manipulation

2.2.1.1 Cell culture

Mouse neuroblastoma (N2A) cells were grown in Dulbecco's Modified Eagle Medium (DMEM; (Life Technologies, #31966-021)) supplemented with 10% Fetal Bovine Serum (FBS) and penicillin/streptomycin (Life Technologies, #15140-122) at 37 °C in a humidified atmosphere with 5% CO_2. The cells were grown to 80-90% confluency between passages (every 3 days).

Mouse embryonic stem cells (E14) were grown in ES cell medium (Glasgow Modified Eagle Medium (GMEM; (Sigma, #G5154)) + 15% Fetal Calf Serum (FCS) + 1 mM Sodium Pyruvate (Life Technologies, #11360-070) +100 μM Non-Essential Amino Acids (NEAA) (Life Technologies , #11140-035) + 1000 U/ml ESGRO® Leukemia Inhibitory Factor (LIF) (Millipore, #ESG1107) + 100 μM ß-mercaptoethanol (Life Technologies, #31350-010) + 50 μg/ml each penicillin-streptomycin) at 37 °C in a humidified atmosphere with 5% CO_2 on gelatinized dishes to 60-80% confluency. Cells were split 1 in 5 every 3 days (seeding cells at approximately 1.5-4.0 x 10^5 cells/cm^2).

Human neuroblastoma (SH-SY5Y) cells were grown in DMEM/F12 medium (Life Technologies, #31330-038) supplemented with 10% FBS, 1% penicillin-streptomycin, and 1% L-glutamine (Life Technologies, #25030-024) at 37 °C in a humidified atmosphere with 5% CO_2. The cells were grown to 80-90% confluency between passages (every 5 days).

2.2.1.2 Retinoic acid differentiation of ESCs

E14 mouse embryonic stem cells (ESCs) were plated in ES cell medium (DMEM/15% FCS + 1 mM Sodium Pyruvate + 100 μM NEAA + 1000 U/ml LIF + 100 μM ß-mercaptoethanol + 50 μg/ml each penicillin-streptomycin) on gelatinized dishes at a density of 4-5 x 10^4 cells/cm^2. The next day, cells were washed in DMEM medium twice and PBS once, and differentiation medium (EC10 (Life Technologies, #11960-044) + 10^{-7} M retinoic acid (Sigma, #R2625)) added. A sufficient number of dishes was treated in parallel to allow cells to be harvested every day during the course of the experiment (normally, 6-10 days). Each day, fresh differentiation medium was added.

Table 2.1 Mouse ES cell differentiation oligos

Embryonic sheet markers
qPCR oligos

Oligo name	**Oligo sequence (5′ to 3′)**
Fgf5F1	GGTTTGGTTTTGTTGTTTTGG
Fgf5R1	GCCCCACTTTACTGTTGTGTC
brachF1	CAGCCCACCTACTGGCTCTA
brachR1	GAGCCTGGGGTGATGGTA
Sox17F1	TTCCTATTTCCCCAAGAGGTC
Sox17R1	GCTTCTCTGCCAAGGTCAAC
CdxF1	CTGCCACACTTGGGCTCT
CdxR1	CTTGGCTCTGCGGTTCTG
Nkx2-9F1	ATGCGAAGCCTAGGGTGA
Nkx2-9R1	TCTGCAGGGCTTGTCTCC

***Dnmt1* qPCR oligos**

Dnmt1F	CAGAGACTCCCGAGGACAGA
Dnmt1R	TTTACGTGTCGTTTTTCGTCTC

2.2.1.3 Retinoic acid differentiation of N2A cells

N2A cells were plated at a density 1.5 x 10^5 cells/well in a 6-well plate to allow for neurite outgrowth. After 24 h, growth medium was replaced with differentiation medium containing 2% FBS and 20 μM all-trans retinoic acid. Cells were differentiated for 48-72 h. Differentiation induced morphological changes were maximal after 72 h.

2.2.1.4 Transfection

1.0 x 10^5 N2A cells were seeded into wells of a 6-well plate in 2 mL of medium and transfected with shRNA-expressing constructs (or scrambled control) the next day. Generally, 1.5 μg of DNA was added to a reaction containing 4.5 μl of FuGENE® 6 Transfection Agent (Promega, #E2691) in 100 μl OPTI-MEM medium (Life Technologies, #31985-062) prepared according to the manufacturer's instructions. The complexes were allowed to form for 30 min, and applied to cells. Cells were harvested 72 hours after transfection.

Mouse E14 ES cells were transfected in 6-well format with Xfect TM Transfection Reagent for Mouse Embryonic Stem Cells (Clontech, #631320) according to the manufacturer's instructions.

2.2.1.5 Establishment of stable N2A cell lines

N2A cells lines with stably integrated shRNA constructs targeting *Dali* transcript were generated by co-transfecting shRNA-expressing vectors with hygromycin resistance plasmid pTK-Hyg (Clontech, #6153-1), at a ratio of 5:1, as described above. The next day, the standard media was replaced with media containing hygromycin (0.4 mg/ml; Life Technologies, #10687-010). Selection was maintained until individual resistant colonies became prominent. Individual clones were isolated and screened for the most efficient reduction of *Dali* expression. Selected clones were then maintained in media containing hygromycin (0.2 mg/ml).

2.2.1.6 Cell fractionation

Approximately $2.5x10^6$ cells were harvested, washed twice with ice-cold PBS, re-suspended in 250 µl Lysis Buffer (15 mM HEPES pH7.5, 10 mM KCl, 5 mM $MgCl_2$, 0.1 mM EDTA, 0.5 mM EGTA, 250 mM Sucrose, 0.4% Igepal, 1 mM DTT, 40 U/ml RNaseOUT (Life Technologies, #10777-019), protease inhibitor cocktail (Roche, #11836153001)), and incubated on ice for 20 min. Nuclei were pelleted at 2,000 x g for 10 min at 4 °C. The supernatant was collected as the cytoplasmic fraction. Nuclei were then re-suspended in 50 µl Nuclei Lysis Buffer (10 mM HEPES pH7.5, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM DTT, 40 U/ml RNaseOUT, protease inhibitor cocktail) and incubated on ice for 5 min. Nuclei were centrifuged at 17,000 x g for 5 min at 4 °C. The supernatant was removed as the nucleoplasm fraction. The pellet was re-suspended in 50 µl Salt Extraction Buffer (25 mM HEPES pH7.5, 10% glycerol, 420 mM NaCl, 5 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT, 40 U/ml RNaseOUT, protease inhibitor cocktail) and incubated for 30

min at 4 °C with rotation. The sample was centrifuged at 17,000 x g for 20 min at 4 °C. The supernatant was collected as the salt extracted fraction and the pellet re-suspended in 50 µl Salt Extraction Buffer to obtain the chromatin fraction. RNA was isolated from each fraction using the TRIzol method.

2.2.1.7 Transformation of competent bacterial cells

TOP10 chemically competent cells (Life Technologies) were used for the propagation of plasmid DNA.

TOP10 cells were thawed on ice. Typically, 25 µl bacterial suspension was used for each transformation. 1 µl of plasmid DNA, 2 µl of TOPO cloning reaction or 3 µl of inactivated ligation reaction was added to the cells and incubated on ice for 30 min. Cells then were heat-shocked for 30 sec at 42 °C, and immediately put back on ice for 2 min. 125 µl of room temperature Super Optimal broth with Catabolic repression (SOC) medium was added to the cells. The cells were incubated for 1 h at 37 °C with shaking. The cell suspension was spread on LB-agar plates (1% bactotryptone, 0.5% yeast extract, 1% NaCl, 1.5% agar) containing 100 µg/ml carbenicillin. For transformation of TOPO cloning reactions 50 mg/ml X-gal in N,N'-dimethylformamide was added to a final concentration of 40 ug/ml. Plates were incubated overnight at 37 °C.

2.2.1.8 Quantification of neurite outgrowth

Control and *Dali* stable knockdown cell lines treated with 20 µM retinoic acid for 72 h (as described above) were imaged using bright field microscopy. Cells with one or more neurites of length greater than twice the cell body diameter were scored as positive. A total

of at least 500-600 cells over at least three different fields of view were counted for each cell line.

2.2.2 DNA manipulation

2.2.2.1 Genomic DNA isolation

Prior to genomic DNA isolation, adherent cells were harvested and rinsed with ice cold PBS twice. Cells were re-suspended at approximately 1.0 x 10^6 cells per ml in 10 mM Tris HCl pH 8.0, 10 mM EDTA. To the samples, SDS and proteinase K solutions (Roche, #03115887001) were added to a final concentration 0.5% and 200 μg/ml, respectively. Samples were pipetted several times and incubated at 55 °C for 2 h to ensure that proteins are digested. Following the incubation, NaCl solution was added to a final concentration 0.2 M. Extraction with equal volume of phenol:chloroform was performed twice. To the aqueous phase, an RNase A (DNase-free) (Thermo Scientific, #EN0531) solution was added to a final concentration 25 μg/ml. Samples were incubated at 37 °C for 1 h to digest RNA, followed by two phenol:chloroform extractions. DNA in the aqueous phase was precipitated by adding 1.5 volumes of absolute ethanol. Samples were incubated at -20 °C for at least 20 min, and centrifuged at ≥ 10,000 rpm for at least 5 min at 4 °C. Pellets were washed with 70% ethanol and centrifuged again. Residual ethanol was removed completely. Pellets were air-dried and re-suspended in 1x TE buffer (10 mM Tris HCl pH8.0, 1 mM EDTA).

2.2.2.2 Rapid Amplification of cDNA Ends (RACE)

5′ and 3′ ends of full-length protein-coding mRNAs and lncRNAs were obtained using GeneRacer™ Core kit (Invitrogen, #450168) according to the manufacturer's protocol.
The protocol is based on RNA ligase-mediated and oligo-capping rapid amplification of cDNA ends (RLM-RACE) methods that achieve selective ligation of an RNA oligo to the 5′ ends of decapped RNA using T4 RNA ligase.

Briefly, to obtain the 5′ end of a transcript of interest, 5′ phosphates were removed from truncated or non-capped transcripts in the total RNA sample using calf intestinal phosphatase (CIP) to ensure that GeneRacer™ RNA Oligo was later ligated only to full length capped transcripts. Dephosphorylated RNA was treated with tobacco acid pyrophosphatase (TAP) to remove the 5′ cap from full-length capped molecules, leaving 5′ phosphate required for ligation to the GeneRacer™ RNA Oligo. The GeneRacer™ RNA Oligo was ligated to the 5′ end of transcripts using T4 RNA ligase. Ligated transcripts were reverse transcribed using SuperScript™ III RT Module (Life Technologies, #450167) and the GeneRacer™ Oligo dT Primer (if both 5′ and 3′ RACE was performed using the same sample), or random primers (if only the 5′ ends were of interest). Following reverse transcription, the first-strand cDNA was amplified using a reverse gene-specific primer and the GeneRacer™ 5′ Primer homologous to the GeneRacer™ RNA Oligo. If needed, additional round of PCR was performed using nested primers.

To obtain 3′ ends, first-strand cDNA was amplified using a forward gene-specific primer and the GeneRacer™ 3′ Primer homologous to the GeneRacer™ Oligo dT Primer. If needed, an additional round of PCR was performed using nested primers. If only 3′ ends were of interest, the first three steps in the protocol above (de-phosphorylation, de-capping

and ligation of the GeneRacer™ RNA Oligo) were omitted, starting the procedure from the reverse transcription stage.

Table 2.2 RACE oligos

5′ RACE (1° PCR)	
Oligo name	**Oligo sequence (5′ to 3′)**
8b 5RACE 2 (561)	TCCATTTACAAACAGCCTGGCCCATTCC
Linc45RACE2R	CACCCCTCCCCAAATGAGGAAGGAAAAA
Linc45RACE2R	TTGGGGAGTGAGAGGCAAGGATCATTCA
5′ RACE (2° PCR)	
8b 5RACE 1 (114)	ATTGGGGGAGGGGAAGCCAGCAAAG
Linc45RACE1R	CCAAACCACCTGGGAAGGAGCCTGTTTT
Linc55RACE1R	CCATGGTGGCAAGCTCAGAGGCTACAAG
3′ RACE (1° PCR)	
8b 3RACE 2 (1310)	CCCCATCTACCCAATCACCTTCCCCTGA
Linc43RACE1L	AAGCTCAGCACCACAGACAACCTCAACC
Linc83RACE2L	CCCACCAGCCGCAGTGAACTCCATTT
Linc53RACE2L	AGGGTTGGACCTCCGTTTAGAGTGCTCA
Pou3f33RACE2L	CGCACTTCCCTTGACTTCTGCCCTCAAT
Mitf3RACE1L	GGCTTGTAGCCTCTGAGCTTGCCACCAT
Meis13RACE1L	CTAGCCTGGGGCTTGAATTTGCATGTCT
3′ RACE (2° PCR)	
8b 3RACE 1 (1530)	TGCTTCTTCCTGGCCTCCAGTGTTTGC
Linc43RACE2L	GAGGAGGAAAAGGGGACAGAGGGAGAGA
Linc83RACE1L	CACTGCCCTGGCTGCAGGTTGAACA

Linc53RACE1L	CCCCGCTGTTGGGTACAGCAATACTTTC
Pou3f33RACE1L	ACTTCTGCCCTCAATGCGTTGCCAACTT
Mitf3RACE2L	CTCAGGGTTGGACCTCCGTTTAGAGTGC
Meis13RACE2L	CGGGAAGGATGGGAAAAACTGCAGTCAC

2.2.2.3 Circular Rapid Amplification of cDNA Ends (cRACE)

To map 5′ and 3′ ends of uncharacterized human transcripts orthologous to mouse *Dali*, total RNA from human fetal brain (Clontech, #636526) was subject to cRACE. Briefly, RNA sample was treated with TURBO^{TM} DNase I kit (Life Technologies, #AM2238) according to the manufacturer's instructions. RNA sample (75 ng/µl) was then de-capped using 0.1 U of yeast inorganic pyrophosphatase (Thermo Scientific, #EF0221) in 1x T4 RNA ligase buffer (Thermo Scientific, EL0021) in the presence of RNasIn RNase inhibitor (Promega, #N2511) for 30 min at 37 °C (total reaction volume 20 µl). To de-capped samples, 1 µl of T4 RNA ligase (Thermo Scientific, #EL0021) was added. Samples were incubated at 37 °C for 30 min, followed by 10 min at 65 °C to inactivate the enzymes. 6 µl of circularized RNA sample (450 ng) was used for reverse transcription using Omniscript RT kit (QIAGEN, #205110). Briefly, 1 µl of RNase-free water, 1 µl of 10x Omniscript RT buffer and 2 µl of 10 mM primer mix was added to 6 µl of circularized RNA sample. The sample was incubated at 70 °C for 10 min and then plunged on ice. 8 µl of RT mix (2.9 µl water, 1.8 µl 5mM dNTPs, 0.8 µl 10x buffer, 0.2 µl RNasIn, 0.5 µl RT, 1.8 µl 10x Actinomycin D) was added. Reverse transcription was performed at 42 °C for 30 min, following by incubation at 65 °C for 5 min to inactivate the enzyme. The sample was then diluted 10-fold by adding 162 µl of water. 10 µl of the diluted sample was used in the 30 µl first round PCR reaction performed with Platinum Taq DNA Polymerase High-Fidelity

PCR kit (Life Technologies, #11304011) and a pair of divergent primers designed to cross the anticipated ligation point of 5′ and 3′ ends (based on the pre-existing sequence tags in the region). The first round PCR reaction was diluted 1:10 with water. 10 µl was used for the second round RCR using nested primers. Product was purified using QIAQuick PCR purification columns (QIAGEN, #28104), TOPO-cloned into pCR4-TOPO vector (Life Technologies, #K4575-02), sequenced conventionally and mapped to the predicted region orthologous to mouse *Dali* using the UCSC Genome Browser (BLAT).

Table 2.3 cRACE oligos

Oligo name	**Oligo sequence (5′ to 3′)**
hlinc81Rn	ATGCCCGGCCCATGTAGCACATTTT
hlinc81R1	GTGCCTGCCACCACACCTGGCTATT
hlinc81L1	CCCAAGGAGTCCTGGACCCAAATGC
hlinc81Ln	CAGCCTGGCCAAAATGGCAAAATCC
hlinc82Rn	GCCCAAAGAAGACATGGGCTGAAGAGC
hlinc82R1	CCCCAGAGAAAGTTGAGCAGCCCAAAG
hlinc82Ln	GGGCTGCTCAACTTTCTCTGGGGTCA
hlinc82L1	TGGGGTTCAAGCTCTTCAGCCCATGT
hlinc83Rn	CTCCAGACCGCTGGGACCATGGATA
hlinc83R1	CAGCGGCAAGCACCAGGTGAGACTA
hlinc83Ln	GGCTCCCACAGCCACAGTCAACTCC
hlinc83L1	TGGCCTCTGTGTCTGTGCCTGCAAT
hlinc84Rn	CAAGGGGCCCCGTTATCACCAATGT
hlinc84R1	CAATGACAGCCCTGGAAAGTGAACCA
hlinc84L1	CGGGGCCCCTTGACTTGCTATTGAC

2.2.2.4 Chromatin Conformation Capture (3C)

The 3C assay was performed as described previously (Hagege et al., 2007). E14 ES cells were seeded on gelatin-coated 15 cm^2 culture dishes at a density of 4-5 x 10^4 cells per cm^2 and either differentiated with retinoic acid as described above for 96 h, or grown to 80-90 % confluency. Cells were then washed twice with PBS and trypsinised with 2.5 ml of 1x 0.05% trypsin-EDTA (Life Technologies, #25300-054) at 37 °C. 10 ml of growth medium was then added to neutralise the trypsin. Cells from two 15 cm^2 dishes were collected into one sterile 50 ml falcon tube (to be used per 3C reaction) and spun down at 1000 rpm for 5 min at room temperature. The cell pellet was re-suspended in 50 ml growth medium. At this stage samples for RT-qPCR analysis of *Dali* and *Pou3f3* expression were collected. Next, the single cell suspension was treated with 2% formaldehyde for 5 min at room temperature with mixing to cross-link chromatin interactions. The cross-linking reaction was quenched with 0.125 M glycine. Cells were then centrifuged at 3500 rpm for 15 min and washed twice with ice-cold PBS. Samples were stored at – 80 °C until further use.

Cross-linked cells were lysed in 50 ml ice cold lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% NP-40, 1x protease inhibitor cocktail) for 90 min at 4 °C, while rotating. Next, cell nuclei were spun down at 2500 rpm for 15 min at 4 °C. Nuclei were re-suspended in 2 ml of 1.2x restriction buffer (NEB3 for *Bgl*II). SDS was added to a final concentration of 0.3% and samples were incubated for 1 h at 37 °C with shaking. Triton X-100 was then added (to sequester SDS) to a final concentration of 1.8% and samples were incubated for 1 h at 37 °C with shaking. Nuclei were counted using a hemocytometer. An aliquot containing 1 x 10^6 nuclei (approximately 15 μg of chromatin) was digested with

400 units of *Bgl*II restriction enzyme (50000 U/ml, 8 µl) overnight at 37 °C with shaking. The following day, SDS was added to a final concentration of 1.6% and samples were shaken for 20-25 min at 65 °C to inactivate the restriction enzyme. Nuclei were diluted in 6.125 ml 1.15x T4 DNA ligation buffer (NEB, # M0202L), then Triton X-100 was added to a final concentration of 1%, and the samples were incubated for 1 h at 37 °C with shaking. Next, the digested chromatin was ligated using 100 Weiss units of T4 DNA ligase for 4 h at 16 °C. Cross-links were then reversed using 300 µg Proteinase K overnight at 65 °C. The following day, 300 µg RNase A was added and samples were incubated for 30-45 min at 37 °C. Finally, DNA was phenol-chloroform extracted and ethanol precipitated. DNA concentration was determined semi-quantitatively by Spectrophotometer, operated using the ND-1000 V3.6.0 software.

As a control template for the 3C assay, an RP23-92N4 (CHORI; BACPAC) Bacterial Artificial Chromosome (BAC) clone covering the *Pou3f3-Dali* locus was used. 30 µg BAC DNA was digested with 300 units *Bgl*II restriction enzyme overnight at 37 °C. Digested DNA was phenol-chloroform extracted and ethanol precipitated. Digested BAC clone was ligated at 250 ng/µl using T4 DNA ligase. The ligation products were phenol-chloroform extracted and ethanol precipitated prior to analysis. Ligation products of 3C and BAC samples were quantified using standard SYBR® Green qPCR techniques (Chapter 2.2.3.5). The standard PCR was carried out using a StepOne™ Real-Time PCR system. The primers used for the analysis are shown in Table 2.4. The primers were 20-22 bp long with a melting temperature between 50-54 °C, and were located within 70-100 bp from *Bgl*II restriction sites in the *Pou3f3-Dali* locus. The conditions for PCR analysis were determined using serial dilutions of 3C and BAC control samples. Efficiencies of primer

pairs were tested using standard curves that cover the same range of amplification as the 3C samples. Standard curves were generated by analysing serial dilutions (5 steps, 1:5 step, 10 ng/µl starting concentration) of BAC clone template mixed with *Bgl*II digested genomic DNA (to mimic the complexity of the 3C reactions). The internal glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*) control was prepared from genomic DNA which was digested and precipitated in the same manner as the BAC clone sample. PCR reactions consisted of 300 ng 3C sample, 0.2 µM test primers and a primer corresponding to the *Dali* promoter (only primer pairs that passed an efficiency test were included), and 1x SYBR® Green PCR Mastermix (Life Technologies, #4309155), in a total volume of 20 µl. The PCR conditions were 94 °C for 20 sec (initial denaturation), 40 thermo-cycles of 94 °C for 1 sec (denaturation), 60 °C for 20 sec (annealing and extension), followed by a melting curve analysis. All reactions were performed in triplicate. The mean threshold cycle (Ct) value was calculated (any outlying Ct value deviating from the other two by 1 Ct in the triplicate was eliminated from calculations) and used to calculate relative amounts of PCR products. To normalise for different primer efficiencies, interaction frequencies were calculated by dividing the amount of PCR product obtained with the 3C sample by the amount of DNA obtained with control BAC DNA. Interaction frequencies were also normalised to the *Gapdh* internal controls.

Table 2.4 3C oligos

Oligo name	**Oligo sequence (5′ to 3′)**
PouR1	CAAGGAGCTGTACAACCACAG
PouR3	AAGATGGAACACAGATTTCATTG

PouR4	CGGGGTGAAAAGATACCATC
PouR5	TTTTGATTTGTGCTCGTCTGA
PouR6	TCAAAAACGTACTCTGCTACCC
PouR7	AAGACTTCGGAGCTAGTTTTGG
PouR9	GGCTTCTCTGCAGCCTAGAT
PouR13	TCTTCAGACACCCCAGAAGAG
PouR16	GCAGACAAAGTGACAAGTCAGA
PouR19	TCCCTCTTGTTAGAACCCAGAT
PouR20	CACTCAACGTGGTTTATGTGTT
PouR21	GGGCATTTTAGTAATCGGATAAT
PouR22	TGGTTTCTCCTTATCTTCAGGAC
PouR23	AGGTGACTTGGTGGTCCAGT
PouR24	CAGAGAGAGTTCCAAAACAACC
GAPDH 1	GTTTCCATAGGACCTGCTGCG
GAPDH 2	GTTTTACACTGGGCACTTGAGGTC

2.2.2.5 TOPO-cloning

DNA fragments were inserted into the pCR4-TOPO vector as described in the TOPO TA Cloning Kit for Sequencing (Life Technologies, #K4575-01) manufacturer's instructions. Briefly, the TOPO-cloning reactions consisted of 1 to 4 µl of PCR product, 1ul of salt solution and 1 µl of TOPO vector, in a total volume of 6 µl. Reactions were incubated for 5 min at room temperature and 2 µl of the reactions were used to transform chemically competent bacterial cells as described in Chapter 2.2.1.7.

2.2.2.6 Restriction enzyme digestion

Generally, 3 units of restriction enzymes were used to digest 1 µg DNA in the presence of the recommended 1x restriction buffer in a total volume of 30 µl. The reactions were incubated for at least 1 h at 37 °C (or other enzyme-specific recommended temperature). Restriction digest reactions were supplemented with 0.1 mg/ml BSA, if required.

2.2.2.7 Blunt-ending linearized plasmid DNA

pBS-U6-GFP plasmid DNA was linearized using the *Apa*I (NEB, #R0114S) restriction enzyme in the presence of 1x NEB4 buffer and 1x BSA at 25 °C for 2 h. The reaction was heat-inactivated at 65 °C for 20 min. Inactivated reaction (30 µl) was supplemented with 1.5 µl of 33 µM each dNTPs, 4.5 µl NEB2 buffer, 1 µl DNA Polymerase I Large (Klenow) fragment (NEB, #M0210S), 8 µl water to a total volume of 45 µl. The reaction was incubated for 30 min at room temperature and inactivated by adding 0.9 µl 10 mM EDTA. The reaction was purified using QIAGEN QIAquick PCR purification kit (as described in Chapter 2.2.2.11).

2.2.2.8 Dephosphorylation of linearised plasmid DNA

Prior to ligation, 5′ termini of linearised plasmid DNA were dephosphorylated to prevent vector re-ligation. Typically, 1.5 µl of Calf Intestinal Alkaline Phosphatase (NEB, #M0290S) was added to 40 µl inactivated restriction digestion reactions. The samples were incubated at 37 °C for 2 h. Plasmid DNA was purified using QIAGEN QIAquick PCR Purification kit as described in Chapter 2.2.2.11.

2.2.2.9 Oligonucleotide annealing and phosphorylation for shRNA expression construct generation

10 µl each of 100 µM Forward and Reverse high performance liquid chromatography-purified oligonucleotides (Table 2.5) were mixed with 2.2 µl of 10x Annealing buffer (100 mM Tris, pH 7.5–8.0, 0.5 M NaCl, 10 mM EDTA). Samples were incubated at 95 °C for 15 min in a thermocycler. Following the incubation, the thermocycler was switched off, and samples were allowed to cool for 1 h. Samples were then diluted 1:100 in water. To phosphorylate annealed oligos, 10 µl of diluted samples were incubated with 1 µl of T4 polynucleotide kinase (NEB, #M0201S) in the presence of 1x T4 polynucleotide kinase buffer in a total volume of 20 µl at 37 °C for 1h. Samples were heat-inactivated at 65 °C for 20 min and purified using QIAGEN QIAquick PCR Purification kit (as described in Chapter 2.2.2.11). 1 µl of each annealed sample was run on a 4% low melting temperature agarose gel as described in Chapter 2.2.2.17. 1 µl of 1 µM unannealed oligo was included to observe a shift in size of annealed oligos.

Table 2.5 shRNA oligos

***Pou3f3* KD**

Oligo name	**Oligo sequence (5′ to 3′)**
Pou3f3sh1941f	AAGGGTGAAGATGCCTAATTCAAGAGACTAGGCATCTTCACCCTTC TTTTTC
Pou3f3sh1941r	TCGAGAAAAAGAAGGGTGAAGATGCCTAGTCTCTTGAATTAGGCA TCTTCACCCTT
Pou3f3sh2687f	CGGATGTGATTATGTAATTTCAAGAGACTTACATAATCACATCCGC TTTTTC

Pou3f3sh2687r	TCGAGAAAAAGCGGATGTGATTATGTAAGTCTCTTGAAATTACATAATCACATCCG
Pou3f3sh2765f	CAATATAGAATCGTTAGATTCAAGAGACCTAACGATTCTATATTGCTTTTTC
Pou3f3sh2765r	TCGAGAAAAAGCAATATAGAATCGTTAGGTCTCTTGAATCTAACGATTCTATATTG
Pou3f3sh3012f	AGTAGTGATTCCGTAAGATTCAAGAGACCTTACGGAATCACTACTCTTTTTC
Pou3f3sh3012r	TCGAGAAAAAGAGTAGTGATTCCGTAAGGTCTCTTGAATCTTACGGAATCACTACT

***Dali* KD**

Linc8sh74For	GCTTGGATTCAGTTCCTATTCAAGAGACAGGAACTGAATCCAAGCCTTTTTC
Linc8sh74Rev	TCGAGAAAAAGGCTTGGATTCAGTTCCTGTCTCTTGAATAGGAACTGAATCCAAGC
Linc8sh862For	AGCTACTGATGAGACATATTCAAGAGACATGTCTCATCAGTAGCTCTTTTTC
Linc8sh862Rev	TCGAGAAAAAGAGCTACTGATGAGACATGTCTCTTGAATATGTCTCATCAGTAGCT
Linc8sh906For	ATCAACACTCATCAAGAATTCAAGAGACTCTTGATGAGTGTTGATCTTTTTC
Linc8sh906Rev	TCGAGAAAAAGATCAACACTCATCAAGAGTCTCTTGAATTCTTGATGAGTGTTGAT
Linc8sh1000For	TGGTCTACAGTAGAGAATTTCAAGAGACTTCTCTACTGTAGACCACTTTTTC
Linc8sh1000Rev	TCGAGAAAAAGTGGTCTACAGTAGAGAAGTCTCTTGAAATTCTCTACTGTAGACCA

2.2.2.10 DNA ligation

Ligations were performed using T4 DNA ligase (NEB, #M0202S). Variable amounts (up to 100 ng) of linearized plasmid DNA were mixed with insert DNA (vector to insert ratio between 1:3 and 1:5), 20 units of T4 DNA ligase and 1x ligase buffer in a final volume of 20 µl. Reactions were incubated overnight at 16 °C. As a control for vector re-ligation, a ligation reaction without an insert was also set up. Ligation reactions were heat-inactivated at 65 °C for 20 min. 3 µl of ligation reactions were used to transform chemically competent bacterial cells as described in Chapter 2.2.1.7.

Table 2.6 Expression vector construction oligos

Oligo name	**Oligo sequence (5′ to 3′)**
PouFLAGrev1	CTGCGCCCCGGCATTCACTGCTCTAGAAGAC
PouFLAGrev2	CGGCATTCACTGCACGCTGGTCTGATCTAGAAGAC
Linc8lforEcoRI	AGACGAATTCGAGTAGTATCAGTATAGATAAGAG
Linc8lrevXhoI	AGACCTCGAGGTATTCTTTCCTTTCCCCTTTATT

2.2.2.11 On-column DNA purification

DNA samples were purified to remove enzymatic reaction components using QIAGEN QIAquick PCR Purification kit (QIAGEN, #28104) according to the manufacturer's instructions. Briefly, DNA was bound to a column in a high salt buffer. Unbound impurities were washed away, and DNA was eluted in Elution buffer. DNA concentration was determined using NanoDrop (NanoDrop).

2.2.2.12 Nucleic acid purification by phenol-chloroform extraction and alcohol precipitation

Equal volume of phenol:chloroform (1:1) was added to the nucleic acid (DNA or RNA) sample to be purified. The samples were vortexed for 15 sec and centrifuged at full speed for 5 min at room temperature to separate the phases. The aqueous phase was transferred to a fresh tube. Phenol:chloroform extraction was repeated until no protein was visible at the interphase of the organic and aqueous phases. An equal volume of chloroform was added to the aqueous phase, the samples were vortexed and centrifuged as above. DNA in the aqueous phase was precipitated with either 2.2 volumes of absolute ethanol (at -20 °C) for at least 1 h or 0.6 volumes of isopropanol (at RT) for 15 min. The samples were either incubated at -20 °C before pelleting the nucleic acids, or pelleted by centrifugation at full speed for at least 30 min at -4 °C immediately. Pellets were washed with 70% (or 80% for shorter RNA or DNA species) ethanol, centrifuged again for 10 min, air-dried and re-suspended in the appropriate buffer or water, as required.

2.2.2.13 Small-scale plasmid DNA isolation

A single bacterial colony isolated from a Luria-Bertani (LB)-agar plate was inoculated in 2 ml of LB medium (1% bactotryptone, 1% NaCl) supplemented with 100 μg/ml carbenicillin overnight at 37 °C with shaking. Cells from 1.5 ml of the culture were spun down in a microcentrifuge at 3,000 rpm for 5 min. Plasmid DNA was isolated using the QIAprep Spin Miniprep kit (QIAGEN, #27104) according to the manufacturer's instructions.

2.2.2.14 Large-scale DNA isolation

A single bacterial colony isolated from an LB-agar plate was inoculated in 3 ml of LB medium supplemented with 100 µg/ml carbenicillin overnight at 37 °C with shaking. 2.5 ml of overnight medium was inoculated into 500 ml of medium and re-grown with vigorous shaking (250 rpm). Cells were divided into 50 ml falcon tubes and pelleted using a Heraeus Multifuge 3S-R Centrifuge (Thermoelectron Corporation) at 4,500 rpm for 20 min at 4 °C. Plasmid DNA was isolated using a QIAquick HiSpeed Plasmid Midi Kit (QIAGEN, #12643) according to the manufacturer's instructions with some modifications. Briefly, each bacterial pellet was re-suspended in 10 ml Buffer P1 (supplemented with 100 µg/ml RNase A). 10 ml of Buffer P2 was added to each tube. Samples were mixed thoroughly by inverting the tubes and incubated for 5 min at room temperature. Then, 10 ml of chilled Buffer P3 was added to each tube, mixed immediately by inverting the tubes and incubated on ice for 15 min. Samples were centrifuged at ≥ 20,000 x g or 30 min at 4 °C. Supernatant containing plasmid DNA was transferred to fresh tubes and centrifuged again at ≥ 20,000 x g or 30 min at 4 °C. Supernatant was transferred to fresh tubes. QIAGEN-tip 100 was equilibrated by applying 4 ml of Buffer QBT, and allowing the column to empty by gravity flow. Two supernatants containing plasmid DNA were pooled together, applied to the QIAGEN-tip, and allowed to enter the resin by gravity flow. QIAGEN-tip was washed twice with 10 ml of Buffer QC. DNA was eluted with 5 x 1 ml Buffer QF pre-warmed to 65 °C (eluting in 5 x 1 ml aliquots prevents Buffer QF from cooling). DNA was precipitated by adding 3.5 ml room temperature isopropanol. Samples were mixed and centrifuged immediately at ≥ 15,000 x g for 30 min at 4 °C. Pellets were

washed with 2 ml of room temperature 70% ethanol, centrifuged at ≥ 15,000 x g for 10 min at 4 °C, air-dried and re-dissolved in a suitable volume of 10 mM Tris-HCl, pH 8.5.

2.2.2.15 DNA sequencing

Sanger sequencing was performed at the Source BioScience Sequencing Service, Department of Biochemistry, University of Oxford, UK. Generally, 5 µl of 100-250 ng/µl plasmid DNA and 5 µl of 3.2 pmol/µl primer solutions were submitted per sequencing reaction. Sequencing was performed using conventional in-house protocol.

For DNA methylation studies, the same sample requirements were observed, but the dGTP chemistry protocol was used for sequencing.

Next-generation sequencing was performed at the High-Throughput Genomics Group at the Oxford Genomics Centre, Wellcome Trust Centre for Human Genetics, University of Oxford.

2.2.2.16 Gel extraction

DNA fragments were visualised using a High Performance UV Transilluminator and excised from agarose gel with a blade. DNA was extracted from the gel and purified using the GeneJET Gel Extraction kit (Thermo Scientific, #K0692) according to the manufacturer's instructions. Briefly, agarose gel fragments were solubilised under high salt conditions to enable binding of DNA to the GeneJET column. Unbound material and impurities were washed away. DNA was eluted in Elution buffer or nuclease-free water (Ambion). The concentration of purified DNA fragments was determined using a

NanoDrop. 2 µl of the samples were also run on an appropriate density agarose gel to check excision precision and DNA quality.

2.2.2.17 Agarose gel electrophoresis

Agarose gel electrophoresis was performed in a horizontal gel electrophoresis system (BioRad). Agarose gels (0.7-4%) were prepared by dissolving agarose (Fisher Scientific, #BP160-500) or low gelling temperature agarose (Sigma, #A9414) in 100 ml of 1x TAE (40 mM Tris-Acetate, 1 mM EDTA, pH 8.0) in a microwave. 1x GelRed Nucleic Acid Stain (Biotium, #41003) was added and the solution was poured into a gel casting tray with sealed ends and inserted combs. Electrophoreses were performed using 1x TAE buffer. A DNA ladder marker (Chapter 2.2.2.18) and DNA samples containing 1x loading dye were loaded into individual wells. Electrophoretic separation of DNA fragments was carried out at 80-120 V (constant voltage). DNA was visualized using a transilluminator (Vilber Lourmet). Gel images were taken using AlphaImager MultiImage Light Cabinet (Alpha Innotech) and AlphaEase FC (Alpha Innotech) software.

2.2.2.18 DNA ladder marker

A DNA ladder marker for gel electrophoresis was obtained by mixing 500 µg/ml 1 kb DNA ladder (NEB, # N3232L) or 500 µg/ml 100 bp DNA ladder (NEB, N3231L) with 1x Blue Gel Loading Dye.

2.2.2.19 Chromatin Immunoprecipitation (ChIP)

N2A cells were seeded on 150 mm dishes and grown to 80-90% confluency (approximately 2 x 10^7 cells). 1/10 volume of cross-linking mix (11% formaldehyde, 100 mM NaCl, 0.5 mM EGTA, 50 mM HEPES, pH 8.0) was added directly to the dishes containing growth medium, giving a final concentration of 1% formaldehyde. Dishes were incubated in a tissue culture incubator at 37 °C for 10 min. Cross-linking reactions were quenched by adding 1/10 volume of 1.25 M glycine to a final concentration of 0.125 M. Dishes were further incubated for 10 min at room temperature with gentle rotation. Medium was removed from the plates, cells were washed with ice cold PBS and scraped off the plates using a cell scraper into a fresh 50 ml falcon tubes. Cells were pelleted at 3,000 rpm for 5 min at 4 °C, and washed twice with ice-cold PBS. Cell pellets were re-suspended in 2 ml of Lysis buffer (10 mM Tris-HCl, pH 7.2, 10 mM KCl, 1.5 mM $MgCl_2$, 0.5% NP-40, 1x protease inhibitor cocktail) per each 150 mm dish of cells. Cell suspension was transferred to a Dounce homogeniser and dounced 10 times with a tight pestle on ice. Cell suspension was incubated on ice for 10 min and dounced 10 times again. Cell suspension was transferred to a centrifuge tube and nuclei were pelleted at 3,000 rpm for 5 min at 4 °C. Supernatant was discarded and nuclei pellets were re-suspended in 300 ul of Sonication buffer (1% SDS, 10 mM EDTA, pH 8.0, 50 mM Tris-HCl, pH 8.0 with protease inhibitor) per 150 mm dish used. Chromatin was sheared to approximately 500-1000 bp using a Bioruptor (Diagenode) for 35 cycles (30 sec ON, 30 sec OFF), level 5. Sheared samples were centrifuged for 10 min at full speed at 4 °C to pellet insoluble material and pooled together. 10 µl aliquot was removed, boiled at 100 °C for 5 min, phenol-extracted and run on a 1.5% agarose gel to check fragmentation. DNA

concentration in the sonicated sample was estimated using a NanoDrop, and adjusted to 1 µg/µl with Lysis buffer. For each immunoprecipitation, 400 µl aliquots were stored at –80 °C until needed or used for further processing. 400 µl aliquot of sonicated chromatin (approximately 400 µg DNA) was diluted to a final volume of 4 ml with Dilution buffer (1% Triton X-100, 150 mM NaCl, 2 mM EDTA, pH 8.0, 20 mM Tris-HCl, pH 8.0, with protease inhibitors). 40 ul aliquot, representing 10% input, was placed in a fresh tube labelled "Input" and stored at -80 °C until required. 4 µg of antibody against a protein of interest, or control normal rabbit IgG, was added per immunoprecipitation and rotated at 4 °C overnight.

Next day, 100 µl of Protein A/G Dynabeads (Life Technologies, #10003D) was washed twice and re-suspended in the original volume of 9:1 Dilution buffer:Lysis buffer. 100 µl of beads was added to each tube incubated overnight, and incubated at 4 °C for further 4 h with rotation. Following the incubation, tubes were pulsed for 2 sec and placed on a magnetic separator. Supernatant was carefully removed. Beads were resuspended in 1 ml of Wash buffer (1% Triton X-100, 0.1% SDS, 150 mM NaCl, 2 mM EDTA, pH 8.0) and transferred to a 1.7 ml tube. Beads were washed an additional three times in 1 ml Wash buffer, once with Final wash buffer (1% Triton X-100, 0.1% SDS, 500 mM NaCl, 2 mM EDTA, pH 8.0, 20 mM Tris-HCl, pH 8.0, with protease inhibitors), and then once with Wash buffer B (20 mM Tris-HCl, pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate, with protease inhibitors). After the washes, all supernatant was removed. Immune complexes were eluted with 2x 225 µl of Elution buffer (1% SDS, 100 mM $NaHCO_3$) at 37 °C for 30 min each with rotation. At this stage, 40 µl of Input sample was diluted to 450 µl with Elution buffer and treated in parallel with immunoprecipitation

samples. RNase A was added to samples to a final concentration of 500 µg/ml. Samples were incubated at 65 °C overnight in the presence of 200 mM NaCl to reverse cross-links. The following day, samples were phenol-chloroform extracted and DNA precipitated with ethanol overnight at -20 °C. Samples were then centrifuged at full speed for 30 min at 4 °C. Pellets were washed with 70% ethanol, centrifuged again, air-dried briefly and re-suspended in 100 µl of water. 25 µl of 5x proteinase K buffer (50 mM Tris-HCl, pH 6.5, 25 mM EDTA, 1.25% SDS), 3 µl of 10 mg/ml proteinase K was added, and samples incubated for 1-2 h at 55 °C. Following the incubation, 15 µl of 3M Sodium Acetate was added to each sample. Samples were purified using QIAGEN PCR Purification kit (QIAGEN, #28104) according to the manufacturer's instructions. DNA was eluted in 50-100 µl of elution buffer and used for qPCR analysis.

The following antibodies were used: anti-DNMT1 (Abcam, #ab87656), anti-BRG1 (Abcam, #ab4081), anti-CTCF (Abcam, #70303), normal rabbit IgG (Millipore) (Chapter 2.1.1). For FLAG-tagged POU3F3 ChIP, mouse monoclonal anti-FLAG M2 beads (Sigma-Aldrich) were used for immunoprecipitation. ChIP from extracts not containing FLAG-tagged POU3F3 was used as control instead of the isotype-matched control antibodies.

Table 2.7 ChIP-qPCR oligos

CTCF ChIP-qPCR

Oligo name	**Oligo sequence (5′ to 3′)**
THChIPF1	GTCAGCCAACATGGGTACG
THChIPR1	GTGGCCTCACACAGAGACTG

Tgfb3chartF2	CCCATTGGCACTCACTCC
Tgfb3chartR2	GCAAACAGCAGGAGGCTAAA
Creb3l2ChIPF1	GCGGAGAGGGTCTTTTACG
Creb3l2ChIPR1	AGCACCCTATTGGTCCATTG
Slc1a4ChIPF1	CTAGGGAGTCCGGAACGAC
Slc1a4ChIPR1	GCCTACAGAGCCCTTAGAAGAA
GanabChIPF1	CCTGGGCCAGTTTCATTAGT
GanabChIPR1	TGCAGACTGTGAACTTACTGGAC
Zfp954ChIPF1	GGGGACATGAACGTCTGC
Zfp954ChIPR1	ATCGCCCACCTACCTGAGTT
AcheChIPloF1	CACATTTCAGGGAGTTGGTTC
AcheChIPloR1	TTGTCGCCGAACATACTTAGG
Linc8 for1821 (DaliChIPF)	CCTCCGATCCTGCAGTTAGC
Linc8 rev1912 (DaliChIPR)	ACAGCAGTTCAGCAGCAAGAAG
Pou3f3 for1025 (Pou3f3ChIPF)	TGGCTCTGGGTACGCTCTATG
Pou3f3 rev1124 (Pou3f3ChIPR)	GGCTTGAGCTTGCACATGTTC
linc5F1	GGCACCTTCCTCCTAAGACAC
linc5R1	CCATTCCCTATGTTGTCACCTC
GAPDH 1	GTTTCCATAGGACCTGCTGCG
GAPDH 2	GTTTTACACTGGGCACTTGAGGTC

POU3F3 ChIP-qPCR

Chga3ChIPF1	TGAATGTGAAATTTGTTCTGATTCTT
Chga3ChIPR1	CCCCTCGGGAGTGAAAAA
Fam92bChIPF1	GGAGATGGGGCTCATTCA
Fam92bChIPR1	GGTACAGGAGGCGGATTCTC

Prr11ChIPF1	CGCAGTGGAAGCTTTTATGTC
Prr11ChIPR1	AGTTTGCGATTGGTGTACTCG
Gins1ChIPF1	CCAGTGCACTTCTATTGGTTGA
Gins1ChIPR1	TCAGAGTCGCACAACACCA
Itga2ChIPF1	CAAACTCCGGATCTGTCCTG
Itga2ChIPR1	TGGATCTTGACTCCTTTTTCAGA
Lgi1ChIPF1	CCAAAATCCACGGATCTCAC
Lgi1ChIPR1	AACTTGGGCAGATGGTTCC
Galnt5ChIPF1	GGCACAAATACTTTGCTGTGAA
Galnt5ChIPR1	TCAGCTCTTACTGACTGCCTGT
DbhChIPF1	CCCCACTGGACAGGCATA
DbhChIPR1	GGCTGAGATGAGCTTGCAT
Foxd1ChIPF1	TGGGCTTCAGTCTCCAACTC
Foxd1ChIPR1	AATCGGCGTGTAGGAAGAGA
Arhgap18ChIPF1	ATGCCTTCCTTGAGTGATGC
Arhgap18ChIPR1	TTTCCCTTCCCCATGTGTTA
Angpt2ChIPF1	ATGTCAGAGGCACCACCCTA
Angpt2ChIPR1	GCTTAGCCTACAAACGAGCAG

2.2.2.20 Bisulfite conversion and sequencing

Bisulfite conversion of genomic DNA was performed using EZ DNA Methylation-Gold kit (Zymo Research, #D5005) according to the manufacturer's instructions. Briefly, 130 µl of the CT Conversion Reagent was added to 20 µl of DNA sample (80-350 ng DNA) and pipetted to mix. The sample was incubated in a thermal cycler as follows: 98 °C 10 min,

64 °C 2.5 h, store 4 °C (up to 20 h). Following the incubation, the sample was loaded onto a Zymo-Spin IC Column pre-loaded with 600 µl of M-Binding Buffer. The column was inverted several times to mix and centrifuged a full speed (14,000 x g) for 30 sec. The flow-through was discarded, 100 µl of M-Wash Buffer was added, and the column was centrifuged again at full speed for 30 sec. The flow-through was discarded, and 200 µl of M-Desulphonation Buffer was added to the column and let stand at room temperature for 15-20 min. Following the incubation, the column was centrifuged at full speed for 30 sec, and the flow-through was discarded. The column was washed twice with 200 µl of M-Wash Buffer, and then transferred to a fresh 1.5 ml tube. Bisulfite converted DNA was eluted by adding 12 µl of M-Elution Buffer directly to the column matrix and centrifuging the column for 30 sec at full speed. 1 to 4 µl of eluted DNA was used for each PCR. PCR products were purified using QIAquick PCR Purification Kit (QIAGEN, # 28104), TOPO-cloned into the pCR4-TOPO vector and sequenced using dGTP chemistry (Source BioScience). At least 20 individual clones were sequenced for each amplified region and cell line. Sequencing data was analysed using BiQAnalyzer HT programme accessible at http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/.

2.2.2.21 Combined Bisulfite Restriction Analysis (COBRA)

80-350 ng of genomic DNA was bisulfite-treated using EZ DNA Methylation Gold kit as described above. 1 µl (or up to 4 µl for 50 µl PCR reactions) of bisulfite converted DNA was used for each 25 µl PCR reaction. PCR reactions contained 1x EpiMark Hot Start *Taq* Reaction Buffer (NEB, #M0490L), 200 µM each dNTPs, 0.2 µM forward primer, 0.2 µM reverse primer, 1.25 units (per 50 µl reaction) EpiMark Hot Start *Taq* Polymerase, 1-4 µl

(30-120 ng) template bisulfite converted genomic DNA. Primers for amplifying bisulfite converted template DNA were designed using MethPrimer software accessible at http://www.urogene.org/methprimer/ (Li and Dahiya, 2002). PCR reactions were performed using the following thermo-cycler programme: 95 °C 30 sec (initial denaturation); 40 cycles of 95 °C 30 sec, 40-47 °C 30 sec, 68 °C 30 sec; 68 °C 5 min (final extension); 4 °C (store).

A second round of PCR was performed in some cases if insufficient product was obtained for the downstream processing after the first round. PCR products were on-column purified with QIAquick PCR Purification Kit. 250 ng to 1 µg of purified products were incubated with appropriate COBRA-compatible (*Bst*UI (NEB, #R0518S), *Msp*I (NEB, #R0106S), *Taq*I (Thermo Scientific, #ER0671), *Hpy*CH4IV (NEB, #R0619S)) or control (*Hsp*92II (Promega, #R7161), *Bfa*I (NEB, #R0568S)) restriction enzymes overnight. Restriction products were analysed on 3% low melting point agarose gels.

Table 2.8 COBRA oligos

Oligo name	**Oligo sequence (5′ to 3′)**
Fbn1COBRAF3	GGTAGAGAGGGTTAGTTGTAGATGTT
Fbn1COBRAR3	AATTAAAACAAAAACTATAAATACCACAAA
Gdpd5COBRAF1	TATTTTTTAGGGGAGGGGTTAAGT
Gdpd5COBRAR1	AAAAAATCTACCCAAATCCC
Gdpd5COBRAF2	TTAGTTTTGGGATTTGAGTGTAGAG
Gdpd5COBRAR2	AAAACCCCTATATTCCCCTAAAAA
Nos1COBRAF1	AAAAGATGTATGTTTTTAAGTTTAGAG

Nos1COBRAR1	AAAAACTACCAACTTCCCCTTAC
Nos1COBRAF2N	AAGGGGAAGTTGGTAGTTTTTAGTT
Nos1COBRAR2N	CAATTCCAAAACTATAAATCAC
Fam173aCOBRAF2	AAATTTTTTTGAGGTTGTATTGTTG
Fam173aCOBRAR2	AACCAACATCTATCTCTAACTACCC
Dlgap5COBRAF2	TTAGGGATTTAAGTGTGAGGTGTAAG
Dlgap5COBRAR2	AAAAACACCTTAAATAAATAAAAACC
Cd9COBRAF1	GAAGAGTAGGTATTTGATGTATTTGTTATT
Cd9COBRAR1	AAATCCCAAAACCTAAATCTAAAAAAA
Mical2COBRAF2	GTTTTAAGTATGGGGTGGAGTTT
Mical2COBRAR2	CAAATCAACACCTACCTTTCTTATC
Zik1COBRAF2	TTGGATAATATTGTTATTGTTAGTGT
Zik1COBRAR1	ATTTTTAAATCCCTCCTTCCAAAT
Kif18aCOBRAF2	AAGTTAATTTTGTTAAGTTGTGA
Kif18aCOBRAR2	AAAAAAAACCAACTTCCCCTC
Padi2COBRAF1-2N	GAATAAAAGAATGAATAAGTTAAATA
Padi2COBRAR1N	CTCCCAAAAATAAACACCTACTACAA
Hmgb2COBRAF1N	GGTAAGGTGAGGGATTTTGGT
Hmgb2COBRAR1	CCCCTAATAATTTAAAAATAAATAAAATAA
MetrnCOBRAF1	GTGTGGTTTGTAGGAAAAGGGT
MetrnCOBRAR1	ATTTAACCCAAAAACCTAACAAC

Table 2.9 COBRA follow-up oligos

Oligo name	Oligo sequence (5′ to 3′)
Nos1Ex1F	GATCCACAGCCCTGGAACT
Nos1Ex1R	AGATGCAGAGGTGATGTAGGG
Nos1Ex2F	TCGCTTCCTTAAGGTCAAGAAC
Nos1Ex2R	CTCTGTGCACCCCGTTTC
Nos1F3	CATCAGGCACCCCAAGTT
Nos1R3	CAGCAGCATGTTGGACACA
Fbn1ExF1	CGCTTCTCAGCCCAGTCTT
Fbn1ExR1	GAGCCTCGGAGCCTACAAT
Fbn1ExF2	TGGGAGGAGAAATTGTAGGC
Fbn1ExR2	CCTGTTGGCTGGCTCAGT
Fbn1ExF3	TGTCCTGGATGGAAAACCTTA
Fbn1ExR3	GAATCCATCCCCACAGGAAT
Fbn1ExF4	CCTTCCTGTGGCTCCAGAT
Fbn1ExR4	GCTGCCCCCATTCATACA
Fbn1ExF5	TGAAGTGTCACAAAAATGTGAAGA
Fbn1ExR5	TGTACATTCCCCGCCATC
Fbn1ExF6	AGTGCTGTTGCGATCTTGG
Fbn1ExR6	GCACAGCTTGTTGAAATCCTC
Fbn1ExF7	CTCCTACAGATGCGAATGCTT
Fbn1ExR7	TCCGCATGTGTGTGTCAA
Fbn1ExF8	CAGTGGACCGGGAATGAC
Fbn1ExR8	TGGGCAAATATCAGGATCTAATG

2.2.3 RNA manipulation

2.2.3.1 RNA isolation

Prior to RNA isolation, adherent cells were harvested and rinsed twice with ice cold PBS twice. Cells were lysed by adding 1 ml of TRIzol Reagent (Life Technologies, #15596-026) per 3.5 cm diameter dish worth of cells (or approximately 5-10 x 10^6 cells). Cell lysate was passed several times through a pipette and then vortexed thoroughly to ensure full cell lysis. Homogenized samples were incubated at room temperature for at least 5 min to allow the complete release of RNA from nucleoprotein complexes, and centrifuged for 10 min at 10,000 x g at 4 °C to remove cell debris. Supernatant was transferred to a new tube. 0.2 ml of chloroform was added per 1 ml of TRIzol Reagent. Samples were vortexed vigorously for 15 seconds and incubated at room temperature for 2 to 3 minutes. Samples were then centrifuged at no more than 12,000 x g for 15 minutes at 4 °C. Following centrifugation, the upper aqueous phase was carefully transferred without disturbing the interphase into a fresh tube. 0.5 ml of isopropanol was added per 1 ml of TRIzol Reagent used for the initial homogenization to the aqueous phase to precipitate the RNA. Samples were incubated at room temperature for at least 10 minutes and centrifuged at 12,000 x g for 10 minutes 4 °C. Following centrifugation, the supernatant was removed completely. Pellets were washed once with 75% ethanol with at least 1 ml of 75% ethanol per 1 ml of TRIzol Reagent used for the initial homogenization. Samples were mixed by vortexing and centrifuged at 7,500 x g for 5 minutes at 4 °C. The washing procedure was repeated one more time, if necessary. After the washes, residual ethanol was removed completely. RNA pellets were air-dried and re-dissolved in RNase-free water.

2.2.3.2 DNA contamination removal from RNA samples

To remove residual DNA contamination from RNA samples prior to RT-qPCR, TURBO DNA-free™ Kit (Cat. No. AM1907) was used according to the manufacturer's instructions. Briefly, 0.1 volume of 10x TURBO DNase Buffer and 1 µl of TURBO DNase was added to RNA samples and mixed gently. The samples were incubated at 37 °C for 20-30 min, followed by the addition of 0.1 volume of DNase Inactivation Reagent. The samples were then mixed well and incubated for 5 min at room temperature, mixing occasionally. The samples were centrifuged at 10,000 g for 1.5 min, and the supernatants were transferred to a fresh tube.

2.2.3.3 First-strand cDNA synthesis

First-strand cDNA synthesis was performed using QuantiTect Reverse Transcription Kit (QIAGEN, #205313) according to the manufacturer's protocol. Briefly, 1 µg RNA (12 µl) was incubated with 2 µl of gDNA Wipeout Buffer at 42 °C for 5 min to digest contaminating DNA, and then plunged onto ice. To the sample, 6 µl of Reverse Transcription mix containing 1 µl Quantiscript Reverse Transcriptase, 4 µl 5x Quantiscript RT Buffer, 1 µl RT Primer Mix was added. First-strand cDNA synthesis was performed for 30 min at 42 °C, before inactivating the reaction for 5 min at 95 °C. The samples were diluted 1:2 with nuclease-free water before using in qPCR reactions.

2.2.3.4 Gene-specific first-strand cDNA synthesis optimized for low-abundance transcript detection

This was performed using the QuantiTect Reverse Transcription Kit (QIAGEN, #205313) according to the manufacturer's protocol with the following modification. Instead of RT Primer Mix supplied with the kit, a gene-specific reverse oligo was used to prime first strand synthesis. The oligo was placed 3′ to the primer pair used for the subsequent qPCR reaction. For every transcript of interest, a range of oligo concentrations (14 μM, 7 μM, 3.5 μM, 1.75 μM, 0.875 μM) was tested. An optimal concentration that provided efficient priming while avoiding PCR artefacts (as judged by the melting curve performed at the end of the amplification programme) was selected. Usually, a 3.5 μM concentration worked well for both the low-abundance transcripts studied and normalization controls.

2.2.3.5 RNA expression analysis using quantitative PCR (qPCR)

Primers used in qPCR analyses throughout the project were designed using the Universal ProbeLibrary Assay Design Center, Roche. Relative expression levels of the genes were measured by real-time quantitative PCR (qPCR) using SYBR® Green Master Mix (Applied Biosystems) with the StepOne™ Real-Time PCR system (Applied Biosystems). Three technical replicates were tested for each sample, including "no RT" negative controls. Expression fold changes in shRNA-transfected samples were calculated relative to the levels observed in scrambled control transfected samples. Abundance of each transcript was normalized against that of a housekeeping gene(s), known not to be affected by conditions of a particular experiment.

2.2.3.6 Adult mouse tissue panel gene expression measurements

Expression of candidate lncRNAs and their protein-coding neighbours was assessed in tissues included on the Mouse Total RNA Master Panel (Clontech, #636644) using RT-qPCR.

Table 2.10 LncRNA and TF qPCR oligos

lncRNA qPCR oligos

Oligo name	**Oligo sequence (5′ to 3′)**
linc4F2	GGGAGTCTCCTGACTCTTTGG
linc4R2	GGCCTTCCCTTAGGAGTATTTG
linc5F1	GGCACCTTCCTCCTAAGACAC
linc5R1	CCATTCCCTATGTTGTCACCTC
linc7F1	ATCCACAGTGACTGGCTAAGG
linc7R1	GGCCTCTGATGCTTTTTCAC
linc8 for1821	CCTCCGATCCTGCAGTTAGC
linc8 rev1912	ACAGCAGTTCAGCAGCAAGAAG
linc8bF1	AGAAGGCCTCCCTGAAAGTG
linc8bR1	ACCAAGATCCCTGTCACACAC
linc8uF1	CGGGACTGTAAGGCGGATA
linc8uR1	AATTGGCAGGAGGAGCATC
linc9F2	CATGGCTCACGGAAAAGC
linc9R2	AAGGACACTTTGCGATCCTG
linc12F1	CTTCTCCTGTTGTCCCATCTG
linc12R1	GCCTCCACCAAATTGTCAG

TF gene qPCR oligos

Meis1F1	TTCACACTGGCCTTAAAGAGG
Meis1R1	CCATAATGGGGTAGGTCGTC
MitfF1	GACACCAGCCATAAACGTCAG
MitfR1	TTTTCCAGGTGGGTCTGC
Smad2F1	AGGACGGTTAGATGAGCTTGAG
Smad2R1	CAGTCCCCAAATTTCAGAGC
Cited1F1	CCGTACCTCAGCTCCTGTG
Cited1R1	ATCGGTGGCAGTTGATGC
Pou3f3 for1025	TGGCTCTGGGTACGCTCTATG
Pou3f3 rev1124	GGCTTGAGCTTGCACATGTTC
Meis2F1	AGACAAGGACGCAATCTATGG
Meis2R1	GCTCGCACTTCTCAAAAACC

2.2.3.7 Transcriptomic analysis

Total RNA was isolated using the Qiagen Mini RNeasy kit (QIAGEN, #74106) following the manufacturer's instructions. RNA integrity was assessed on a BioAnalyzer (Agilent Technologies). 200 ng RNA was used to produce labelled sense single stranded DNA (ssDNA) for hybridization with the Ambion WT Expression Kit (Affymetrix, #4411973), the Affymetrix WT Terminal Labelling and Controls Kit (Affymetrix, #901525) and the Affymetrix Hybridization, Wash, and Stain Kit (Affymetrix, # 900720) according to the manufacturer's protocols. Sense ssDNA was fragmented and the distribution of fragment lengths was analysed on a BioAnalyzer. Fragmented ssDNA was then labelled and hybridized to the Affymetrix GeneChip Mouse Gene 1.0 ST Array (Affymetrix, #901169).

Microarray chips were processed on an Affymetrix GeneChip Fluidics Station 450 and Scanner 3000.

Table 2.11 Microarray validation oligos

Stable arrays

Oligo name	**Oligo sequence (5′ to 3′)**
Tmsb4F1	AGCAGATCAGACTCTCCTCGTT
Tmsb4R1	GCCATATCGGGTTTGTCAGA
Capn6F1	ATGGCCGTACCAGTGAATTT
Capn6R1	GCATGTCCAAGGTCAGTTCC
Lix1F1	GGATCACCAAGGAGTTCATCA
Lix1R1	CATCCGCATCATCTAACGTG
IslrF1	TGACTTCAGAAGAACAGTTTGGAG
IslrR1	GTTCAGGAGAACAGCCCAAC
ArhgdibF1	GAGACACAGGGCCGACTG
ArhgdibR1	GCTTGCTGTCCAGGTCGT
Sept3F1	CCAGCGCAATATCCCAAT
Sept3R1	CGTCCGTCCTAGCCTCTG
NrcamF1	TTAACAAATGGAGTCCCAATAGAAA
NrcamR1	TGAAAATATAATGGTATCGCCATC
Lima1F1	CAAGAACAAGTCATCCGCAAT
Lima1R1	TTCCTTTCCATGTTGGCTTC
SctrF1	CAGCATCAGCCATCGTATTC
SctrR1	GTCCCAAAACCTCTCACAGC
Efhc2F1	GTTTGAGCCTATAGAGAATAATTCAGG

Efhc2R1	CAGGCTTTTTCACACGACTTC
CamkvF1	CTGCTGAGCGGGTTATCG
CamkvR1	ACCCAAACGGCATTGCTA
Casp8F1	TTGAACAATGAGATCCCCAAA
Casp8R1	CCATTTCTACAAAAATTTCAAGCAG
CluF1	AAGTACTACCTTCGGGTCTCCA
CluR1	AGCTTCACCACCACCTCAGT
Gstm5F1	CAGCTATGAGGAGAAACGGTACA
Gstm5R1	TCTTCCCGTCCATGAGGTAG
Nos1F1	CATCAGGCACCCCAAGTT
Nos1R1	CAGCAGCATGTTGGACACA
SctrF1	CAGCATCAGCCATCGTATTC
SctrR1	GTCCCAAAACCTCTCACAGC
Unc5aF1	TTCTGTGAGGGGCAGAATG
Unc5aR1	GCAGGCTGACCACTTACTCC
Rbms2F1	CTAGTGGGGCCAGCAGAG
Rbms2R1	GGTGATGATGGCTTCACACTT
Zfp827F1	TCCAGTTTCCAAGGGAAGAG
Zfp827R1	AATGGCTTCCCGCTTTCT
Pcp4F1	CCAACGGAAAAGACAAGACG
Pcp4R1	TGTCGATATCAAATTCTTCTTGGA
Irgm1R1	AAGGCCACTAACATCGAATCA
Irgm1F1	TGCCTTATCTCACTTAATACTCCTCA
Fbn1F1	GGACGGAAAGAACTGTGAAGAT
Fbn1R1	ACACATTCCGTTTAGGCACA
Padi2F1	GATCCTCATCGGAAGCAGTT

Padi2R1	AGTCGCGTACCACCTTGG
Slc16a3F2	CTCGCTGCAGCTCAGTTCT
Slc16a3R2	GACCCTCGTCCACCACAG
Nos1shF1	GATCCACAGCCCTGGAACT
Nos1shR1	AGATGCAGAGGTGATGTAGGG
Nos1loF1	CATCAGGCACCCCAAGTT
Nos1loR1	CAGCAGCATGTTGGACACA
Sec11cF1	GCTTGTACAAAGAAGGCCAGA
Sec11cR1	TGACCATGCCAACATAAGGTAA
Ddit3F1	GCGACAGAGCCAGAATAACA
Ddit3R1	GATGCACTTCCTTCTGGAACA
Hmgb2F1	CTCTGCGGAGGTCTGAGG
Hmgb2R1	GTCACCCTTGCCCATGAC
Dlgap5F1	ATGCCACCTTCTTGAACCAC
Dlgap5R1	GAACTGTCTGCTGCGATCTG
MelkF1	TCTGTCCTGTTCGGTCCTGT
MelkR1	TCAAAATGACAACCAGAACCTG
Fam72aF1	TCCCTTTCCACCAGGAAAAT
Fam72aR1	GCCTCTCGTTCTCTACGCTCT
Mcm3F1	CAGGACTCCCAGAAAGTGGA
Mcm3R1	TAAGAGGGCCGCCTTAAAA
Zfand2aF1	GCACAGGCATCCTCTGGA
Zfand2aR1	TCAGCAGCTCTGGATGTAGAAG
JunF1	CCAGAAGATGGTGTGGTGTTT
JunR1	CTGACCCTCTCCCCTTGC
Tinf2F1	TGGTTTTGGACTCTGATGAGG

Tinf2R1	GGGGATATAGGTGTCAAACTTCG
Chek2F1	TTATTCCTGAAGTCTGGACAGATG
Chek2R1	CTAACAGTTTCTTGACAAGGTCCA
NfkbiaF1	ACGAGCAAATGGTGAAGGAG
NfkbiaR1	ATGATTGCCAAGTGCAGGA
Pmp22F1	AGCTGTCCCTTTGAACTGAAAC
Pmp22R1	CCCCAACAAGAGTAGGAGCA
Psmc2F1	AAAACCCGACGTCACATACAG
Psmc2R1	CGCAATTTCTCAATCTGTTCC
Nusap1F1	TCTAAACTTGGGAACAATAAAAGGA
Nusap1R1	TGGATTCCATTTTCTTAAAACGA
Zfpm1F1	AGGCCCCAGGATGAAGAG
Zfpm1R1	ATCCTGCAGTGCCAGCTC

Transient arrays

Rrm2F	GCAGCCAGTGATGGAATTG
Rrm2R	GGAACCTGCACCTCCTGAC
Anxa2F	GGAAATATGGCAAGTCCCTGT
Anxa2R	TCTGGTAGTCACCCTTGGTGT
NaspF	TGGGTGACATTCCAGCAG
NaspR	GCCGTTTCTCCATACTTCTTACC
HellsF	TGATCATTTATGACAGTGATTGGAA
HellsR	TTGTCTGACCAATTCTATGACATCTA
Mcm3F	CAGGACTCCCAGAAAGTGGA
Mcm3R	TAAGAGGGCCGCCTTAAAA
GanabF	TGATCCATGAAGTCACCAAGG

GanabR	CCTGATCCGAGTCATGTTCTT
AcheF	GCTGTCACTGTCGGCTCA
AcheR	AGGACAGGCTGGTGTCTGA

2.2.3.8 Capture Hybridization Analysis of RNA Targets (CHART-Seq)

CHART Enrichment and RNase H Mapping experiments were performed as described in (Simon, 2013). I designed 10 biotinylated DNA capture (C)-oligos: 5 oligos complementary to the most accessible regions of *Dali*, as determined by RNase H mapping, and 5 oligos targeting the most evolutionarily conserved regions of the transcript (Figure 5.1 B). These oligos were used as two cocktails of 5 oligos, and as a pool of all 10. As controls, we used an oligo designed to target the antisense *Dali* sequence (absent from the N2A transcriptome) or *E. coli lacZ* sequence (absent from the mouse genome). Compared to controls, all three cocktails of *Dali* oligos showed significant enrichment of the *Dali* transcript (10-fold compared to *lacZ*), but no enrichment of the abundant mRNA *Gapdh* (Figure 5.2 A). Without any prior information about *Dali* genomic binding, we considered its endogenous site of synthesis to assess the enrichment of transcript-associated DNA loci. Specific enrichment of *Dali* at its locus was observed as expected (Figure 5.2 B).

CHART extract was prepared from approximately 3×10^8 N2A cells per pull down and hybridized overnight with 810 pmol biotinylated oligonucleotide cocktail (Table 2.12) at room temperature with rotation. 250 µl MyOneC1 streptavidin beads (Invitrogen) were used to capture the complexes overnight at room temperature with rotation. After extensive

washes, bound material was eluted using RNase H (New England Biolabs) for 30 min at room temperature. Samples were treated with Proteinase K and cross-links were reversed. RNA was purified from 1/5 total sample volume using the QIAGEN miRNeasy kit. DNA was prepared from the remaining sample using the phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation method. DNA was further sheared to an average fragment size of 150–300 bp using a Bioruptor and sequenced on an Illumina HiSeq (50 bp paired end).

CHART-Seq was performed with three independent pull down samples (using two independent cocktails of 5 C-oligos, and one cocktail containing all 10 C-oligos) and sequenced simultaneously with a matched input sample. Two control samples using non-targeting *lacZ* C-oligos (GSE52571) (Vance et al., 2014) were prepared and sequenced separately.

Table 2.12 CHART oligos

RNase H sensitivity assay

qPCR

Oligo name	**Oligo sequence (5′ to 3′)**
linc8-367F	CAGAGATCTGCCTACCTCTTCTG
linc8-367R	GAGTCACCTCCCCAAAATCTATAA
linc8-855F	GCACTCCTTCATGGCCTCT
linc8-855R	GTCAAGGCAGGAACTCAACC
linc8-1223F	CAAGGACAAATGAGCTGCCTA
linc8-1223R	GTCACAGGGAACTCCACCAT
linc8-1594F	CCTCTGTCTCTGTGTACCAGGATT

linc8-1594R	GGCCATAACTACCTAAGCATCATT
linc8 for1821	CCTCCGATCCTGCAGTTAGC
linc8 rev1912	ACAGCAGTTCAGCAGCAAGAAG
linc8-2025F	CTGGAGCCCCAGAGTCTTAG
linc8-2025R	TCCTTCACTAGGCAATTTGGA
linc8-2365F	CCTGTGGTCTGCACTCACTG
linc8-2365R	CTGGTGGGAAGCATCTGAG
linc8-357595F	CTCTTCTGCTCCACCTCCAC
linc8-357595R	TGTGCATGCCTGGTTTATGT
linc8-12501463F	CCCTGTGACTGTGGGATTCT
linc8-12501463R	TGTGCAGTGTTCAGGCTAGG
linc8-8861089F	CTGCCTTGACTTTCCCTCAG
linc8-8861089R	TGGTCAAGGGGCATTTTAAG

RNase H mapping

linc8FISH7	AAAAACTATAGCTGGGCAGG
linc8FISH8	GTCACCTCCCCAAAATCTAT
linc8FISH9	GGAGAGTTGCAAATGCTCTT
linc8FISH10	TAGGAACTGAATCCAAGCCT
linc8FISH11	ATGATTGTGAGCCACCATGT
linc8FISH13	GTGCATGCCTGGTTTATGTA
linc8FISH14	GGTGTTTGCCTGCACAATAT
linc8FISH21	AGAACACTCTATCACTGAGG
linc8FISH22	CGAGGGTTTACTTCATCTAC
linc8FISH23	TCATGACCAAGGGCAGTTTA
linc8FISH24	GGGGTTCCATTGCTATGATA
linc8FISH25	TGAAGCAAACCTGTCTGGAA

linc8FISH26	ACTAGGAAGGCTTTGCTGAA
linc8FISH27	GACTCTTGCATTCTCTACTG
linc8FISH28	AATTGGTGTCTGTGTGCAGT
linc8FISH29	GGTAGATTGTCACAAGTTCC
linc8FISH30	AATCCTGGTACACAGAGACA
linc8FISH31	TGAAGAAGAATGTGTGCCCA
linc8FISH32	GCCATAACTACCTAAGCATC
C-oligos	
linc8-C-oligo10	TAGGAACTGAATCCAAGCCTCCTTG-HEG-BIOTIN-TEG
linc8-C-oligo11	GGTGGAGGATTTTTCTCAACTGAGG-HEG-BIOTIN-TEG
linc8-C-oligo25	TGAAGCAAACCTGTCTGGAAGCACA-HEG-BIOTIN-TEG
linc8-C-oligo27	GACTCTTGCATTCTCTACTGTAGAC-HEG-BIOTIN-TEG
linc8-C-oligo30	AATCCTGGTACACAGAGACAGAGGC-HEG-BIOTIN-TEG
cons1-1	GCTTTTCATCACCCATCTGGCTCTC-HEG-BIOTIN-TEG
cons1-2	GACTCTGCTTTCAAGAAGTGACAGC-HEG-BIOTIN-TEG
cons2-1	GAAGCATCTGAGAAGTAGCCTGACC-HEG-BIOTIN-TEG
cons2-2	TCAAAATGGAGTTCACTGCGGCTGG-HEG-BIOTIN-TEG
noncons1	TTCCAAAGGGACTTTGGAACAGGTG-HEG-BIOTIN-TEG
lacZ	GACGGCCAGTGAATCCGTAATCATG-HEG-BIOTIN-TEG
sense-linc8	GTCTACAGTAGAGAATGCAAGAGTCCATCC-HEG-BIOTIN-TEG
CHART-seq peak validation	
Tgfb3chartF2	CCCATTGGCACTCACTCC
Tgfb3chartR2	GCAAACAGCAGGAGGCTAAA
E2F2chartF1	TTCCCACCAGGGTCTCAG
E2F2chartR1	CTGGAGCGAAAGGGGAAG
ActBchartF1	GCCATCCTATCCCAAGCATA

ActBchartR1	TCTTCTTGCAACACCTCCAG
Aimp2chartF1	TGGGTGTTCCAGGAGACG
Aimp2chartR1	ACCTCACGCCTTCGGTTT
DolkchartF1	TGTAGCAGATGTTTGCCCTCT
DolkchartR1	CAAGATGGGGCTGTCAAGAT
Ube2cchartF1	TTTAATGGTCGGCGTCGT
Ube2cchartR1	AAAGCTGCCATTAACTAACGAATC
Dpm2chartF1	GACCCTCCGTTAGTGCTTGT
Dpm2chartR1	TGGCCTGGAGAAAAATAAAGG
Phox2bchartF1	AGTGGGGAGATGTGCACTG
Phox2bchartR1	GCTCGGCTGGTAGTAAGGAG
Pax6chartF1	CTCATTTCCCGCTCTGGTT
Pax6chartR1	CATTAGCGAAGCCTGACCTC
AK146698chartF1	CGGAAAGAGTGGAAGTCGAT
AK146698chartR1	CCCTCAGTGAACTCCTCCAC
Pou3f3 for1025	TGGCTCTGGGTACGCTCTATG
Pou3f3 rev1124	GGCTTGAGCTTGCACATGTTC

2.2.3.9 RNA pull-down assay using nuclear protein extract

Transcripts to be assayed were *in vitro* transcribed, purified and concentrated as described in Chapter 2.2.3.13. Purified RNA was then 5′ end labelled with biotin-maleimide using 5′ EndTag Nucleic Acid Labelling system (Vector laboratories, #MB-9001) according to the manufacturer's instructions. Briefly, 8 µl of RNA (approximately 1.5-3.0 µg) was combined with 1 µl of universal reaction buffer and 1 µl of alkaline phosphatase. The

reaction was incubated at 37 °C for 30 min. Following the incubation, 2 µl of universal reaction buffer, 1 µl ATPγS, and 2 µl T4 polynucleotide kinase were added. The sample was incubated 37 °C for a further 30 min. 10 µl of thiol-reactive label was then added, and the sample was transferred to 65 °C for 30 min. After that, 70 µl of water and 100 µl of buffered phenol were added. The sample was vortexed briefly and centrifuged to separate the phases. The aqueous phase was transferred to a fresh tube. 5 µl of precipitant and 263 µl of absolute ethanol were added. Precipitated nucleic acid was pelleted by centrifugation at full speed for 30 min at 4 °C. The pellet was washed with 70% ethanol and centrifuged again. Air-dried pellet was re-suspended in water.

2.5 µg labelled RNA was denatured at 95 °C for three min, plunged on ice, supplied with RNA structure buffer (10 mM Tris pH 7.0, 0.1 M KCl, 10 mM $MgCl_2$), and then shifted to room temperature for 20 min to allow proper secondary structure formation. Streptavidin coated Dynabeads M-280 (Life Technologies, 11205D) were washed, prepared for RNA manipulation and bound by 5′ end labelled RNA using Dynabeads kilobaseBINDER Kit (Life Technologies, #60101) according to the manufacturer's instructions. 20 µl of beads was used per immunoprecipitation. Double the amount of beads needed for immunoprecipitations was prepared so that half of them could be used for nuclear protein extract pre-clearing. Beads were incubated with 5′ end labelled RNA at room temperature for three hours on a roller. Following the incubation, beads were placed on a magnet and supernatant was discarded. Beads were washed twice in the Washing solution (10 mM Tris HCl, pH 7.5, 1 mM EDTA, 2 M NaCl) and once in water. Beads where then re-suspended in the original volume of 1x Binding Buffer (150 mM NaCl, 50 mM HEPES, pH 8.0, 0.5% Igepal, 10 mM MgCl2).

E14 ES cells were differentiated for four days with retinoic acid as described in Chapter 2.2.1.2. To each 10 cm^2 dish of cells, 500 µl of Buffer A (10 mM HEPES, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 0.05% NP-40, supplemented with protease inhibitors) were added. Cells were scraped off the dish and left on ice for 10 min. Lysates were centrifuged at 3,000 rpm for 10 min at 4 °C to pellet the nuclei. Pellets were re-suspended in 374 µl of Buffer B (5 mM HEPES, 1.5 mM $MgCl_2$, 0.2 mM EDTA. 0.5 mM DTT, 26% glycerol (v/v)) and 26 µl of 4.6 M NaCl per 10 cm^2 dish of cells. Samples were disrupted by at least 20 strokes in the Dounce homogeniser and left on ice for 30 min before centrifuging at maximum speed for 20 min at 4 °C. Supernatant representing nuclear protein extract was removed and used in further manipulations. Protein concentration was measured using BCA Protein Assay Kit (Pierce, #23225) according to the manufacturer's instructions as described in Chapter 2.2.4.1.

Protein extract was pre-cleared on beads prepared as described above for 60 min at room temperature with rotation in the presence of protease inhibitors. Pre-cleared protein extract was diluted in 2x Binding buffer 300 mM NaCl, 100 mM HEPES, pH 8.0, 1.0% Igepal, 20 mM $MgCl_2$) to a final concentration of 1x of Binding buffer. Diluted protein extract was supplemented with 100 µg/ml tRNA (Life Technologies, #AM7119), 40 U/ml RNaseOUT (Life Technologies, #10777-019), and incubated with RNA-coated beads at room temperature for two hours with rotation. 1 mg of nuclear protein extract was used per pull-down assay. Following the incubation, the sample was placed on a magnet, supernatant removed, and beads washed six times (10 min each) with 1x Binding buffer. The proteins that remained bound were eluted by incubating the beads with 20 µl of 1x Laemmli sample buffer (Sigma, #S3401) at 100 °C for 5 min. Eluted protein samples were loaded onto a 4-

20% Tris-glycine polyacrylamide gel (BioRad, #456-1093) and separated under SDS-PAGE denaturing conditions (Chapter 2.2.4.2) until the sample just entered the gel. Protein-containing bands were excised from the gel with a clean blade, diced and washed 3 times with nanopure water.

2.2.3.10 Native RNA Immunoprecipitation (RIP)

RNA Immunoprecipitation (RIP) was performed using Magna RIP™ RNA-Binding Protein Immunoprecipitation Kit (Millipore, #17-700) according to the manufacturer's instructions. Briefly, cells were harvested by trypsinization. Cell pellets were washed twice with ice-cold PBS, re-suspended in an equal volume of RIP Lysis Buffer, and pipetted several times to homogenise the mixtures. Lysates were incubated on ice if used immediately, or frozen at – 80 °C until needed. Meanwhile, 50 µl of magnetic beads per immunoprecipitation were dispensed into individual microcentrifuge tubes and washed twice with 0.5 ml of RIP Wash Buffer, vortexing briefly between washes. After the second wash, each aliquot of beads was re-suspended in 100 µl of RIP Wash Buffer, and 5 µg of the antibody of interest (or normal rabbit IgG as a negative control) was added. Beads were pre-coated with antibodies by incubating them with rotation for 30 min at room temperature. After the incubation, tubes were placed on a magnetic separator and the supernatant was discarded. Beads were washed twice with 0.5 ml of RIP Wash Buffer as before. After the second wash, beads were re-suspended in 0.5 ml of RIP Wash Buffer and left on ice while the RIP Immunoprecipitation Buffer was prepared. When ready, beads were placed on a magnetic separator and the supernatant was discarded. Each aliquot of

beads was re-suspended in 900 µl of RIP Immmunoprecipitation Buffer. RIP lysates were centrifuged at maximum speed (14,000 rpm) for 10 min at 4 °C.

A 100 µl aliquot of the supernatant was added to each beads-antibody complex in RIP Immunoprecipitation Buffer (to a final volume of 1 ml). 10 µl of the supernatant in RIP lysate buffer representing 10 % input was removed, placed into a new tube, and stored at –80 °C until the RNA purification stage. Immunoprecipitation samples were incubated overnight at 4 °C with rotation. Following the overnight incubation, the tubes were centrifuged briefly, placed on the magnetic separator, and the supernatant was discarded. Six washes with 0.5 ml of ice cold RIP Wash Buffer were performed, as described previously. After the last wash, the supernatant was removed completely, and beads were re-suspended in 150 µl of proteinase K buffer containing 117 µl of RIP Wash Buffer, 15 µl of 10% SDS, 18 µl of 10 mg/ml proteinase K. Input sample was thawed and re-suspended in 107 µl of RIP Wash Buffer, 15 µl of 10% SDS, 18 µl of 10 mg/ml proteinase K (total volume 150 µl). All samples were incubated at 55 °C for 30 min with shaking to digest the protein. After the incubation, tubes were centrifuged briefly and placed on the magnetic separator. The supernatant from each tube was transferred into a new tube. To the supernatant, 250 µl of RIP Wash Buffer was added, followed by 400 µl of phenol:chloroform:isoamyl alcohol. The samples were vortexed for 15 sec and centrifuged at maximum speed for 10 min at room temperature to separate the phases. 350 µl of the aqueous phase was carefully removed to a new tube. 400 µl of chloroform was added, followed by 15 sec vortexing and 10 min centrifugation at maximum speed at room temperature. 300 µl of the aqueous phase of each sample was removed into a new tube. To each tube, 50 µl of Salt Solution I, 15 µl of Salt Solution II, 5 µl of Precipitate Enhancer

and then 850 µl of absolute ethanol were added. Tubes were turned several times to mix and incubated overnight at – 80 °C to precipitate the RNA. Next day, samples were centrifuged at maximum speed for at least 30 min at 4 °C. Supernatant was carefully discarded and pellets were washed once with 80% ethanol. Samples were centrifuged again at maximum speed for 15 min at 4 °C. Pellets were air-dried and re-suspended in 10 to 20 µl of RNase-free water. RNA was reverse transcribed and analysed by qPCR as described previously.

2.2.3.11 UV cross-linked RNA Immunoprecipitation (UV-RIP)

Mouse neuroblastoma N2A or human neuroblastoma SH-SY5Y cells were harvested by trypsinization, washed and resuspended in ice-cold PBS (15 ml of PBS per two 80% confluent 15 cm^2 dishes worth of cells). Cell suspension was transferred into a new 15 cm^2 dish ensuring that the suspension completely covered the bottom of the dish. Cells were UV-irradiated twice at 25 nm, 400 mL/cm^2 (Stratagene Stratalinker), mixing the cells between irradiations by gently swirling the suspension. Cells and nuclei were lysed in RSB-Triton buffer (10 mM Tris-HCl, 100 mM NaCl, 2.5 mM $MgCl_2$, 35 µg/ml digitonin, 0.5% Triton X-100) with disruptive sonication (Bioruptor, 5 min, level 5). Nuclear lysates were pre-cleared on protein A/G magnetic beads (Pierce, #88802) in the presence of sheared salmon sperm DNA (Life Technologies, #AM9680) for 1 h at 4 °C and incubated with 5 µg of antibodies overnight. RNA/antibody complexes were then precipitated with protein A/G beads, washed first in low-stringency buffer (1x PBS [150 mM NaCl], 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40), then twice in a high-stringency buffer (5x PBS

[750 mM NaCl], 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40), and treated with proteinase K. RNA was extracted using TRIzol and RT-qPCR was performed as described above.

The antibodies used were: anti-DNMT1 (Abcam, #ab87656), anti-BRG1 (Abcam, #ab4081), anti-P66 beta (Abcam, #ab76924), anti-SIN3A (Active Motif, #39865), anti-CTCF (Abcam, #70303), anti-ORC3 (MyBiosource, #MBS421369), anti-POU3F3 (Santa Cruz, #sc-6028-R (C-17)), normal rabbit IgG control (Millipore) (Chaper 2.1.1).

To perform UV-RIP of FLAG-tagged POU3F3, nuclear extracts were incubated overnight with mouse monoclonal anti-FLAG M2 beads (Sigma-Aldrich) instead. Extracts from cells not expressing FLAG-tagged POU3F3 were used as control instead of isotype-matched IgG.

Table 2.13 Human qPCR oligos

Human control oligos

Oligo name	**Oligo sequence (5′ to 3′)**
Q_hsatubfor	CTTCGTCTCCGCCATCAG
Q_hsabtubrev	TTGCCAATCTGGACACCA
Q_hsRNAPIIfor	TTGTGCAGGACACACTCACA
Q_hsRNAPIIrev	CAGGAGGTTCATCACTTCACC

Human *DALI* ortholog oligos

hlinc8-6-8F	TTGAGCAAGAGTGAGGAAGATTT
hlinc8-6-8R	TGGATTTAAATAAACCTGAGTGGTT
hlinc8-68310F1	GTCTGTGCCTGCCAATCAG
hlinc8-68310R1	CAAAATGGGAGTTGACTGTGG

hlinc8-68310F2	CCCTTCACAGCTGCCTCT
hlinc8-68310R2	CAAAATGGGAGTTGACTGTGG

2.2.3.12 Transcript stability assessment by nascent transcript pulse-chase labelling

To assess the stability of transcripts of interest, nascent RNA produced in N2A cells was labelled with 0.2 mM ethinyl uridine (EU) for 24 h. After 24 h ("pulse"), the media containing the label was removed. Cells were washed twice with label-free media and cultured in EU-free media for variable amounts of time (1, 2, 4, or 6 h "chase"). Cells were harvested, and total RNA was extracted using the TRIzol method as described previously. Labelled RNA was isolated using Click-iT® Nascent RNA Capture Kit (Invitrogen; #C-10365) according to the manufacturer's instructions. The abundance of transcripts of interest in the labelled fraction was quantified by RT-qPCR.

2.2.3.13 *In vitro* transcription

DNA sequences for *in vitro* transcription were cloned into pCR4-TOPO. RNA polymerases (T3 or T7) used for *in vitro* transcription depended on the orientation of the insert and the desired strand to be transcribed. Plasmids were linearized at, or near, the terminus of the insert with a suitable enzyme. Linearized plasmid DNA was extracted with an equal volume of phenol:chloroform:isoamyl alcohol, followed by extraction with an equal volume of chloroform, prior to ethanol precipitation. Linearized and purified plasmid DNA was re-suspended in nuclease-free water at a concentration approximately 200 ng/µl. Standard 20 µl reactions were set up to contain the following quantities: 2 µl (400 ng)

linearized plasmid DNA, 20 U ribonuclease inhibitor, 2 µl 100 mM DTT, 4 µl 5x T3/T7 RNA polymerase buffer (Thermo Scientific, #EP0101; NEB, #M0251S), 4 µl 2.5 mM NTPs (NEB, #772451PK), nuclease-free water to a final volume of 19.5 µl, and 25 U (0.5 µl) of T3/T7 RNA polymerase. The reactions were incubated at 37 °C for 1 h. DNA template was removed by the addition of 25 U of RNase-free DNase I and a further incubation at 37 °C for 30 min. The reactions were extracted with phenol:chloroform:isoamyl alcohol twice, ethanol precipitated, air-dried and re-suspended in 25 µl of nuclease-free water.

2.2.3.14 Transcript copy number quantification

To quantify the population-average copy number per cell of the *Dali* transcript, a calibration curve was derived as follows. Full-length *Dali* transcript was cloned into the pCR4-TOPO vector. The resulting construct was linearized with *Pme*I (NEB, #R0560S). *Dali* RNA was *in vitro* transcribed from T7 promoter and purified as described above. The amount of RNA to be spiked into 500,000 ES E14 cells not expressing *Dali* to achieve 100, 50, 25, 10, 5, 2.5, 1 and 0.5 copies per cell was calculated using the nucleic acid calculator found at http://www.endmemo.com/bio/dnacopynum.php. Non-expressing cells provided a genome-matched background to control for losses during RNA extraction. Aliquots of 500,000 ES cells were lysed by adding 1 ml of TRIzol to the cell pellets. 1 µl of appropriate dilutions of *in vitro* transcribed *Dali* RNA was added to each cell lysate in TRIzol. Total RNA was extracted as described previously. 1 µg of total RNA was used for reverse transcription using Quantitec Reverse Transcription kit (QIAGEN, #205313) according to the manufacturer's instructions except for one modification. A mix of gene-

specific oligos (3.5 μM each) designed to prime cDNA synthesis of all transcripts of interest (*Dali* and control *Gapdh* mRNA) was used instead of primer mix supplied with the kit. Subsequently, qPCR was performed as described previously using primer pair positioned 5′ with respect to the oligo used to prime reverse transcription and non-overlapping with that oligo. The data were used to create a linear calibration curve. Total RNA was extracted in triplicate from aliquots of 500,000 N2A cells using the TRIzol method. Reverse transcription and qPCR were performed as described for the spike-in samples. *Dali* copy number was determined using the calibration curve.

2.2.4 Protein manipulation

2.2.4.1 Protein concentration measurement

Protein concentration was measured using the BCA Protein Assay Kit (Pierce, #23225) according to the manufacturer's instructions. Briefly, a set of protein standards was prepared by diluting Albumin Standard (BSA) in Diluent supplied with the kit. 25 μl of each standard and unknown sample were pipetted into individual wells of a microplate. To these, 200 μl of Working Reagent was added, and mixed well on a plate shaker for 30 sec. The plate was covered with foil and incubated at 37 °C for 30 min, and allowed to cool at room temperature. The absorbance was measured at 562 nm on a FLUOstar OPTIMA plate reader (BMG Labtech).

2.2.4.2 Protein electrophoresis

Equal amounts of protein samples (up to 15 µl) and 2x Laemmli sample buffer were mixed together, boiled at 95°C for 5 min and plunged on ice. Samples were loaded into the wells of a mini (8.6 x 6.7 cm) format 4-20% Mini-PROTEAN TGX gel (Biorad, #456-1093) along with Full Range Molecular Weight Marker Rainbow (GE Healthcare, #RPN800E). Electrophoresis was performed in 1x TBE buffer (89 mM Tris-borate, 2 mM EDTA, pH 8.3). Gels were run for 5 min at 50 V. Then, voltage was increased to 100V (constant voltage) and electrophoresis was allowed to proceed for one hour.

2.2.5 Other manipulations

2.2.5.1 TALEN-mediated up-regulation

Target regions were selected and TAL effector constructs were designed using software, tools, and information found on the TAL Effector Nucleotide Targeter 2.0 website accessible at https://tale-nt.cac.cornell.edu/. Construction of custom TALETFs designed to target promoter-proximal region of *Dali* to up-regulate transcription from the locus was performed as described by Sanjana et al. (Sanjana et al., 2012).

Briefly, bacterial clones carrying monomer plasmids were grown up, plasmid DNA was isolated (Chapter 2.2.2.13), and used to prepare a library of monomers for TALE DNA-binding domain construction. Monomer library was PCR-amplified with ligation adaptors (Table 2.14) in reactions containing 50 pg/µl monomer template plasmid, 1 mM dNTPs

(0.25 mM each), 1x Herculase II PCR buffer, 100 nM forward primer, 100 nM reverse primer, 2 µl Herculase II Fusion polymerase (per 200 µl reaction). 2x 100 µl reactions were set up for each monomer (in 2 96-well plates). Reactions were carried out using the following thermo-cycler programme: 95 °C 2 min (initial denaturation), 30 cycles (95 °C, 20 sec; 60 °C, 20 sec; 72 °C, 10 sec), 72 °C 3 min (final extension), 4 °C (store). Selected products (representing each template and each primer pair, no need to run all products) were run on 2% TAE gel to check for successful amplification. Both 100 µl reactions for each monomer were pooled and PCR-purified using QIAquick PCR purification columns. Products were initially eluted in 100 µl. DNA concentrations were measured on a NanoDrop and then adjusted to 18 ng/µl (for monomers 1, 6, 7, 12, 13, 18) or 15 ng/µl (the other monomers) to be used for hexamer assembly.

TALE DNA-binding domain sequence was divided into hexamers. Each hexamer was assembled using Golden Gate digestion-ligation approach in 10 µl reactions containing 0.375U/µl *Esp*3I (*Bsm*BI), 1x Tango buffer, 1 mM DTT, 75 U/µl T7 ligase, 1 mM ATP, and 1 µl of each monomer product prepared previously. Reactions were incubated in a thermo-cycler as follows: 15 cycles (37 °C, 5 min (digest); 20 °C, 5 min (ligate)), 4 °C (store). Ligation products were run on a 2% TAE agarose gel to check for approximately 700 bp bands corresponding to hexamer products. Incomplete (non-circular, as only complete ligation products of the Golden Gate reactions are able to circularize) ligation products were degraded by adding 1 µl each PlasmidSafe DNase, 10x PlasmidSafe readtion buffer, and 10 mM ATP to 7 µl of each hexamer reaction and incubating the tubes at 37 °C for 30 min, followed by 30 min at 70 °C to inactivate the enzyme. Each PlasmidSafe-treated hexamer was amplified in a 50 µl PCR reaction using high-fidelity

Herculase II polymerase and hexamer forward and reverse primers (Hex-F and Hex-R, Table 2.14). Each PCR reaction consisted of 1mM dNTPs (0.25 mM each), 1x Herculase II reaction buffer, 200 nm primer mix, 1x Herculase II Fusion DNA polymerase, 1 µl of PlasmidSafe-treated hexamer and nuclease-free water to 50 µl total volume. PCR reactions were performed using the following cycling conditions: 95 °C, 2 min; 35 cycles (95 °C, 20 sec; 60 °C, 20 sec; 72 °C 30 sec); 72 °C 3 min; 4 °C hold. Each hexamer was gel-purified (Chapter 2.2.2.16) from other non-specific amplicons and its concentration was adjusted to 20 ng/µl.

To assemble the full TALE backbone, hexamers and the appropriate TALE backbone vector were combined in a Golden Gate digestion-ligation reaction that included 100 ng of TALE backbone vector, 15 U *Bsa*I-HF, 1x NEBuffer 4, 1x BSA, 1mM ATP, 750 U T7 Ligase, 20 ng each of the three hexamers, and nuclease-free water to 10 µl total volume. A control reaction without hexamers was also set up. The reactions were carried out using the following cycling conditions: 20 cycles (37 °C, 5 min; 20 °C, 5 min); 80 °C, 20 min; 4 °C hold. 5 µl of ligation reactions were transformed into chemically competent bacterial cells (Chapter 2.2.1.7). Colony PCR was performed to verify correct assembly of the TALE backbone. PCR reactions contained 1 µl of 100 µl colony suspension in distilled water, 0.8 mM dNTPs (0.2 mM each), 1x AmpliTaq 360 Buffer, 1.5 mM $MgCl_2$, 0.2 µM forward primer (TALE-Seq-F1), 0.2 µM reverse primer (TALE-Se-R1), 1.25 U AmpliTaq 360 DNA Polymerase, and water to 25 µl total volume. PCR reactions were performed using the following cycling conditions: 94 °C, 3 min; 30 cycles (94 °C, 30 sec; 60 °C, 72 °C, 2 min); 72 °C, 5 min; 4 °C hold. PCR products were run on a 1% TAE agarose gel. Positive colonies were grown up, plasmid DNA was isolated (Chapter 2.2.2.13) and sequenced

using TALE-Seq-F1, TALE-Seq-F2, TALE-Seq-R1 (or TALE-Seq-R2 for TALEs with more than 18 full monomer repeats).

Table 2.14 TALEN oligos

Colony PCR

Oligo name	**Oligo sequence (5′ to 3′)**
pCR8_F1	TTGATGCCTGGCAGTTCCCT
pCR8_R1	CGAACCGAACAGGCTTATGT
TAL_F1	TTGGCGTCGGCAAACAGTGG
TAL_R2	GGCGACGAGGTGGTCGTTGG

Monomer amplification

Ex-F1	TGCGTCCGTCTCCGAACCTTAAACCGGCCAACATACCGGTCTCCTGACCCCAGAGCAGGTCGTG
Ex-F2	TGCGTCCGTCTCCGAACCTTAAACCGGCCAACATACCGGTCTCGACTTACACCCGAACAAGTCGTGGCAATTGCGAGC
Ex-F3	TGCGTCCGTCTCCGAACCTTAAACCGGCCAACATACCGGTCTCGCGGCCTCACCCCAGAGCAGGTCG
Ex-F4	TGCGTCCGTCTCCGAACCTTAAACCGGCCAACATACCGGTCTCGTGGGC
	TCACCCCAGAGCAGGTCG
Ex-R1	GCTGACCGTCTCCGTTCAGTCTGTCTTTCCCCTTTCCGGTCTCTAAGTC
	CGTGCGCTTGGCAC
Ex-R2	GCTGACCGTCTCCGTTCAGTCTGTCTTTCCCCTTTCCGGTCTCAGCCGT
	GCGCTTGGCACAG
Ex-R3	GCTGACCGTCTCCGTTCAGTCTGTCTTTCCCCTTTCCGGTCTCTCCCAT
	GGGCCTGACATAACACAGGCAGCAACCTCTG

Ex-R4	GCTGACCGTCTCCGTTCAGTCTGTCTTTCCCCTTTCCGGTCTCTGAGTC CGTGCGCTTGGCAC
In-F2	CTTGTTATGGACGAGTTGCCCGTCTCGTACGCCAGAGCAGGTCGTGGC
In-F3	CCAAAGATTCAACCGTCCTGCGTCTCGAACCCCAGAGCAGGTCGTG
In-F4	TATTCATGCTTGGACGGACTCGTCTCGGTTGACCCCAGAGCAGGTCGTG
In-F5	GTCCTAGTGAGGAATACCGGCGTCTCGCCTGACCCCAGAGCAGGTCGTG
In-F6	TTCCTTGATACCGTAGCTCGCGTCTCGGACACCAGAGCAGGTCGTGGC
In-R1	TCTTATCGGTGCTTCGTTCTCGTCTCCCGTAAGTCCGTGCGCTTGGCAC
In-R2	CGTTTCTTTCCGGTCGTTAGCGTCTCTGGTTAGTCCGTGCGCTTGGCAC
In-R3	TGAGCCTTATGATTTCCCGTCGTCTCTCAACCCGTGCGCTTGGCACAG
In-R4	AGTCTGTCTTTCCCCTTTCCCGTCTCTCAGGCCGTGCGCTTGGCACAG
In-R5	CCGAAGAATCGCAGATCCTACGTCTCTTGTCAGTCCGTGCGCTTGGCAC

Hexamer amplification

Hex-F	CTTAAACCGGCCAACATACC
Hex-R	AGTCTGTCTTTCCCCTTTCC

Sequencing

TALE-Seq-F1 (aka colony PCR forward)	CCAGTTGCTGAAGATCGCGAAGC
TALE-Seq-F2	ACTTACACCCGAACAAGTCG
TALE-Seq-R1 (aka colony PCR reverse)	GCCACTCGATGTGATGTCCTC
TALE-Seq-R2	CCCATGGGCCTGACATAA

2.2.5.2 Mass spectrometry

Mass spectrometry of protein samples obtained in the RNA pull-down assay (Chapter 2.2.3.9) was performed at the Central Proteomics Facility, Sir William Dunn School of Pathology, University of Oxford.

2.2.6 Animal work

2.2.6.1 Mouse brain dissection

All animal experiments were conducted in accordance to schedule one UK Home Office guidelines (Scientific Procedures Act, 1986). C57BL/6J, postnatal day P56 male and pregnant females were killed by cervical dislocation; whole brains were dissected in ice-cold phosphate-buffered saline (PBS) from adult (n=2), and intrauterine stages E9 (n=6), E10.5 (n=6), E13.5 (n=6), E15.5 (n=6) and E18.5 (n=6) mice. Brains were embedded in 5% agarose (low melting, Bioline) and sectioned using a vibrating microtome (Leica, VT1000S) into 200 μm coronal sections using a chilled solution of 1:1 mixture of RNAlater (Ambion) and PBS. Regions of interest (adult: dentate gyrus, subventricular zone and olfactory bulb; embryos: preplate, proliferative compartments combining ventricular and subventricular zones, and cortical plate from lateral and dorsal tiers) were dissected from individual sections using 27 gauge needles under visual guidance, using transillumination on a dissecting microscope (MZFLIII, Leica). Dissected samples were rinsed in RNAse free PBS/RNAlater 1:1, submerged in ice-cold RNAlater, kept for 24 hours at 4°C and stored at -80°C in RNAlater until processing.

2.2.7 Computational analysis

2.2.7.1 Transcriptomic analysis

Affymetrix CEL files were analysed using the Limma, oligo, and genefilter R Bioconductor packages (Carvalho and Irizarry, 2010; Smyth, 2004). Arrays were RMA background corrected and quantile normalised. Summary expression values were calculated for Refseq and full length mRNAs. Genes whose expression changed upon *Dali* and *Pou3f3* knockdown, as well as upon RA induced differentiation of control and stable *Dali* knockdown cells, were filtered to remove genes showing little variation in expression (variance cut off of 0.5) before the identification of significant changes. In every case, the Limma Ebayes algorithm was used to identify differential expression between three knockdown and three control samples (biological replicates). Multifactorial analysis to identify genes differentially responding to retinoic acid treatment was performed using the Limma R Bioconductor package. GOToolbox was used to perform Gene Ontology analyses ((Martin et al., 2004); http://genome.crg.es/GOToolBox/). Representative significantly enriched categories were selected from a hypergeometric test with a Benjamini-Hochberg corrected P-value threshold of 0.05.

2.2.7.2 CHART-Seq analysis

50 bp, paired-end reads were mapped to the mouse genome (mm9) using bowtie with the options '–m1 –v2 –best –strata –a'. For each *Dali* sample, peaks were called against the

matched N2A input sample (4208 peaks) and against two *lacZ* controls (1928 peaks). Peak calls were made using the MACS2 algorithm ((Zhang et al., 2008); https://github.com/taoliu/MACS/blob/master/README) with the options '–mfold 10 30 –gsize=2.39e9 –qvalue=0.01' using the CGAT pipeline 'pipeline_mapping.py' (https://github.com/CGATOxford/cgat). Peak calls were then filtered such that only peak calls with a −log10 q value > 5 were retained (FDR 0.001%).

2.2.7.3 Characterisation of *Dali* binding sites

The chromosomal distribution of *Dali* peaks was visualised using the R Bioconductor package 'ggbio' (Yin et al., 2012). Genome territory enrichment analysis was performed using the Genome Association Tester (GAT; (Heger et al., 2013)). 10,000 simulations were performed using a mappability filtered workspace and an isochore file partitioning the genome into 8 bins based on regional GC content. For the chromosomal enrichment analyses, chromosomal territories were proportionally assigned to a single virtual meta-chromosome before using GAT to test for GC and mappability corrected enrichments as above. Peak shapes were visualised using read count normalised (MACS2–SPMR), background subtracted (MACS2–bdgcmp) coverage tracks. Regions covering peaks were extracted and centred on the location of the peak maximum. Gene Ontology categories enriched for *Dali* binding were identified by intersecting regulatory regions for known coding genes with *Dali* binding sites. Regulatory regions for genes were defined following the GREAT definition (McLean et al., 2010) as a basal domain surrounding the TSS (from −5 kb to +1 kb) and extending domains upstream and downstream to the nearest gene's

basal domain or to a maximum distance of 1 Mb. Enrichments were identified using GOToolbox.

Dali peaks were characterised using DNase I hypersensitivity (HS) data generated by the Stamatoyannopoulos lab at the University of Washington and chromatin features identified by the Ren lab at the Ludwig Institute for Cancer Research (ENCODE, 2011; Shen et al., 2012). Enrichments of DNase I HS and chromatin features overlapping *Dali* peaks were assessed using GAT to control for mappability and regional GC content as above.

Complementarity between *Dali* sequence and binding locations was assessed using the EMBOSS Water algorithm (Rice et al., 2000) which performs Smith-Waterman alignment with a range of gap opening and extension penalties. RNA-DNA:DNA triplex formation was assessed using the Triplexator search software suit (Buske et al., 2012). The MEME-ChIP (Machanick and Bailey, 2011) algorithm was used to perform *de novo* motif discovery analysis by examining the unmasked DNA sequence of the central regions of peak locations. MEME-ChIP was run with the options '-meme-mod zoops -meme-minw 5 -meme-maxw 30–meme-nmotifs 50' using a custom background file prepared from regions flanking the peak locations using the command 'fasta-get-markov -m 2'. Enrichment of known vertebrate transcription factor binding sites from the TRANSFAC Professional database (Matys et al., 2006) was assessed using the AME algorithm (McLeay and Bailey, 2010) with the options '–method fisher –length-correct' using the sequence and background file prepared for MEME-ChIP analysis.

Chapter 3 LncRNA loci often occur in the genomic vicinity of transcription factor genes and are co-expressed with them temporally and spatially.

3.1 Introduction

Thousands of long non-coding transcripts have been described to date from various eukaryotic genomes (Cabili et al., 2011; Carninci et al., 2005; Derrien et al., 2012; Young et al., 2012). Many of these transcripts originate from loci outside of protein-coding genes (Marques et al., 2013; Tang et al., 2013). The vast majority of these are of unknown function due to experimental studies lagging behind largely computational efforts. In fact, what proportion of these transcripts are functional is a subject of much debate and on-going research (Nobrega et al., 2004; Rands et al., 2014; Sauvageau et al., 2013)). Several criteria have been proposed to identify the likely functional lncRNAs for functional studies. One of the most widely accepted hallmarks of functionality is evolutionary conservation (Guttman et al., 2009; Ponjavic et al., 2007). In 2009, Ponjavic and colleagues identified 659 evolutionarily constrained long ncRNAs significantly enriched in predicted RNA secondary structure (Ponjavic et al., 2009). The study found that these transcribed loci tended to reside in close genomic proximity to either CNS-expressed developmental transcriptional factor genes or signalling genes expressed outside CNS in the adult. Furthermore, lncRNAs and their neighbouring protein-coding genes were both spatially and temporary co-expressed with each other when examined experimentally for a small number (11 out of 12 tested) of lncRNA- protein-coding gene pairs. The authors hypothesized that such spatiotemporal co-expression of lncRNAs and protein-coding gene neighbours might represent a general phenomenon, which is particularly pronounced in the

brain. These findings led to a hypothesis that intergenic lncRNAs encoded close to developmental neuronal transcription factors may be involved in regulating these genes *in cis*. If so, such regulation may occur via transcript-dependent or transcript-independent modes of action. LncRNAs can also regulate smaller genomic regions *in cis*. For instance, enhancer-like RNAs have been reported to activate genes in their genomic vicinity in a transcript-dependent manner (Orom et al., 2010). Nascent *HOTTIP* lncRNA transcripts originating from the *HOXA* cluster also act *in cis* to activate transcription of the neighbouring genes by recruiting the MLL complex that deposits H3K4me3 marks on promoters of these genes (Wang et al., 2011b). A transcription-dependent but transcript-independent mode of action was proposed for the intergenic lncRNAs originating from the Polycomb Response Elements (PREs) in the *BITHORAX* locus in *Drosophila*. The act of transcription itself was postulated to evict repressive chromatin complexes and, therefore, activate transcription of the genes in the locus (Schmitt et al., 2005).

Here, I investigated six genomic loci composed of a lncRNA locus or loci and a transcription factor protein-coding gene in its vicinity. I was looking for indications that lncRNAs in these loci might be functional and involved in regulating their neighbouring protein-coding genes *in cis*. I performed a screen to select a candidate locus for further detailed functional studies. The screen aimed to establish whether the lncRNA locus in question lies fully outside of known protein-coding genes nearby, whether the lncRNA loci and the protein coding neighbouring genes are co-expressed spatially in adult mouse organs and temporally during neuronal differentiation of ES cells in culture, and whether the lncRNA is amenable to shRNA-mediated knockdown and is likely to regulate its protein-coding gene neighbour.

Overall, in agreement with the earlier findings by Ponjavic et al., I found that lncRNAs considered here do tend to be co-expressed both spatially and temporarily with their neighbouring transcription factor genes. Also, although my candidate lncRNAs were previously found to evolve under individual evolutionary constraint, I observed that their sequence is conserved in a patchy manner, with regions of relatively high conservation interspersed with often larger regions of much more divergent sequence. Only one of my candidate lncRNAs (namely *linc12*) contained more than one exon. In this case, both exon-intron junctions coincided with short regions of high sequence conservation (Haerty and Ponting, 2014). The rest of the transcript displayed the same patchy conservation pattern described above. This observation contrasts with the conservation pattern of the neighbouring protein-coding genes. The monoexonic *Pou3f3* gene exhibited high conservation throughout the ORF, with the lncRNA-like patchy pattern observed only in the 3′ UTR. The multi-exonic protein-coding genes had highly conserved exons and exon-intron junctions, with the patchy pattern observed in the introns. However, it should be noted that for all the lncRNAs examined, putative promoter regions exhibited the highest sequence conservation, which may indicate conservation of expression profiles (similar to the early studies by the FANTOM project (Carninci et al., 2005)).

In terms of subcellular localisation, all four of the intergenic lncRNA candidates from the Ponjavic et al. dataset remaining after the screening procedure were enriched in the chromatin fraction of the nucleus, which may support the hypothesis by the original authors that these transcripts may play a role in transcriptional regulation of their neighbouring genes. Only one of the intergenic lncRNAs (not from the Ponjavic et al.

dataset, but chosen because it resides in one of the candidate loci, *Pou3f3*) was found to be localised predominantly to the cytoplasm.

As the result of the screen, I identified a capped and polyadenylated lncRNA that I termed *Dali* that satisfied all of the selection criteria, namely fully intergenic; positively correlated in expression with its neighbouring gene temporarily and spatially; and its susceptibility to shRNA-mediated knockdown.

3.2 Results

3.2.1 LncRNA loci found next to CNS-expressed transcription factor genes are often spatially and temporally co-expressed

This project started with six candidate loci containing a transcription factor gene and putatively functional lncRNA locus in its genomic vicinity selected from the Ponjavic et al. (Ponjavic et al., 2009) dataset (Table 3.1). The Ponjavic et al. dataset contained 659 transcripts subselected from mouse cDNA libraries (Carninci et al., 2005; Okazaki et al., 2002) as being CNS-expressed, putatively intergenic, and unlikely to code for proteins. To exclude the possibility that signatures of purifying selection on these transcripts resulted from them containing unannotated protein-coding exons, the authors also removed candidates if BLASTX (Altschul et al., 1990) analysis returned a significant hit of their nonrepetitive sequence to a known protein in the National Center for Biotechnology Information's nonredundant (nr) protein database (E-value < 10^{-3}, hits to "hypothetical," "unknown," or unnamed sequences were not considered). I further confirmed that

transcripts encoded by the selected loci are most likely to be non-coding using the Coding Potential Assessment Tool (CPAT) (Table 3.2). These loci were selected on the basis of the developmental importance of the transcription factor genes they contained. Consequently, if the hypothesis proposed by Ponjavic et al. that the intergenic lncRNA transcribed in their vicinity may be involved in regulating the expression of these transcription factors, it could be important to study the molecular mechanisms of the lncRNAs from these loci. Because the aim of the project was to focus on intergenic transcripts that might be acting in a transcript-dependent manner via novel mechanisms (as opposed to mechanisms already described for antisense transcripts that regulate their overlapping genes *in cis*), I performed a screen to identify candidates for further in depth functional study.

Table 3.1 Genomic coordinates of candidate lncRNAs

ncRNA	ncRNA ID	Chromosomal coordinates (mm9)	strand	Locus 5' end of	Locus 3' end of
linc4	*AK042766*	chr11: 18769475-18770502	-		*Meis1*
linc5	*AK018196*	chr6: 97970397-97971348	+		*Mitf*
linc7	*AK078612*	chr18: 76469748-76471401	+		*Smad2*
linc8 (*Dali*)	*AK034039*	chr1: 42807557-42809726	+		*Pou3f3*
linc8b	*AK053590*	chr1: 42758001-42759804	+		*Pou3f3*
linc9	*AK090266*	chrX: 99447990-99453028	+	*Cited1*	
linc12	*AK144485*	chr2: 115893704-115911646	+	*Meis2*	
Not in Ponjavic et al., 2009 dataset					
linc8u (l*inc-Brn1a*)	*AK019154*	chr1: 42705045-42751670	-	*Pou3f3*	

Table 3.2 Protein coding potential of candidate lncRNAs using CPAT

Transcript	RNA size, nt	ORF size, aa	Ficket score	Hexamer score	Coding probability	Coding label
linc4	2008	198	0.6918	-0.115192479811	0.045050913054107	no
linc5	952	114	0.7398	-0.0944428857959	0.029922329069065	no
linc7	1654	309	0.4939	-0.238419416387	0.044056889540756	no
Dali (linc8)	2580	183	0.4436	-0.27522605573	0.014060822434686	no
linc8b	1804	282	0.8828	-0.204152137395	0.084342274202671	no
linc8u	569	198	0.881	-0.00563049992337	0.095088836448284	no
linc9	1761	270	0.5816	0.129155362141	0.13666398705341	no
linc12	3399	162	0.496	-0.594818411902	0.004123901393116	no

The first stage in the initial screen designed to test the working hypothesis that such transcripts function *in cis* and might regulate their neighbouring genes was to confirm co-expression of these transcripts, both spatial and temporal. To test temporal co-expression, an *in vitro* system of mouse ES cell retinoic acid (RA)-induced differentiation into the neuronal lineage was established (Chapter 2.2.1.2). In this protocol, mouse ES cells treated with RA give rise first to cycling neuronal progenitors (day 2-3) which then undergo cell cycle exit (day 4-5) and commit to a terminal neuronal fate. Fully differentiated neurons first appear at around day 4-5 and by day 6-7 the culture largely consists of neurons, and then gradually becomes populated with proliferating mesodermal cell types(Smith, 1991).

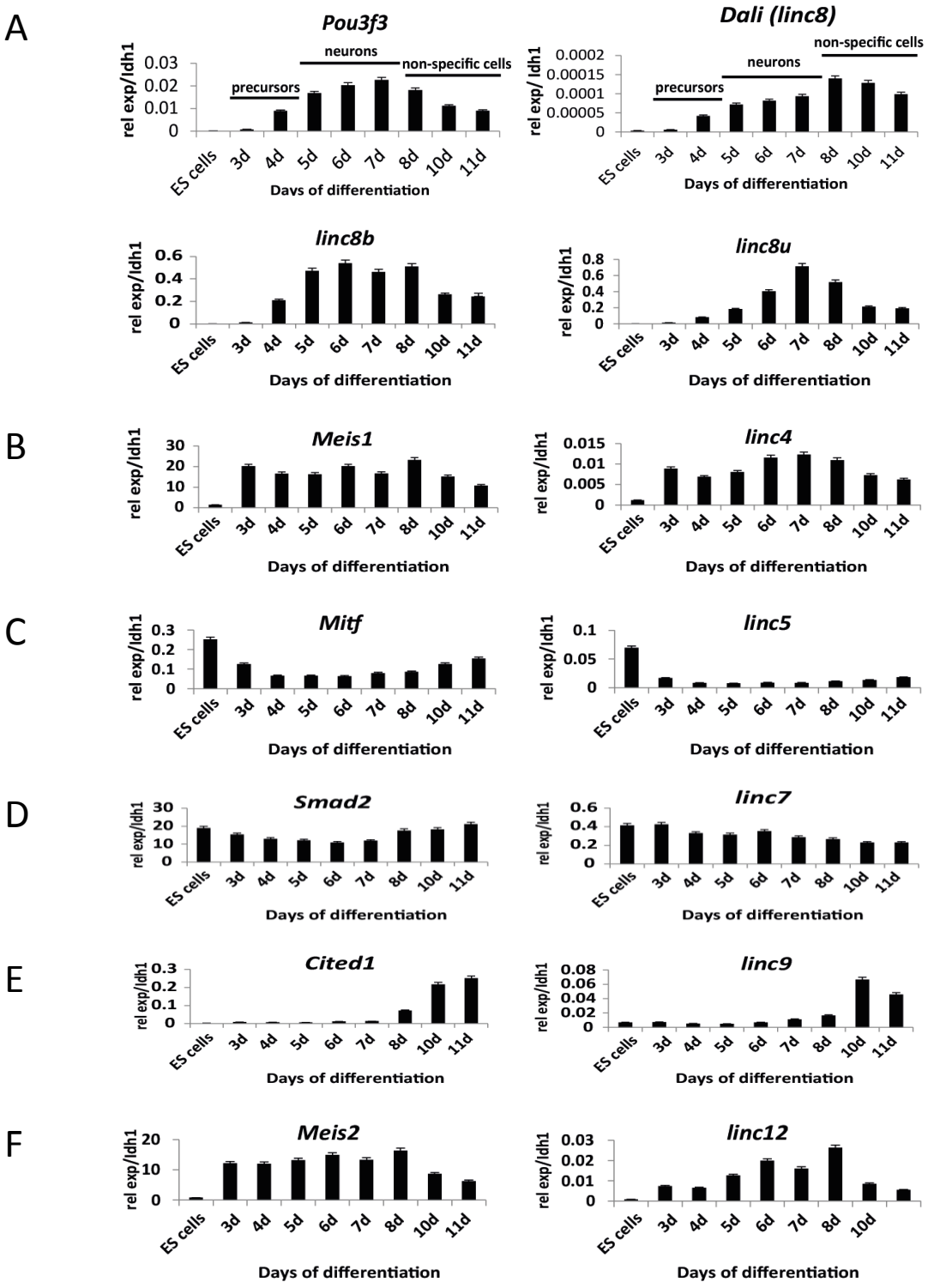


Figure 3.1 RA-induced differentiation of mouse ES cells: lncRNA (left) and neighbouring TF gene expression. Neuronal differentiation of mouse ES cells was induced using RA. The levels of lncRNAs and TF mRNAs were measured in total RNA samples by qRT-PCR. Results are presented relative to an *Idh1* reference gene which does not change significantly during differentiation. Mean +/- s.e., n = 3.

I used this system to monitor the expression of both coding and non-coding transcripts from the chosen six loci. The results of this experiment are shown in Figure 3.1 and reveal good concordance of expression profiles of the transcription factor genes and lncRNAs for all loci tested. The exception (*linc7-Smad2*) was expressed at the same levels across the whole duration of the experiment. Thus, these loci exhibited evidence of regulation of expression during the time-course. As tight regulation in development is considered to be one of the hallmarks of lncRNA functionality, I excluded the *linc7- Smad2* locus from further study. I also excluded the *linc9-Cited1* locus, as the onset of expression of its transcripts lay beyond the time-frame of neuronal differentiation (Figure 3.1 E). Among the remaining loci, *Mitf-linc5* locus transcripts were expressed highest in ES cells and were down-regulated with the onset of differentiation. *Meis1-linc4*, *Pou3f3-linc8* (*Dali*), and *Meis2-linc12* transcripts were absent from ES cells, and were rapidly up-regulated upon differentiation. The expression of *Meis1-linc4* and *Meis2-linc12* loci preceded that of the *Pou3f3-linc8* (*Dali*) locus, which was induced at day 4 at the stage when neuronal progenitors start exiting the cell cycle and the first few mature neurons appear (Figure 3.1 A).

To examine spatial expression profiles of the candidate loci, I assayed the expression of coding and non-coding transcripts in a panel of mouse adult tissues by RT-qPCR (Figure 3.2; Chapter 2.2.3.6). LncRNA-protein coding gene transcripts generated from the *Meis1*, *Meis2*, and *Pou3f3* loci were correlated in their tissue restricted expression patterns and their relative expression between tissues (Figure 3.2 A, C, D). For *Mitf* and *linc5*, a good correlation of expression levels was observed in those tissues where both transcripts were

detectable. However, in skeletal muscle, small intestine, stomach, and testes, only *Mitf* mRNA was expressed, while *linc5* expression was undetectable (Figure 3.2 B).

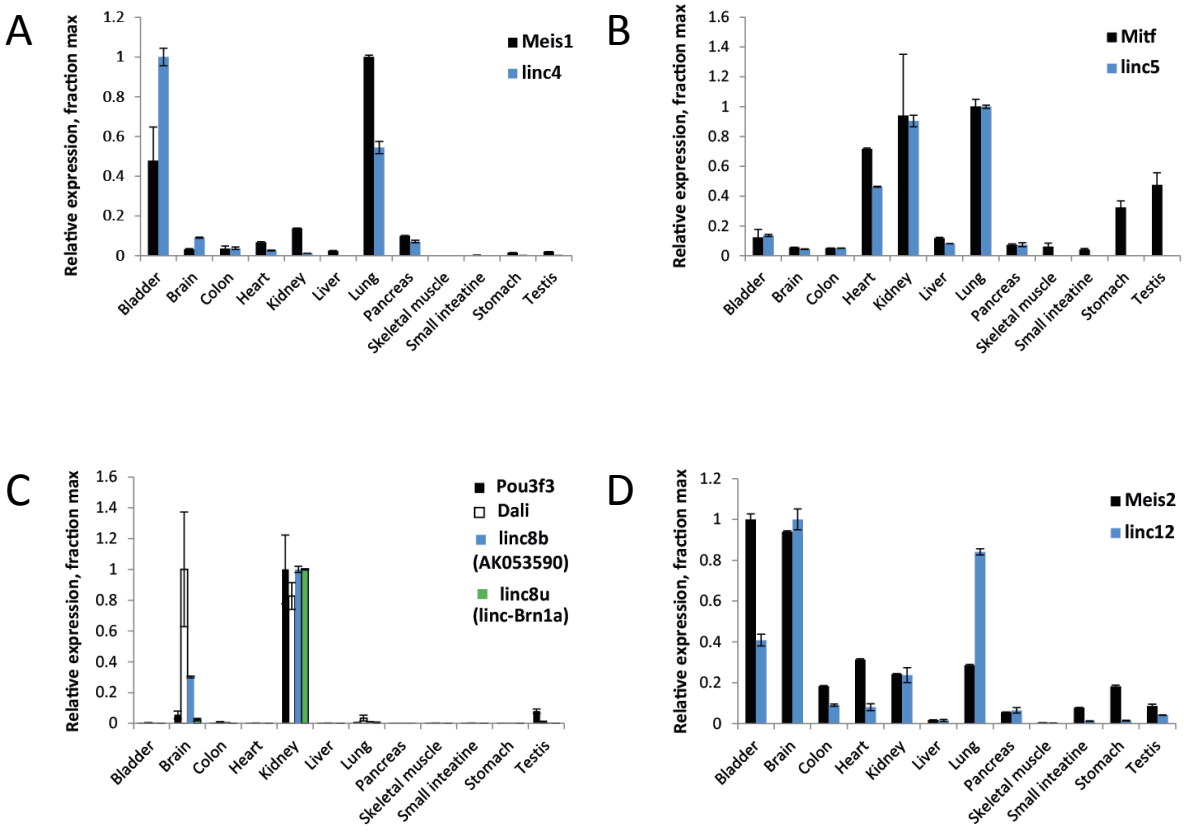


Figure 3.2 Adult mouse organs: lncRNA and neighbouring TF gene expression. The levels of lncRNAs and TF mRNAs were measured in a panel of adult mouse tissues by quantitative RT-PCR (qRT-PCR). Results were normalized by the average value of *Gapdh* and *Tbp* reference genes. Mean values +/- standard error (s.e.) shown, n = 3 replicates.

Therefore, for the four remaining loci, *Meis1-linc4*, *Mitf-linc5*, *Pou3f3-Dali*, and *Meis2-linc12*, I observed both temporal and spatial co-expression between the protein-coding and

lncRNA transcripts. However, as exemplified by *Mitf* and *linc5*, lncRNAs may display a more tissue-restricted expression pattern than their neighbouring protein-coding genes.

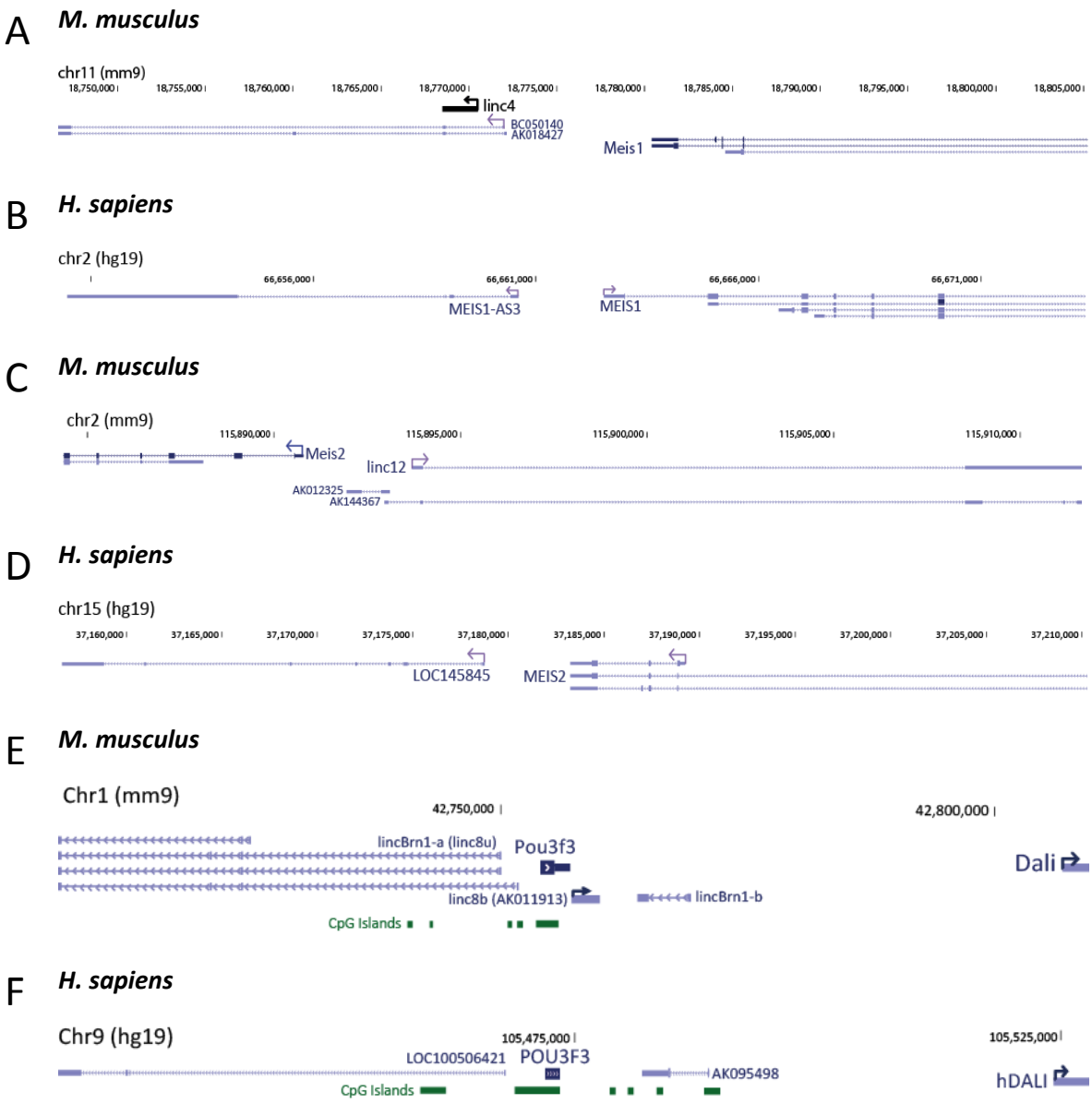


Figure 3.3 Transcription of the candidate TF-lncRNA loci in mouse and human. Schematic representations of the candidate loci in mouse and human (UCSC genomic browser). The positions of lncRNAs and neighbouring TF genes are shown. Mouse *Meis1* is orthologous to human *MEIS1*, *Meis2* to *MEIS2*, *Pou3f3* to *POU3F3*. In both species, lncRNAs are found in similar chromosomal positions relative to a protein-coding gene, i.e. display conserved synteny.

3.2.2 Genomic loci of CNS-expressed transcription factors often exhibit syntenic transcription of lncRNAs

If these evolutionarily constrained intergenic lncRNAs act to regulate their protein-coding neighbours with which they are co-expressed, one would expect to find them in an equivalent, or similar, position in other mammalian genomes. I therefore examined the genomic structures of the human MEIS1, MEIS2, and POU3F3 loci, orthologous to the mouse *Meis1*, *Meis2*, and *Pou3f3* genes, respectively, in the UCSC genomic browser and found that lncRNA transcription in their vicinity is preserved between mouse and human (Figure 3.3, above). Moreover, the transcript structures of the orthologous lncRNAs were similar between the two species (Figure 3.3). Therefore, the chosen loci exhibit syntenic transcription, a second "dimension of non-coding RNA conservation", which may be indicative of these lncRNA transcripts being functional.

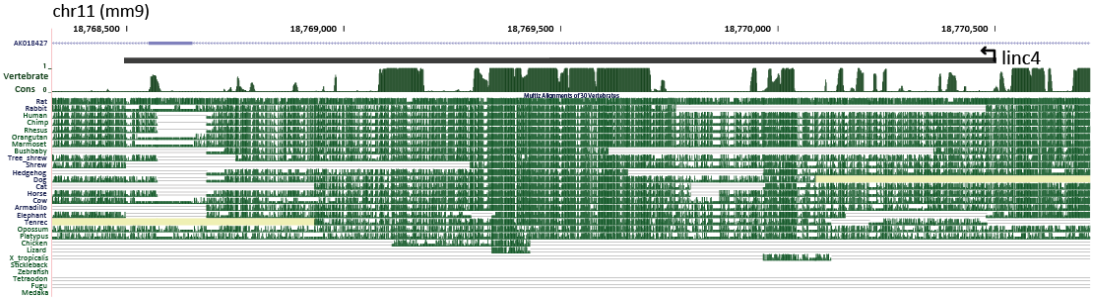
3.2.3 Conservation of lncRNA sequence is confined to short stretches interspersed with regions of more divergent sequence

As sequence conservation is considered one of the major hallmarks of functionality, I examined the patterns of sequence conservation of *linc4*, *Dali*, and *linc12* using MultizAlignments of 60 Vertebrates and Placental Mammal Conservation by PhastCons tracks of the UCSC genomic browser and compared these to sequence conservation patterns of their neighbouring protein-coding genes (Figure 3.4). I noted that all three lncRNAs had regions well conserved across mammals just upstream of their transcriptional start sites (TSSs), which might indicate potential conservation of expression. The transcript

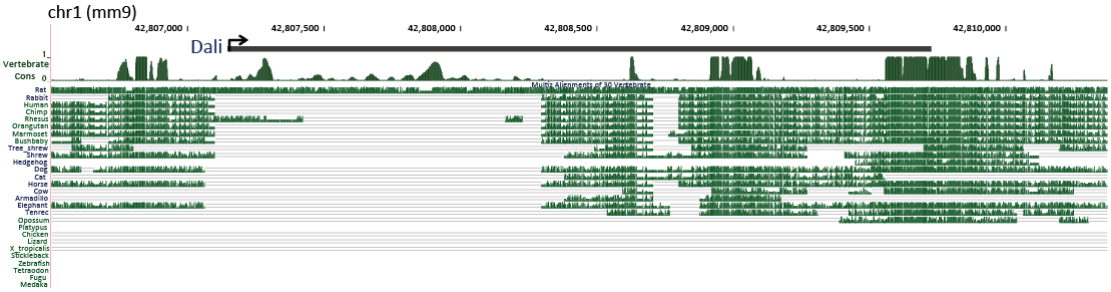
bodies of mono-exonic *linc4* and *Dali* displayed regions of relatively high conservation in therian mammals interspersed with much more divergent regions. Portions of the *linc4* transcript could also be aligned to platypus, as well as chicken, lizard and *Xenopus* genomes (Figure 3.4 A). *Dali* is somewhat less conserved, with portions of it aligned to the opossum genome, but not to more distantly related non-mammalian vertebrate genomes (Figure 3.4 B). In contrast, the mono-exonic protein-coding gene *Pou3f3* displayed high sequence conservation throughout its coding region as well as the 3′ portion of its 3′ UTR. The pattern of conservation observed across the 5′ portion of the 3′ UTR is patchy and resembles that of *linc4* and *Dali*. The overall sequence conservation of *Pou3f3* is higher than that of the lncRNAs discussed here, with most of its sequence aligned to not only mammalian, but also other vertebrate genomes (Figure 3.4 E).

Linc12 contains two exons. However, its conservation pattern is different from multi-exonic genes considered here, *Meis1* and *Meis2* (Figure 3.4 C, D, F). While both of the protein-coding genes had regions of higher conservation scores corresponding to their exons, as well as flanking regions presumably containing splicing regulatory elements (Figure 3.4 D, F), both exons of *linc12* were not better conserved than its intron. However, a higher conservation score was indeed observed just upstream of its TSS, as well as adjacent to both exon-intron junctions (Figure 3.4 C). A pattern of short conserved regions interspersed between low conservation regions resembling that of *linc4* and *Dali* was observed across the much longer second exon. This is in agreement with previous studies showing that at most only 5% of lncRNA sequence is conserved (Ponjavic et al., 2007).

A


chr11 (mm9)
18,768,500
18,769,000
18,769,500
18,770,000
18,770,500
linc4
Vertebrate
Cons
Multiz Alignments of 30 Vertebrates

B


chr1 (mm9)
42,807,000
42,807,500
42,808,000
42,808,500
42,809,000
42,809,500
42,810,000
Dali
Vertebrate
Cons
Multiz Alignments of 30 Vertebrates

C


chr2 (mm9)
115,894,000
115,896,000
115,898,000
115,900,000
115,902,000
115,904,000
115,906,000
115,908,000
115,910,000
AK144367
2810405F15Rik
linc12
Vertebrate
Cons
Multiz Alignments of 30 Vertebrates
chr2 (mm9)
115,893,400
115,893,500
115,893,600
115,893,700
115,893,800
115,893,900
115,894,000
115,894,100
AK144367
linc12
Vertebrate Cons
Vertebrate Conservation by PhastCons
chr2 (mm9)
115,908,000
115,908,500
115,909,000
115,909,500
115,910,000
115,910,500
115,911,000
115,911,500
AK144367
linc12
Exon2
Vertebrate Cons
Vertebrate Conservation by PhastCons

D

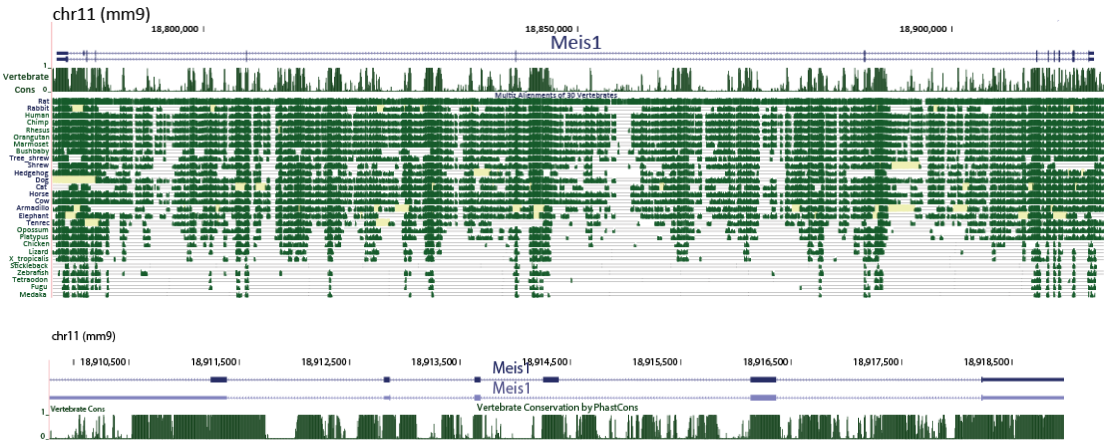

chr11 (mm9)
18,800,000
18,850,000
18,900,000
Meis1
Vertebrate
Cons
Multiz Alignments of 30 Vertebrates
chr11 (mm9)
18,910,500
18,911,500
18,912,500
18,913,500
18,914,500
18,915,500
18,916,500
18,917,500
18,918,500
Meis1
Meis1
Vertebrate Cons
Vertebrate Conservation by PhastCons

Figure 3.4 Sequence conservation of the candidate lncRNAs and neighbouring TF genes in vertebrate evolution. A detailed view of the mouse *linc4*, *Dali*, *linc12*, and their neighbouring *Meis1*, *Pou3f3*, and *Meis2* gene loci indicating regions of vertebrate DNA sequence conservation (Multiz Alignments of 60 Vertebrates and Placental Mammal Conservation by PhastCons tracks from the UCSC brwoser). Promoter region appears to be most conserved for all lncRNAs shown. Within the transcript body, highly conserved patches are interspersed with regions or more divergent sequence. For multi-exonic *linc12* (and its isoform, above), exon-intron boundaries appear more conserved (A, middle and bottom panels). Mono-exonic *Pou3f3* is highly conserved throughout the ORF, with lncRNA-like patchy conservation in the 3′ UTR (E). For the multi-exonic protein-coding *Meis1* and *Meis2*, exons and exon-intron boundaries are highly conserved, with patchy conservation across introns (D, F).

Taken together, these observations suggest that in the case of lncRNAs, it is the expression pattern and possibly the secondary structure and/or presence of functional motifs (such as RNA-binding protein interaction elements, etc) in the transcript which are important for function and are, therefore, preserved in evolution. Similar observations on the genome-

wide scale led others to propose that lncRNAs are modular in structure and function (Johnsson and Morris, 2014; Mercer and Mattick, 2013). This means that these transcripts are composed of short functional elements encoded in their primary sequence and/or secondary structure (for example, stems, loops, etc) which serve distinct roles within the overall molecular function of the transcript. For instance, so-called scaffolding lncRNAs may contain regions responsible for binding distinct and otherwise non-interacting with each other proteins which are brought together by virtue of their binding to the same lncRNA molecule.

The lncRNA *ANRIL* which binds both the SUZ12 subunit of the PRC2 complex and the chromodomain of CBX7 in the PRC1 complex provides an example of a scaffolding lncRNA that works to repress the *CDKN2A*/*CDKN2B* tumor suppressor locus via trimethylation of H3K27 (Yap et al., 2010). Similarly, *Kcnq1ot1* binds to both PRC2 and G9a, which impart H3K4 trimethylation and H3K9 methylation, respectively (Pandey et al., 2008), while *HOTAIR* recruits the EZH2 methyltransferase of PRC2 via its 5′ structured domain and lysine-specific demethylase 1 (LSD1), which catalyzes demethylation of H3K4, via its 3′ end (Tsai et al., 2010).

3.2.4 Syntenic transcription and sequence conservation allow the identification of orthologous intergenic lncRNAs in mouse and human

Although *Dali* is evolving under evolutionary constraint (Ponjavic et al., 2009), it is unknown if this locus is expressed in other mammals, in addition to mouse. I therefore

identified its human ortholog based on two assumptions: that the relative genomic positions of *Pou3f3* and the hypothetical human *DALI* are preserved, and that at least portions of its sequence are conserved between mouse and human and transcribed in both species. Guided by local sequence alignment, a cluster of sequence tags expressed in human neuronal samples (optic nerve (*EB386236*), hypothalamus (*AV728332*), hippocampus (*DA337013*, *DB309936*)) was identified 53 kb downstream of the 3′end of human *POU3F3* (Figure 3.5 A). Using circular RACE (cRACE; Chapter 2.2.2.3) in human fetal brain RNA sample, I showed that these tags represent a continuous 3706 bp long transcript (Figure 3.5 A), which I termed human *DALI* (*hDALI*). *hDALI* contains two regions of >75% sequence identity with mouse *Dali* (approximately 560 bp in total), as well as a region of >50% of sequence identity of approximately 350 bp (Figure 3.5 B).

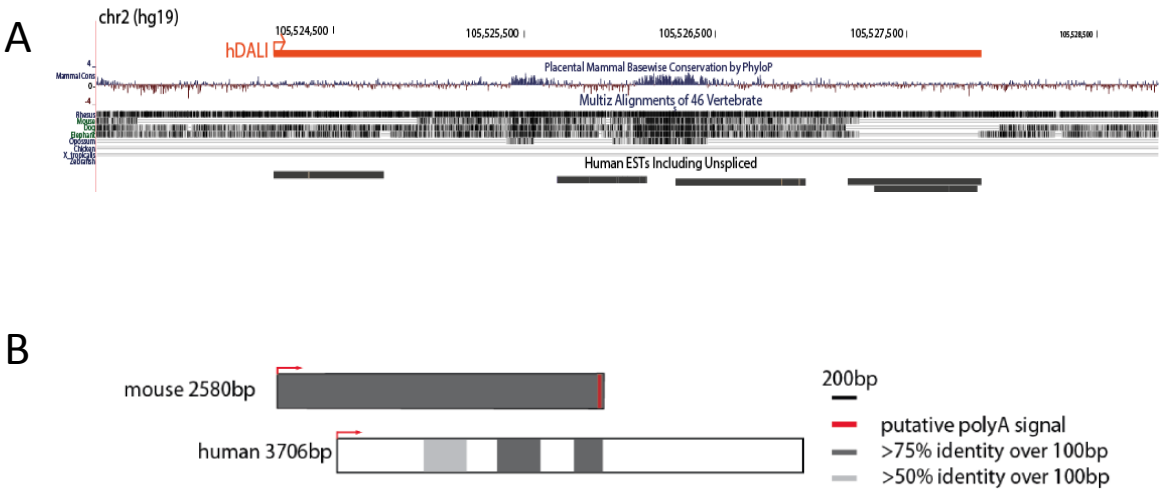


Figure 3.5 Human *DALI* ortholog identification and sequence conservation. (A) Full-length human *DALI* ortholog was identified using cRACE in the human embryonic brain sample. (B) Local sequence alignment (BLAST) tools were used to compare mouse and human *Dali*/*DALI* transcripts. Two regions of high and one region of medium sequence conservation were identified.

3.2.5 Intergenic lncRNAs whose loci reside in genomic proximity to developmental transcription factors exhibit variable sub-cellular localization

To further characterise these transcripts Rapid Amplification of cDNA Ends (RACE) was performed in order to confirm that candidate lncRNAs do not overlap with protein coding gene sequence and are thus truly intergenic (Figure 3.6; see Chapter 2.2.2.2).

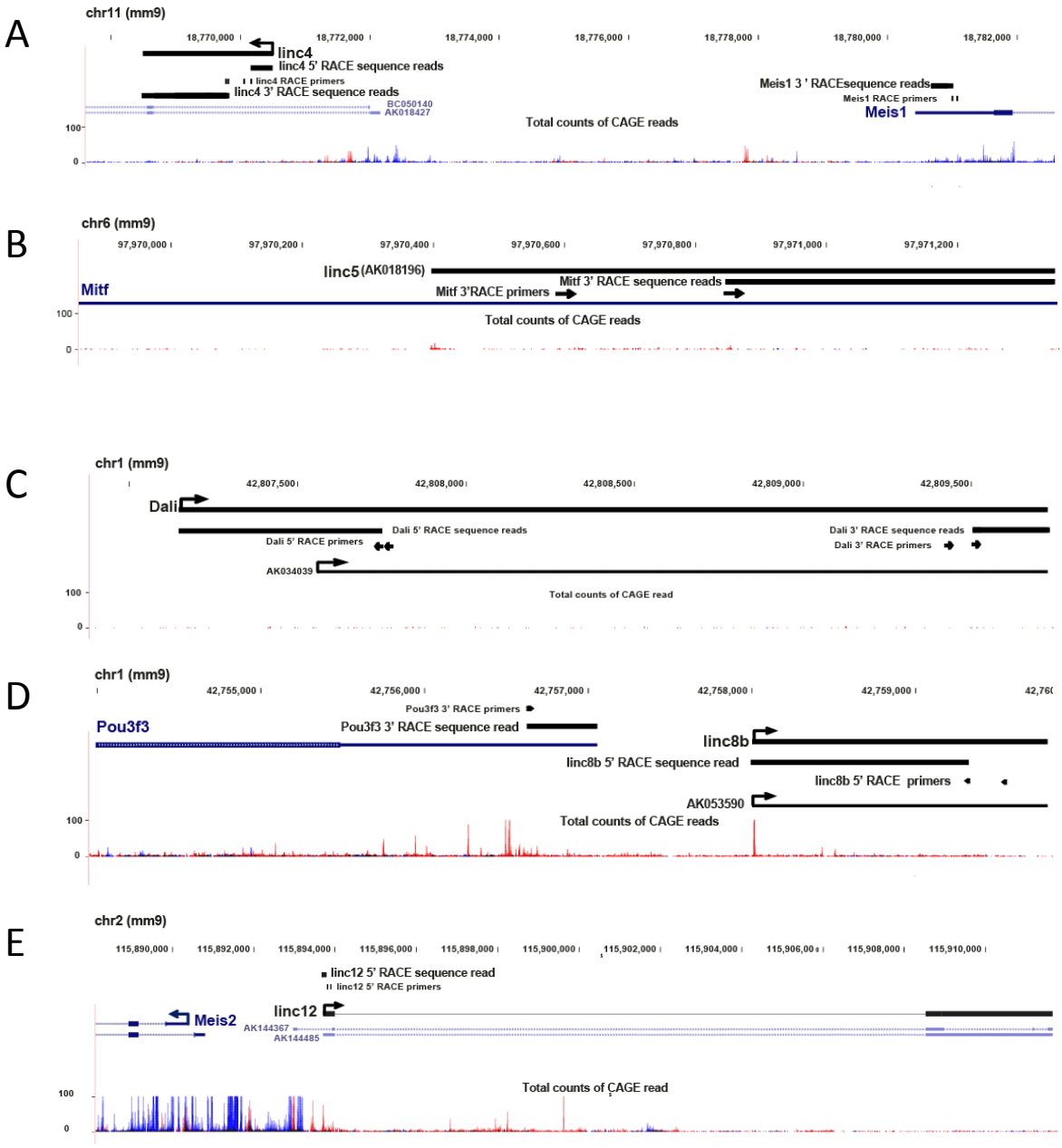


Figure 3.6 The intergenic status of the candidate lncRNAs confirmed by RACE in N2A cells. Boundaries of lncRNA and/or neighbouring protein-coding genes were mapped using Rapid Amplification of cDNA Ends (RACE) in mouse neuroblastoma N2A cells. *Linc4, Dali, AK053590,* and *linc12* were found to be fully intergenic. *Linc5* appears to be a part of the *Mitf* 3′ UTR expressed separately and capped.

The lncRNAs residing in the other three loci were confirmed to be fully intergenic (Figure 3.6 A, C, D, E). *Linc5* was found to lie fully within the extended 3′ UTR of *Mitf* (Figure 3.6 B). Therefore, this locus was excluded from further analyses. 3′ RACE in mouse neuroblastoma N2A cells extended *AK042766* (*linc4*) by 980 bp, merging together two previously known ESTs in the region (*AK042766* and *AK060420*; Figure 3.6 C). I will refer to the resulting transcript as *linc4*, as before. 5′ RACE also added approximately 500 bp to the 5′ end of *AK034039*. I termed the resulting transcript *Dali* (Figure 3.6 C). The intergenic status of another lncRNA from the *Pou3f3* locus, *linc8b* (*AK053590*), was also confirmed using 3′ RACE of *Pou3f3* and 5′ RACE of the lncRNA itself (Figure 3.6 D). *Meis2* and *linc12* are transcribed in opposite orientations in a head-to-head orientation separated by approximately 2.5 kb of intergenic sequence. 5′RACE confirmed the boundaries of previously annotated *Meis2* and *linc12* (*AK144485*) transcripts (Figure 3.6 E).

As sub-cellular localization of a lncRNA transcript may indicate a range of potential molecular functions, I decided to assay the sub-cellular distribution of the lncRNAs from the *Meis1*, *Meis2*, and *Pou3f3* loci in mouse neuroblastoma N2A cells. I biochemically fractionated N2A cells into cytoplasmic, nucleoplasmic, salt-extracted, and chromatin fractions and measured the levels of candidate intergenic lncRNAs in each fraction and in the total N2A RNA input sample using RT-qPCR (Chapter 2.2.1.6). The mono-exonic

Gapdh mRNA was included as a control to assess the fidelity of the fractionation. *Linc4, Dali*, and *linc12* were found to be strongly enriched in the chromatin fraction. *Linc8b* was more equally distributed across all fractions, with twice as much of it being present in the chromatin fraction compared to other fractions, while *linc8u* (another intergenic lncRNA from the *Pou3f3* locus included in this analysis due to its genomic position rather than being selected from the Ponjavic et al. dataset; later termed *lincBrn1-a* by (Sauvageau et al., 2013)) accumulated predominantly in the cytoplasm. *Gapdh* mRNA, as expected, exhibited cytoplasmic localization (Figure 3.7). Therefore, intergenic lncRNAs display variable sub-cellular localization profiles, suggesting that these transcripts are likely to perform different molecular roles in the cell.

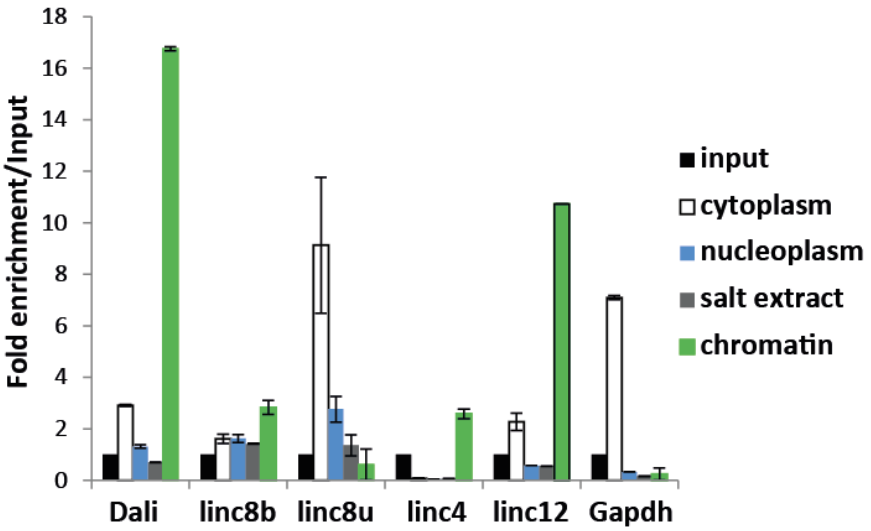


Figure 3.7 The subcellular localization of the candidate lncRNAs in N2A cells. *Dali, linc4* and *linc12* are chromatin-associated transcripts. *LincBrn1-a* (*linc8u*) is a nucleoplasmic transcript. *AK053590* (*linc8b*) is distributed across all fractions, with a small enrichment in the chromatin fraction. The relative amounts of *Dali* (G) and a control mRNA (*Gapdh*) (H) in the indicated fractions were measured by qRT-PCR. Mean values +/- s.e. of three independent experiments.

3.3 Discussion

Evolutionary conservation is widely accepted as a hallmark of functional significance of genes and other genomic features. Fruitful as it has been for protein-coding genes, simply searching for homology at the nucleotide sequence level appears to be insufficient to fully capture the selection process acting on non-coding portions of the genome, and lncRNAs in particular. For lncRNAs, conservation of sequence, structure, function, and syntenic expression can all be important, together or in combination.

Taken together, DNA sequences of mammalian lncRNAs are conserved less well than those of protein coding-genes, but more so than intronic or random intergenic sequences (Guttman et al., 2009; Ponjavic et al., 2009). However, a subset of these transcripts appear to evolve under purifying selection and display evolutionary conservation at the sequence level (Guttman et al., 2009; Marques and Ponting, 2009; Ponjavic et al., 2007). When compared to more distant vertebrate species, such as zebrafish, only a few mammalian lncRNAs displaying significant sequence similarity (usually restricted to relatively short regions) were observed (Ulitsky et al., 2011). Hence, lncRNAs evolve rapidly and either lack orthologs in other species, or the identification of their orthologs is non-trivial.

To identify potentially functional lncRNAs, additional aspects of conservation may be useful. For example, it has been noted that many lncRNAs originate from syntenic loci (regions whose order along the chromosome is preserved between different species) in different, sometimes highly evolutionarily divergent species (Ulitsky et al., 2011; Young et al., 2012). It has been proposed that despite their dissimilar sequences such transcripts may be related to each other functionally (Ulitsky et al., 2011). For example, if the act of transcription at equivalent positions affects the gene expression locally in different species,

the transcripts produced can be considered functional orthologs. In practice, positional information has been used for the identification of potentially functional lncRNAs in several published studies (Haerty and Ponting, 2013; Ulitsky et al., 2011; Young et al., 2012).

Furthermore, secondary structures of mammalian lncRNAs have been found to evolve under purifying selection (Smith et al., 2013). This finding is particularly significant, because lncRNAs with RNA-dependent functions are expected to adopt complex secondary and tertiary structures important for their activity. Of note, conservation of structure can be connected to conservation of sequence, but not necessarily always so. Structurally, the preservation of base-pairing is more important than nucleotide sequence and, therefore, reciprocal mutations can accumulate as long as the same secondary structure is preserved. Supporting this hypothesis, Smith et al. observed that almost 90% of the four million secondary structure elements they identified as conserved in mammals occur outside of known regions of conserved sequence (Smith et al., 2013).

Lastly, the function of a lncRNA can be conserved between species. For example, a lncRNA may interact with the same protein complex in human and mouse. Likewise, tissue specificities of lncRNA expression have been found to be often conserved (Necsulea et al., 2014). Functional conservation is expected to be linked to a conservation of either the sequence or the structure. Unfortunately, only very few lncRNAs have been comprehensively functionally studied in multiple species, which at present precludes the estimation of how often a function of a lncRNA is preserved in evolution.

A restricted expression profile indicative of active regulation has also often been cited as evidence for functionality of some lncRNAs. Several studies in different species found that lncRNAs tend to be expressed in a narrower range of tissues and at more restrictive developmental windows compared to protein-coding genes (Derrien et al., 2012; Mercer et al., 2008; Pauli et al., 2012).

In this Chapter, I examined two aspects of lncRNA conservation discussed above, i.e. sequence conservation patterns and syntenic transcription. I observed that, unlike for the neighbouring protein-coding genes, the sequence of my candidate lncRNAs was not equally conserved throughout their loci, but instead was confined to relatively short stretches interspersed with much more divergent regions. Such a pattern may indicate selection acting to preserve essential components of the overall fold of these lncRNAs, rather than the whole sequence. At the same time, promoter regions in all cases displayed the highest conservation, indicating potential conservation of expression. In agreement with this hypothesis, comparison of mouse candidate loci to their syntenic regions in human revealed that in all cases, transcripts of similar length and structure can be observed at orthologous positions.

I also investigated if the selected lncRNAs display evidence for another hallmarks of functionality, i.e. active expression regulation. For this, I examined their temporal and spatial expression by studying the levels of the candidate transcripts during RA-induced neuronal differentiation of ES cells and in adult mouse organs. 7 out of 8 lncRNAs were actively regulated and either significantly up- or down-regulated during the time-course of neuronal differentiation of ES cells. Of note, for all of the loci examined, expression patterns of lncRNAs tightly correlated with those of their neighbouring transcription factor

genes. Likewise, lncRNAs and TF were generally expressed in the same small set of organs in adult mice, with their relative expression levels well correlated with each other. In 2 out of 4 loci whose organ profiling was performed, both lncRNAs and TFs were expressed only in two tissues. For the other 2 loci expressed in more than two tissues, lncRNA expression was more tissue-restricted, as might have been expected.

Taken together, the work presented in this chapter supports the notion that evolutionary conservation might be distinct between coding and noncoding genes. Different aspects of conservation can be observed in different combination for individual transcripts. Therefore, useful as they might be for the prioritisation of likely functional lncRNAs, global interpretations of lncRNA conservation patterns and conclusions derived from them need to be validated by detailed and thorough functional studies using genetic and molecular biological tools in different species, if possible. For that reason, the work described in subsequent chapters focuses on a single-exon evolutionarily conserved lncRNA, *Dali*. *Dali* is transcribed from a wider locus of the transcription factor gene *Pou3f3*. Like *Pou3f3*, *Dali* is up-regulated at the precursor stage during neuronal differentiation and is expressed in brain and kidney.

Chapter 4 *Dali* is a novel chromatin-associated lncRNA involved in neural differentiation by controlling gene expression locally and distally

4.1 Introduction

A growing number of nuclear localised lncRNAs are known to regulate gene transcription and chromatin organisation (reviewed in (Vance and Ponting, 2014a, b)). A subset of these transcripts appear to act near to their site of synthesis to regulate the expression of genes locally on the same chromosome in an allele specific manner (*cis*-acting), while others have been demonstrated to act *in trans* to regulate gene expression across multiple chromosomes, and on either allele. *Cis*-acting lncRNA regulatory mechanisms have been described in detail for a number of enhancer associated nuclear lncRNAs, as well as lncRNAs involved in the processes of genomic imprinting and X chromosome inactivation (Melo et al., 2013; Monnier et al., 2013; Mousavi et al., 2013; Santoro et al., 2013; Tian et al., 2010; Vallot et al., 2013).

However, not all of the many thousand mammalian intergenic lncRNAs that have been identified will be functional. Single-exon lncRNAs, in particular, can be artefacts arising from genomic DNA contaminating sequencing libraries. Also, transcripts that are expressed at average levels lower than 1 copy per cell are less likely to confer function. At the same time, highly and broadly expressed, and *bona fide* mono-exonic intergenic lncRNAs, such as *Neat1* and *Malat1*/*Neat2*, appear not to have essential roles either, because their knockout mouse models are viable and fertile (Eissmann et al., 2012; Zhang et al., 2012). Transcript sequences and levels are thus not reliable predictors of mechanism. Instead, the significant temporal and spatial co-expression of genomically

adjacent intergenic lncRNA and transcription factor genes might suggest that such lncRNAs commonly modulate transcriptional programmes that are initiated by these transcription factors (Ponjavic et al., 2009). Indeed, several intergenic lncRNAs have well-documented *cis*-acting regulatory roles (Berghoff et al., 2013; Wang et al., 2011b; Zhang et al., 2012).

Spatiotemporal co-expression of intergenic lncRNA and transcription factor genes is most pronounced during the development of the mouse CNS (Ponjavic et al., 2009). To investigate the mechanistic basis of this physical linkage, among the evolutionarily conserved intergenic lncRNAs differentially expressed during RA-induced neuronal differentiation of ES cells (described in Chapter 3) I chose to study a 3.5-kb, CNS-expressed, monoexonic, intergenic lncRNA *Dali*. This was initially termed *linc8* (Chapter 3) but was eventually named *D*NMT1-*A*ssociated *L*ong *I*ntergenic RNA, or *Dali* (see Chapter 5).

As it was decided to focus on neural differentiation as a model system to study potential roles of intergenic lncRNAs, I excluded *linc4* that was expressed at much higher levels in bladder and lung compared to brain (where it was expressed at a level similar to a number of other organs). Also, the single-exon *linc4* lay fully within another intergenic lncRNA, *AK018427* (Figure 3.6 A), which would have complicated quantitative expression studies. *Linc12* was maximally expressed in brain and lung, and was excluded because its expression was the broadest of the remaining candidate lncRNAs. Furthermore, *linc12* fully overlapped another transcript, *AK144367*, which also would have complicated the interpretation of experimental results. *Linc8b* (*AK053590*) that had similar temporal and spatial expression profile with *Dali* and *Pou3f3* was excluded due to its close proximity to

the 3′ end of *Pou3f3* and its overlap with several other mono-exonic non-coding transcripts.

Dali, on the other hand, is expressed in a sense orientation from a locus 50 kb away from its nearest protein-coding neighbour *Pou3f3* (also known as *Brn1* or *Oct8*), which encodes a class III POU family transcription factor. *Dali* does not overlap any other transcripts or previously identified regulatory elements in the region (Figure 4.1 A). As I shall show, *Dali* is a single-exon 2.58 kb long chromatin-associated transcript conserved across therian mammals. It is expressed in the brain and kidney only, the two organs where *Pou3f3* is known to perform important roles during development (Nakai et al., 2003). Also, the onset of *Dali* and *Pou3f3* expression during RA-induced ES cell differentiation occurred later compared to the other loci, which might be indicative of a specific role at more advanced stages of neuronal differentiation.

The extended *Pou3f3* locus is characterised by near pervasive transcription in neuronal lineages (Ramos et al., 2013). Sauvageau et al. recently generated mouse knockout models for two of the intergenic lncRNAs residing in this locus, *linc-Brn1a* (*linc8u*) and *linc-Brn1b* (Figure 4.1). Genomic deletion of the *linc-Brn1b* locus resulted in significant (~50%) down-regulation of the upstream *Pou3f3* gene, and *linc-Brn1b*$^{-/-}$ mice exhibited abnormalities of cortical lamination and barrel cortex organization (Sauvageau et al., 2013). These abnormalities may derive from loss of the *linc-Brn1b* RNA transcript or from deletion of DNA functional elements (Figure 4.1 A) (Bassett et al., 2014a; Bassett et al., 2014b). Therefore, understanding the properties of *Dali* will extend the current knowledge about the *Pou3f3* locus important for brain development.

Pou3f3 is a single exon gene whose protein binds to DNA in a sequence-specific manner. *Pou3f3* contributes to both neuronal and kidney development by regulating the proliferation and differentiation of progenitor cells (Nakai et al., 2003). Mouse mutants with homozygous loss of *Pou3f3* die of renal failure within 36 hours *post partum* (Nakai et al., 2003), with severe defects of the hippocampus and forebrain among others (McEvilly et al., 2002). In the developing neocortex, *Pou3f3* is expressed in late neuronal precursors and in migrating neurons and, together with its closely related paralog *Pou3f2*, is required in ventricular zone progenitors for deep-to-upper layer fate transition, sustained neurogenesis and cell migration (Dominguez et al., 2013).

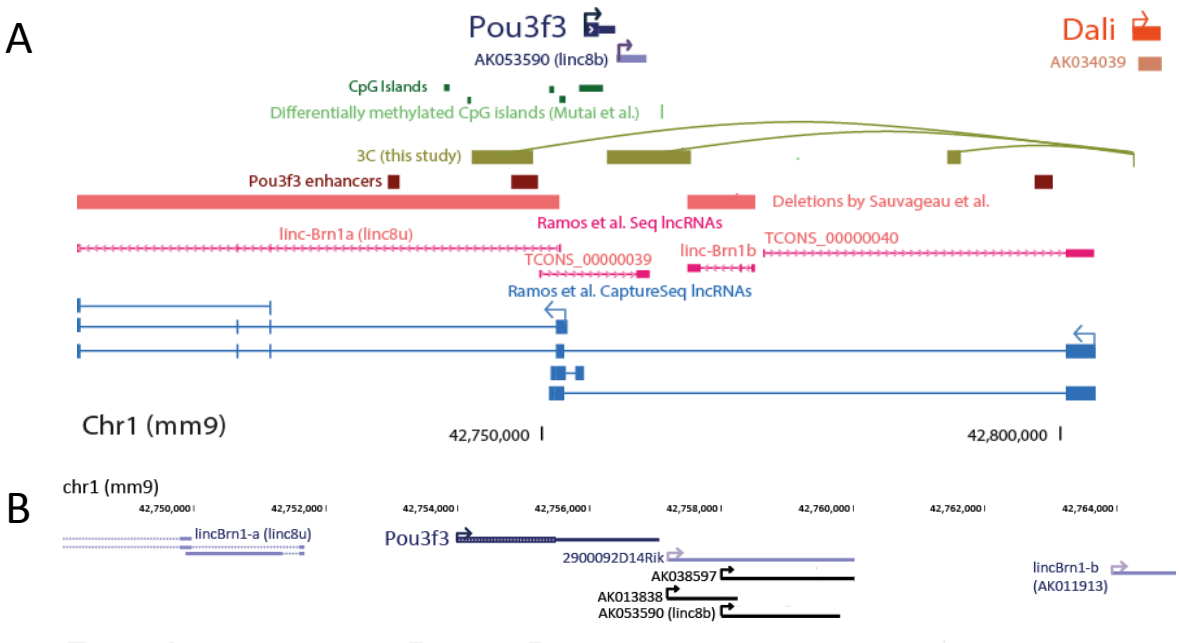


Figure 4.1 Schematic illustration of the mouse *Pou3f3* genomic region. (A) Coding and non-coding transcripts, enhancer elements from Vista Enhancer Browser (Visel et al., 2007), CpG islands, and published genomic deletions (Sauvageau et al., 2013) are shown. (B) Magnified view of the immediate genomic vicinity of the *Pou3f3* gene showing non-coding transcripts on either side of *Pou3f3*.

In this chapter, I show that *Dali* is required for the correct differentiation of neural cells in culture. Furthermore, it functions in an RNA-dependent manner to modulate the expression of its neighbouring *Pou3f3* gene, as well as other genes *in trans* involved in the neuronal differentiation programme. I also demonstrate that *Dali* is likely to regulate some of these distal targets in a *Pou3f3*-independent manner. The data presented therefore demonstrate that a lowly expressed single-exon lncRNA transcript can regulate multiple genes situated away from its site of synthesis.

4.2 Results

4.2.1 *Dali* is co-expressed with *Pou3f3* in the developing mouse brain

Dali is co-expressed with its neighbouring protein-coding gene *Pou3f3* both spatially in adult mouse organs (brain and kidneys; Figures 3.1 A) and temporally during RA-induced neuronal differentiation of mouse ES cells (Figures 3.2 C). In order to assess if these two transcripts are also spatially and temporally co-expressed in the developing mouse embryonic brain, lateral cortical plate (LCP), dorsal cortical plate (DCP), lateral ventricular zone (LVZ) and dorsal ventricular zone (DVZ) were dissected out at five sequential stages of embryonic development: E9.0, E10.5 (lateral pre-plate (LPP), dorsal pre-plate (DPP)), E13.5 (LCP, DCP, LVZ, DVZ), E15.5 (LCP, DCP, LVZ, DVZ), E18.5 (LCP, DCP, LVZ, DVZ) (Chapter 2.2.6.1). Levels of *Dali* and *Pou3f3* were measured in these samples by RT-qPCR, and expressed relative to the endogenous *Gapdh* mRNA control (which was

expressed at similar levels across all stages and regions). The results are summarized in Figure 4.2. Consistent with the previous RA-induced ES cell differentiation results (Figures 3.1 A), both transcripts are first up-regulated with the appearance of first cortical neurons (E10.5), and increase in expression further as the ratio between neurons and progenitors grows (Figure 4.2) (Sahara et al., 2012). As differentiation progresses, *Dali* is up-regulated approximately 6-7 fold, and *Pou3f3* 4.5-7.5 fold between E10.5 and E18.5 in cortical plate (Figure 4.2), similar to 2-3 fold observed for both factors during the ES differentiation time-course (Figure 3.1 A). Therefore, *Dali* and *Pou3f3* are co-expressed spatially and temporarily *in vivo* in the developing mouse embryonic brain. Of note, in differentiated ES cells, *Pou3f3* levels are approximately 10-fold higher relative to *Dali* levels than in the developing mouse brain (Figures 3.1 A and 4.2). This difference may be caused by real differences in expression levels of both transcripts in individual cells *in vitro* and *in vivo*, or alternatively, may indicate that the *in vivo* cellular composition is heterogeneous and includes cells expressing *Pou3f3* but not *Dali*.

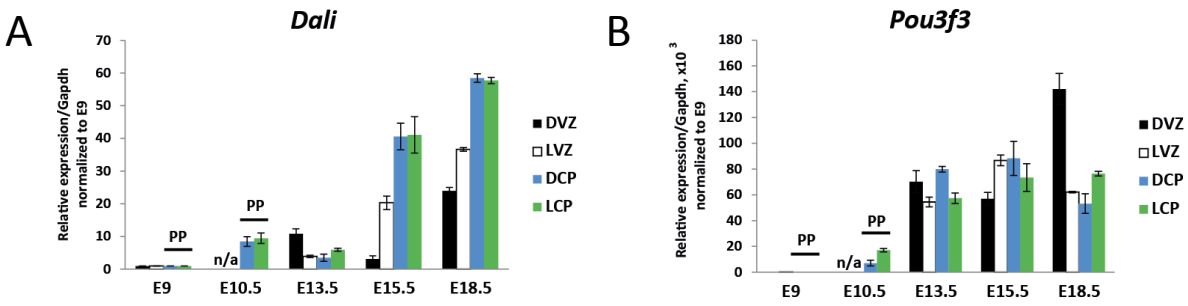


Figure 4.2 *Dali* and *Pou3f3* are co-expressed temporally and spatially in the developing mouse brain. *Dali* (A) and *Pou3f3* (B) expression profiles in developing embryonic brain. DVZ: Dorsal ventricular zone; LVZ: Lateral ventricular zone; DCP: Dorsal cortical plate; LCP: Lateral cortical plate; PP: pre-plate. The levels of *Dali*, *Pou3f3* were measured by qRT-PCR. Results are normalised to *Gapdh* and presented relative to expression in E9.0 sample (set arbitrarily to 1). Mean +/- s.e., n = 3 (technical replicates).

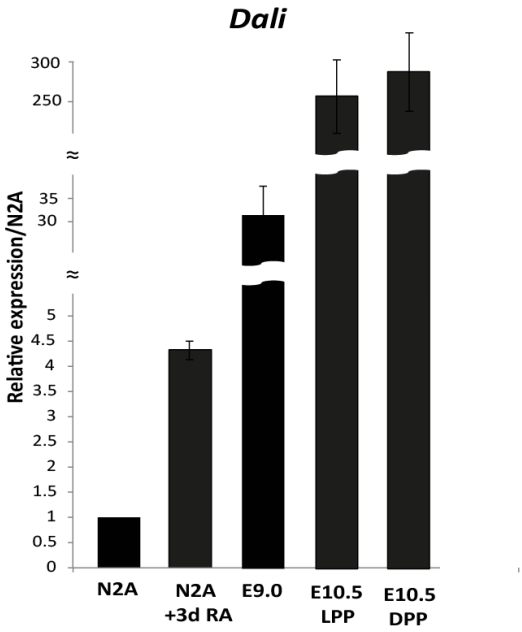


Figure 4.3 *Dali* is expressed higher *in vivo* than in N2A cells. Even at the earliest stages when it is detected (E9.0 and E10.5), *Dali* levels are higher in the developing mouse brain than in both proliferating and differentiated N2A cells. *Dali* levels were measured by RT-qPCR, normalised against *Gapdh* control mRNA, and presented relative to expression in proliferating N2A cells (set arbitrarily to 1). Mean values +/- s.e., n=3 (biological replicates for N2A cells, technical replicates for developing brain samples).

In order to study *Dali* function, a suitable cell culture system was required. Mouse neuroblastoma N2A cells are a fast-growing easy-to-transfect cell line that is frequently used as a neuronal progenitor-like cell type (Olmsted et al., 1970). Among other aspects of neuronal cell biology, these cells are often used to study terminal stages of neuronal differentiation in culture. When treated with retinoic acid (RA) under reduced serum conditions, N2A cells undergo terminal differentiation by exiting cell cycle and developing very long, usually singular, neurite outgrowths (Chapter 2.2.1.3). N2A cells express both *Dali* and *Pou3f3*. In these cells *Dali* is expressed at a level at least two orders of magnitude lower than when first detected in neuronal progenitor-dominated areas of the developing brain (E10.5) (Figure 4.3). Using RT-qPCR and a calibration curve approach (Chapter 2.2.3.14), I estimated N2A cells to express *Dali* at a population-average copy number of 2 (Figure 4.4). However, when N2A cells are treated with retinoic acid (RA) for 72 h, *Dali* is up-regulated approximately 4.5-fold, similar to the approximately 6-7 fold up-regulation observed in embryonic cortical plate (both dorsal and lateral) between days E10.5 to E18.5

(Figure 4.5). Therefore, despite *Dali* expression level differences in N2A cells and the *in vivo* system, N2A cells and their RA-induced differentiation represent an appropriate model system in which to study *Dali* function and were subsequently used in all experiments described below, unless otherwise stated.

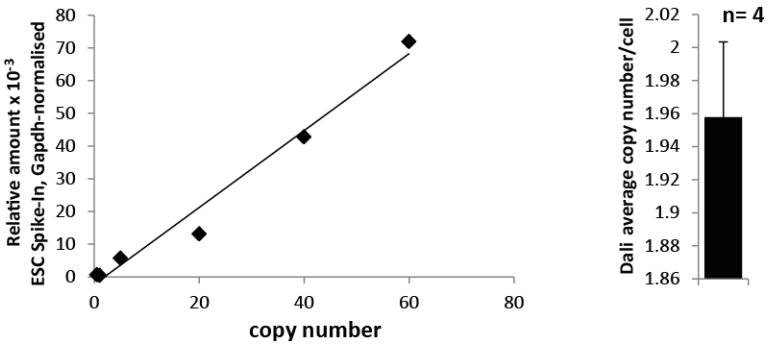


Figure 4.4 Population-average copy number of *Dali* in N2A cells. *Dali* is expressed at an estimated 2.0 ± 0.4 copies per cell in N2A cells. A standard curve of known *Dali* copy number was constructed by spiking *in vitro* transcribed *Dali* into RNA from ES cells, which do not express it (left). Mean *Dali* expression per cell was calculated from four independent RT-qPCR experiments using RNA extracted from a defined number of cells. This value was used to estimate *Dali* copy number form the standard curve. Mean copy number per cell ± s.e. is shown (right).

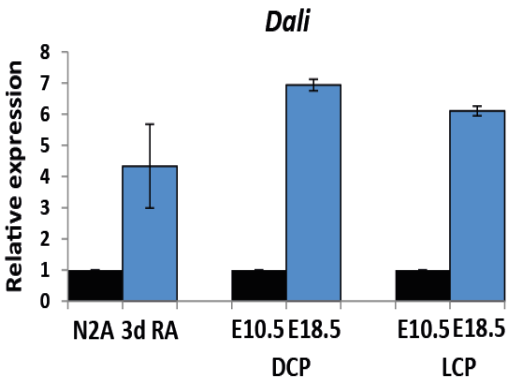


Figure 4.5 *Dali* is up-regulated during neuronal differentiation both in culture and in the developing mouse brain. Despite differences in expression levels in culture and *in vivo*, *Dali* levels increase by similar amounts during RA-induced differentiation of N2A cells and in the developing cortical plate between E10.5 when first neurons emerge and E18.5 when they dominate developing cortical plate by cell number. *Dali* levels were measured by RT-qPCR, normalised against *Gapdh* control mRNA, and presented relative to expression in proliferating N2A cells (set arbitrarily to 1). Mean values +/- s.e., n=3 (biological replicates for N2A cells, technical replicates for developing brain samples). DCP = Dorsal cortical plate; LCP = Lateral cortical plate.

4.2.2 *Dali* is co-expressed with *Pou3f3* in the adult mouse brain

In adult mouse (P56), *Dali* and *Pou3f3* are expressed in all three regions of adult neurogenesis, i.e. sub-ventricular zone (SVZ), olfactory bulb (OB), and dentate gyrus (DG) (Chapter 2.2.6.1) (Ming and Song, 2011)). The highest expression was observed in the sub-ventricular zone, followed by dentate gyrus and olfactory bulb (Figure 4.6). However, the expression of both genes in all three regions was not dramatically different, with *Dali* being approximately 2.8-fold, and *Pou3f3* 3.9-fold higher expressed in SVZ compared to OB. Therefore, *Dali* and *Pou3f3* are also co-expressed in the adult brain regions examined.

Overall the two transcripts are co-expressed in all contexts studied here, i.e. across the time-course of RA-induced neuronal differentiation of mouse ES cells, in adult mouse organs, including brain, as well as in developing mouse brain. These observations support the hypothesis that *Dali* and *Pou3f3* may have regulatory or functional connections with each other.

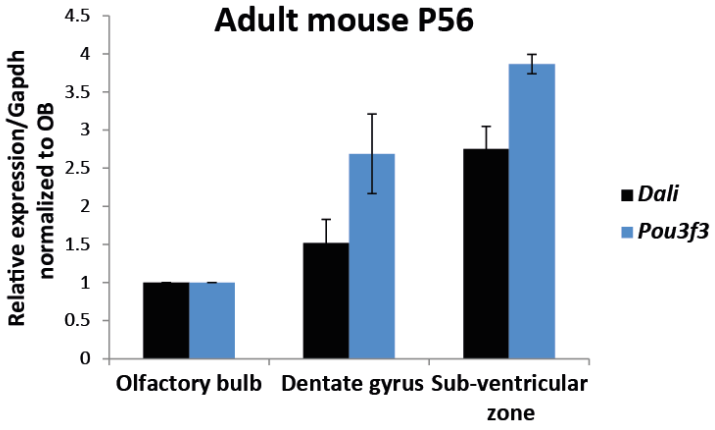


Figure 4.6 *Dali* and *Pou3f3* are co-expressed temporally and spatially in the adult mouse brain. *Dali* and *Pou3f3* levels were measured in the adult mouse brain (P56) by RT-qPCR in three regions of adult neurogenesis, i.e. olfactory bulb, dentate gyrus and sub-ventricular zone. The indicated regions were dissected from a single male adult mouse brain per sample. Measurements were normalised using *Gapdh* and presented relative to expression in the olfactory bulb samples (set arbitrarily to 1). Mean values +/- s.e., n = 2.

4.2.3 *Dali* regulates neural differentiation of N2A cells

Although evolutionary conservation (of sequence, structure and/or syntenic transcription) might be indicative of functionality, a mutant phenotype derived from loss- and/or gain-of-function models provides definitive proof that a particular locus is functional. *Dali* was previously shown to be expressed only in the CNS and kidney (Figure 3.2 C). These are the two organs where its neighbouring *Pou3f3* gene plays an important role during the development of progenitor cells (Nakai et al., 2003). Also, similar to *Pou3f3*, *Dali*'s expression is tightly regulated during neuronal differentiation of ES cells in culture, with its onset corresponding to the stage when cycling progenitors exit cell cycle and undergo terminal differentiation (Figure 3.1 A). Therefore, I decided to test if depletion of *Dali* gives rise to an abnormal neural differentiation phenotype using RA-induced differentiation of N2A cells. I generated three independent stable *Dali* knockdown N2A cell lines carrying different shRNA constructs targeting *Dali* (Chapter 2.2.1.5). These cell lines showed reduction of *Dali* levels by approximately 50-70% (Figure 4.7 A). I performed differentiation experiments with all three independent stable *Dali* knockdown cell lines and a stable non-targeting control line (Chapter 2.2.1.3). 72 h after post-differentiation, significantly fewer differentiated cells of all three *Dali* knockdown lines developed neurites (Figures 4.7 B; Chapter 2.2.1.8). Those that did, displayed shorter neurites, often with multiple short outgrowths emanating from the same cell, compared to one or two long neurites developed by differentiated control cells (Figures 4.7 C). Prior to differentiation, control and knockdown cell lines did not exhibit noticeable phenotypic differences (Figures 4.7 D).

These results suggest that *Dali* is important for normal differentiation of N2A cells in culture, and for neuronal projection development in particular. Therefore, *Dali* exhibits two hallmarks of functionality: it is evolving under evolutionary constraint, and its depletion gives rise to a cellular phenotype of abnormal terminal neural differentiation.

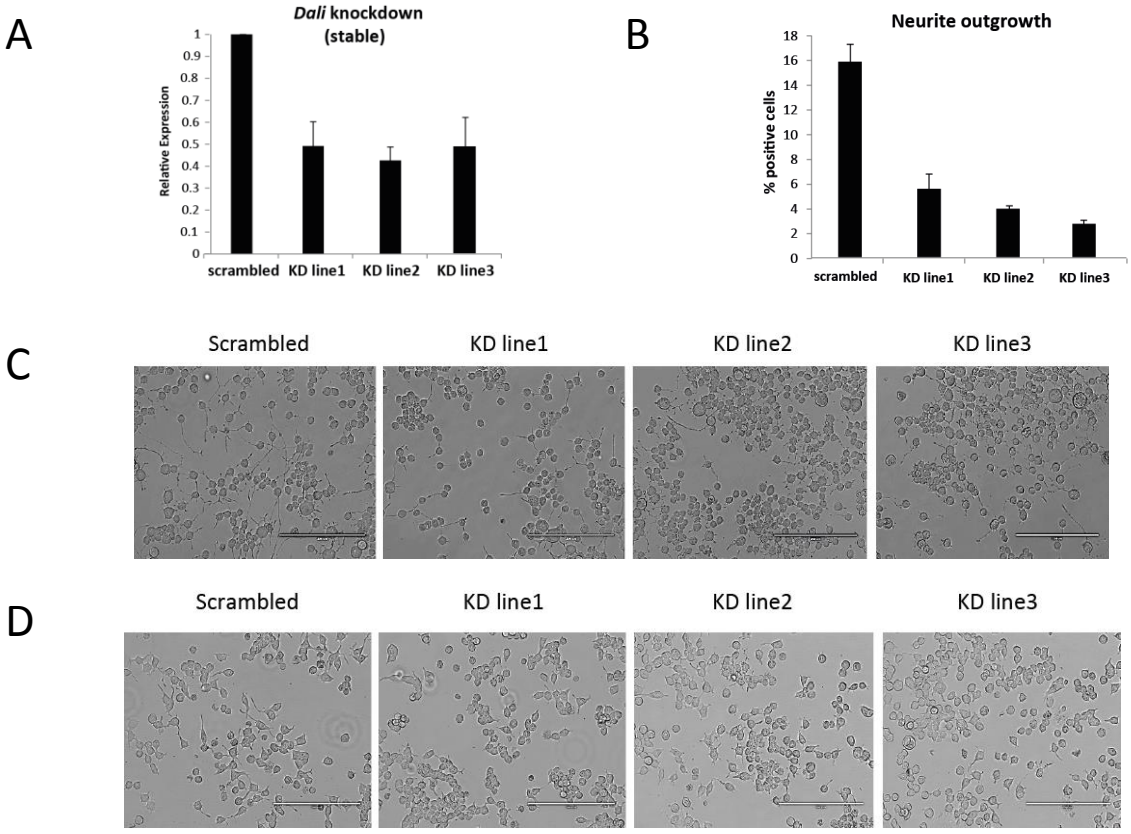


Figure 4.7 *Dali* plays a role in neurite outgrowth development during RA-induced neuronal differentiation of N2A cells. (A) *Dali* knockdown (KD) or control cell lines were generated by stable co-transfection of either shRNA expression plasmids targeting *Dali* or a non-targeting (scrambled) control and a selection vector. qRT-PCR analysis shows that three clonal cell line expres reduced levels of *Dali*. Mean values +/- s.e. (B) *Dali* KD cells display reduced neurite outgrowth when differentiated with RA. Cells were imaged using bright field microscopy. Cells with one or more neurites of length greater than twice the cell body diameter were scored as positive. Average values +/- s.e., n=3. A total of 500-600 cells were counted in each case in at least 3 non-overlapping fields of view. (C) Representative images of control and stable *Dali* KD cells differentiated with RA for 72 h. (D) Control and stable *Dali* KD proliferating cells show no apparent morphological differences prior to differentiation. Scale bar = 200 μm on both C and D.

4.2.4 *Dali* locus in both mouse and human is marked with chromatin marks characteristic of a weak (or poised) enhancer

Figure 4.8 *Dali* loci in both mouse and human bear enhancer chromatin marks. (A) Promoter region of *Dali* in mouse is associated with DNase I hypersensitivity sites in tissues expressing *Dali* (kidney and brain) but not in ES cells where the *Dali* locus is silent. (B) *DALI* locus in human is annotated as a poised (or weak) enhancer by the ENCODE project (Hoffman et al., 2013).

As a first step to functionally characterize the *Dali* locus, I examined its chromatin marks in both mouse and human using the ENCODE data. For the mouse locus, I considered tissues where *Dali* was expressed (namely, brain and kidney), as well as embryonic stem cells where it was silent. I observed a DNase I hypersensitive site in the *Dali* promoter region in both brain (cerebrum and whole adult brain at 8 weeks) and kidney samples. No such signal was observed in the ES cell data, which is consistent with the *Dali* expression pattern (Figure 4.8 A). Also, in both brain (cortex) and kidney samples, I observed a high H3K4me1 to H3K4me3 ratio, as well as low levels of H3K27ac. No such marks were observed in the ES cell data (Figure 4.8 A). These observations suggest that the mouse *Dali* locus may have properties of a weak (or poised) enhancer in both brain and kidney. Consistent with this hypothesis, the ENCODE project annotated the human *Dali* locus as a weak (or poised) enhancer (Figure 4.8 B) (Hoffman et al., 2013).

Therefore, based on the chromatin status of the mouse and human loci, both orthologous *Dali* transcripts may have similar properties in the two species. The chromatin state of the *DALI* loci suggests that this lncRNA may originate from a previously uncharacterised enhancer. This is consistent with the highly positively correlated expression profiles of *Dali* and *Pou3f3* (as well as several non-coding transcripts nearby) in mouse described earlier (Chapter 3.2.1, 4.2.1, 4.2.2).

4.2.5 *Dali* transcript is quickly turning over in mouse neuroblastoma N2A cells

Previous studies reported that nuclear lncRNAs are often less stable than both protein-coding and cytoplasmic non-coding transcripts (Clark et al., 2012). To start gaining

insights into a mechanistic function of *Dali* and its post-transcriptional regulation, I used nascent RNA pulse-chase labelling (see Chapter 2.2.3.12) to estimate the stability of this nuclear-localized transcript in proliferating N2A cells. Although more precise calculations are precluded by the low expression level of *Dali* in N2A cells, as well as the kinetics of the uptake and clearance of the label from metabolising mammalian cells, the turnover time of the *Dali* transcript was estimated to be approximately 30 min (Figure 4.9). This is significantly faster than the estimated 90 min turnover time of the mRNA for the TATA-box binding protein (*Tbp*), which is considered to be relatively unstable, which is characteristic of many genes involved in transcriptional regulation (Sharova et al., 2009). Therefore, *Dali* is an unstable lowly expressed transcript in N2A cells. However, a growing body of literature indicates that low stability is not suggestive of the transcript being non-functional. On the contrary, similar to the instability of transcription factor protein-coding genes (Friedel et al., 2009), rapid turnover of nuclear-localised lncRNAs, *Dali* included, is consistent with the idea that they may act as regulatory molecules and that this may contribute to their rapid and dynamic regulation (Berretta et al., 2008; Clark et al., 2012; Seidl et al., 2006; van Dijk et al., 2011). For example, *Neat1*, a lncRNA involved in nuclear paraspeckle organization and function (Mao et al., 2011), was found to be one of the least stable transcripts in N2A cells. The high turnover rate was proposed to contribute to the dynamic regulation of paraspeckles (Clark et al., 2012). Also, as nuclear lncRNA are not translated, and may be subject to less extensive co-/post-transcriptional processing, they can function almost immediately after transcription. Consequently, nuclear lncRNAs may not require long half-lives (Dinger et al., 2009). Of note, the half-life of several highly unstable nuclear lncRNAs involved in cell proliferation was found to be positively regulated by RA treatment (Tani et al., 2012). Although not investigated here, it would be

interesting to determine whether the stability of *Dali* is also increased upon RA treatment and may contribute to its observed up-regulation.

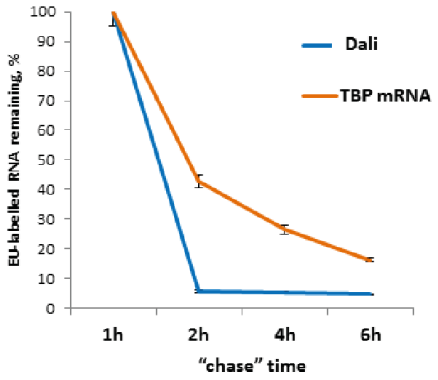


Figure 4.9 *Dali* is an unstable transcript in proliferating N2A cells. Stability of the *Dali* transcript was estimated using a nascent transcript labelling pulse-chase experiment. N2A cells were cultured in the presence of the uridine analog ethyl-uridine for 24 h and returned to the label-free medium. Samples were taken at the indicated intervals and the labelled fractions were isolated. The levels of *Dali* and control transcripts were measured by RT-qPCR, and normalised against the amounts present in the total unfractionated RNA samples prepared in parallel. Fractions of labelled *Dali* and unstable control *Tbp* mRNA are presented relative to the 1h sample (arbitrarily set to 100%). The turnover time for *Dali* is estimated to be approximately ½ h. Mean +/- s.e., n=3.

4.2.6 *Dali* regulates gene expression *in cis*

Previous studies showed frequent genomic co-occurrence and temporal and spatial co-expression of lncRNAs with neighbouring transcription factor genes (Ponjavic et al., 2009). Also, numerous *cis*-acting regulatory roles for a growing number of long non-coding transcripts have been described (for example, *HOTTIP* (Wang et al., 2011b), *Evf2* (Berghoff et al., 2013), *Malat1* (Zhang et al., 2012) and others). Based on these observations, Ponjavic et al., who assembled the set of evolutionarily constrained CNS expressed putatively intergenic lncRNAs from which my initial candidate loci were selected, hypothesized that intergenic lncRNA loci found next to and co-expressed with brain-specific transcription factor genes might be involved in regulating these genes

(Ponjavic et al., 2009). Also, I previously showed that *Dali* is a nuclear localized chromatin-associated transcript (Figure 3.7) and, therefore, may play a role in transcriptional regulation. Consequently, I decided to investigate whether *Dali* knockdown had an effect on the expression of its neighbouring *Pou3f3* gene. I used three different shRNA constructs transiently transfected into N2A cells to reduce *Dali* levels by an average of 60-70% (Figure 4.10 A). After 72 h, I measured *Pou3f3* transcript levels using RT-qPCR and found them to be reduced by approximately 40% (Figure 4.10 A). Three independent stable *Dali* knockdown clones also showed approximately 15-40% reduction of *Pou3f3* levels (Figure 4.10 B). Importantly, the down-regulation of *Pou3f3* using shRNAs did not have a significant effect on *Dali* levels (Figure 4.10 C), suggesting that the regulation of *Pou3f3* transcript levels by *Dali* is unidirectional.

Although the molecular mechanism of nuclear shRNA-mediated knockdown remains unknown and for some nuclear localised transcripts this technique can be inefficient, numerous publications have previously described the knockdown of strictly nuclear lncRNAs (Berezhna et al., 2006; Chen and Carmichael, 2009; Clemson et al., 2009; Ohrt and Schwille, 2008; Vance et al., 2014; Zhang et al., 2013). In fact, both Ago2 and Dicer have been found to localise to the nucleus in human (Ahlenstiel et al., 2012; White et al., 2014). Although *Dali* levels can be reduced by 50-70% using shRNAs, the efficiency of knockdown could potentially be improved using different approaches, such as complementary antisense oligodeoxynucleotides that recruit nuclear RNase H to degrade the RNA (Bennett and Swayze, 2010).

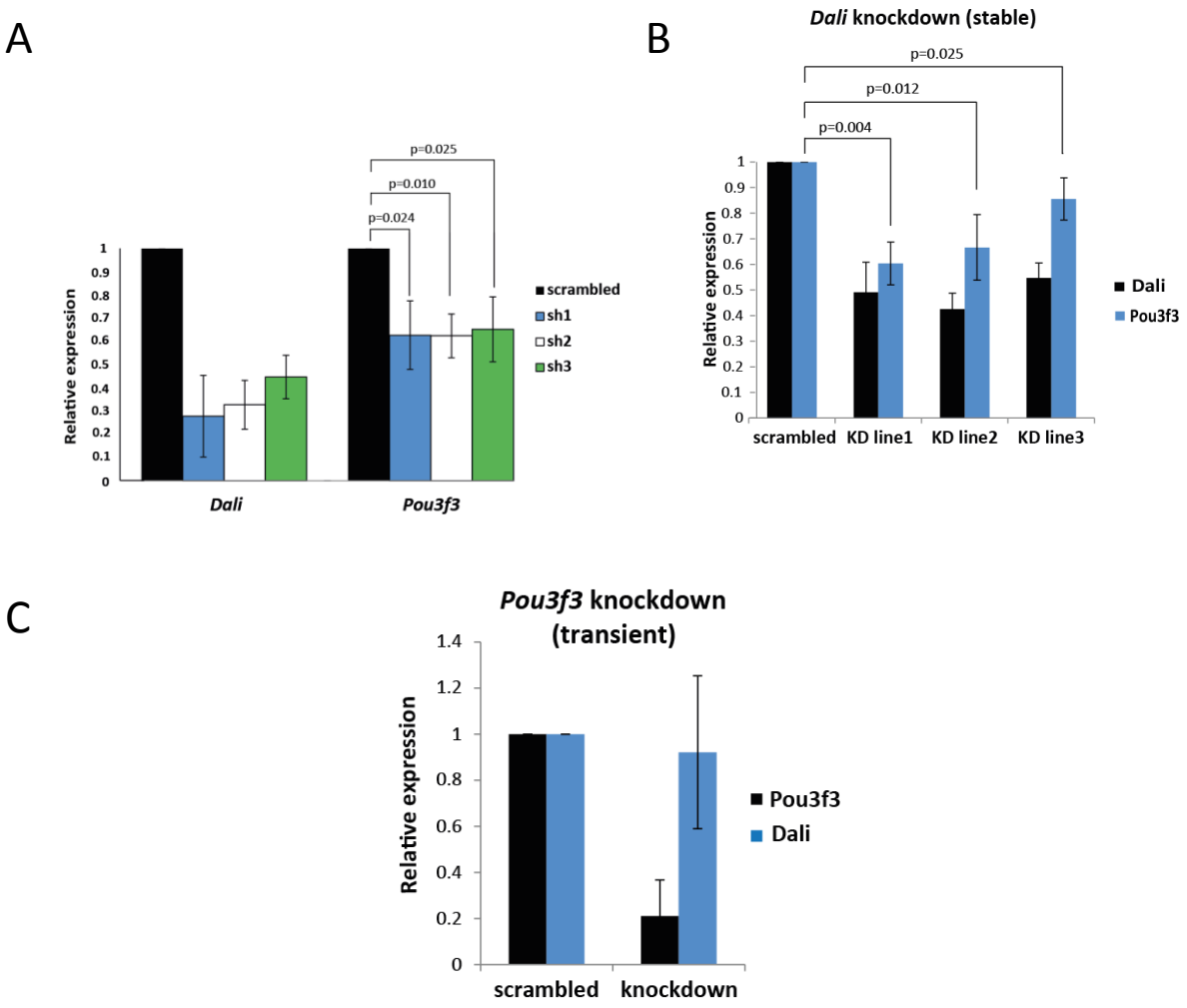


Figure 4.10 *Dali* regulates *Pou3f3* and non-coding transcripts in its genomic vicinity. (A) *Dali* knockdown leads to a decrease in *Pou3f3* expression. N2A cells were transfected with three independent shRNA expression vectors targeting different regions along the *Dali* transcript. The levels of *Dali* and its nearest protein-coding neighbour *Pou3f3* were measured by qRT-PCR three days later. Samples were normalised using *Gapdh* and the results are shown relative to a non-targeting control (set at 1). Mean values +/- s.e., n = 3, one tailed *t*-Test, unequal variance. (B) The relative levels of *Pou3f3* were measured in stable *Dali* knockdown and control cells by qRT-PCR, normalised using *Gapdh* and presented relative to expression in control cells (set arbitrarily to 1). Mean values +/- s.e, n = 3, one tailed *t*-Test, unequal variance. (C) *Dali* levels unchanged upon *Pou3f3* knockdown. N2A cells were transfected with either a non-targeting control (scrambled) or a *Pou3f3*-targeting shRNA expression vector (knockdown). *Pou3f3* and *Dali* levels were measured by RT-qPCR after 72 h. Mean values +/- s.e, n = 3.

I also checked whether *Dali* knockdown has an effect on the expression of *AK053590*, an intergenic lncRNA encoded just downstream of *Pou3f3*. Reduction of *Dali* levels by 30-60% resulted in the down-regulation of *AK053590* by approximately 40-45% (Figure 4.11 A). However, in this case, the regulatory effect was bidirectional. The reduction of *AK053590* levels by approximately 60% resulted in the down-regulation of *Dali* by approximately 70% (Figure 4.11 B). *Pou3f3* levels were also reduced by approximately 80% of the original, while *linc8u* (termed *linc-Brn1a* by (Sauvageau et al., 2013)) levels were up-regulated by 90% following *AK053590* depletion (Figure 4.11 B). Of note, Sauvageau et al. also investigated the effect of another intergenic lncRNA, *linc-Brn1b*, encoded just downstream of *AK053590*, on the expression of *Pou3f3* and *linc8u* (*linc-Brn1a*). These authors found that deletion of *linc-Brn1b* leads to a down-regulation of *Pou3f3* by 2-fold and a significant up-regulation of *linc-Brn1a* which shares a bi-directional promoter with *Pou3f3* (Sauvageau et al., 2013). Therefore, the closely situated lncRNAs *AK053590* and *Brn-1b* loci have similar regulatory effects on the neighbouring transcripts. Depletion of any of these transcripts results in down-regulation of *Pou3f3* and transcript(s) situated to the 3′ of it, and concomitant up-regulation of *linc-Brn1a* situated to the 5′ of *Pou3f3*.

In summary, *Dali* positively regulates both neighbouring protein-coding (*Pou3f3*) and non-coding genes (*AK053590*). Overall, *Pou3f3* and transcripts in its extended locus appear to be regulated by a complex network of local interactions involving non-coding transcripts on both sides of this gene.

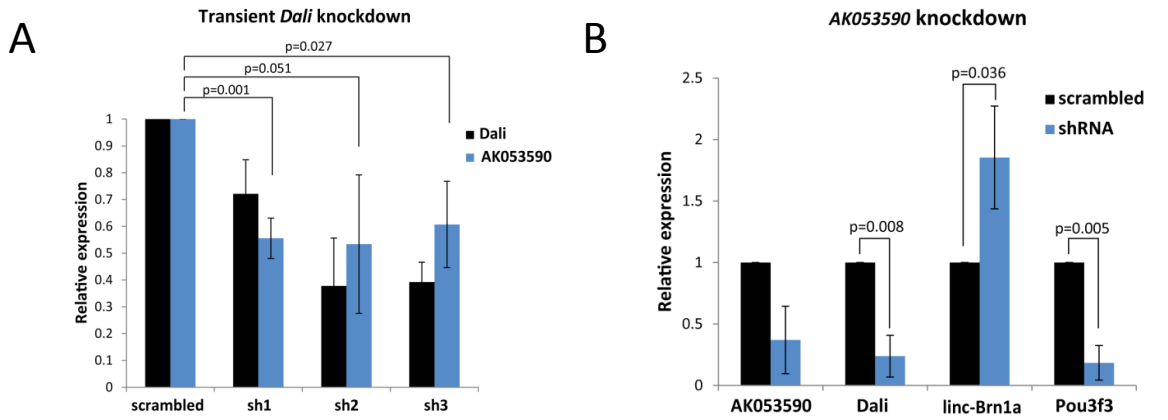


Figure 4.11 Non-coding transcripts in the *Pou3f3* locus form a network of regulatory interactions. (A) *Dali* regulates expression of lncRNA *AK053590*. N2A cells were transfected with either a non-targeting control (scrambled) or three independent *Dali* targeting shRNA expression vectors. *Dali* and *AK053590* levels were measured by RT-qPCR 72 h post- transfection. Mean values +/- s.e., n = 3, one tailed *t*-Test, unequal variance. (B) *AK053590* regulates expression of *Dali* and *Pou3f3* positively and *linc-Brn1a* negatively. N2A cells were transfected with either a non-targeting control (scrambled) or a *AK03590* targeting shRNA expression vector (shRNA). *AK053590*, *Dali*, *linc-Brn1a* and *Pou3f3* levels were measured by RT-qPCR 72 h post-transfection. Mean values +/- s.e., n = 3, one tailed *t*-Test, unequal variance.

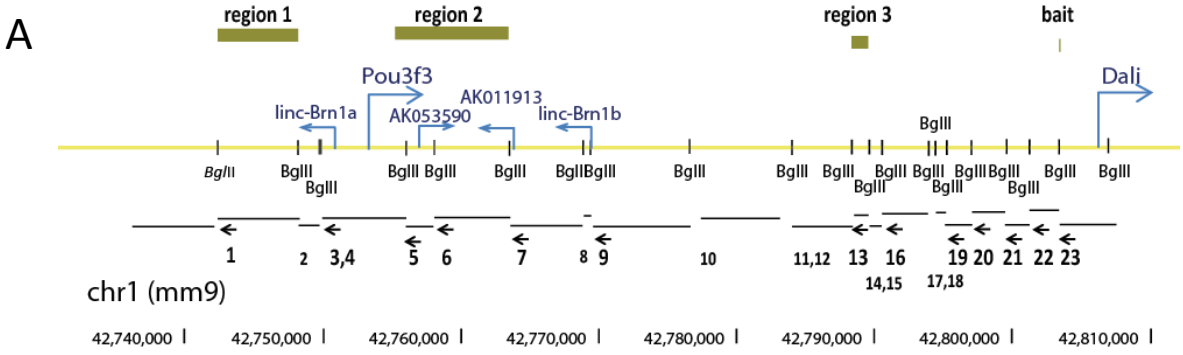
4.2.7 The *Pou3f3* locus exists in a folded nuclear conformation independent of its expression

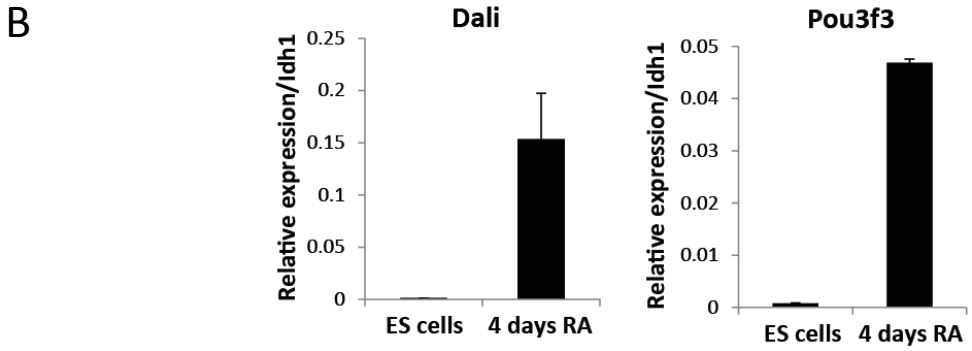
In order to gain structural insights into the genomic region spanning *Pou3f3* and *Dali*, I used Chromatin Confirmation Capture (3C; Chapter 2.2.2.4) to determine the local genomic contacts made by *Dali* across the *Pou3f3* locus (Figure 4.12 A) in ES cell-derived neuronal precursor cells, where both *Pou3f3* and *Dali* are expressed (Figure 4.12 B). Fragment containing the *Dali* promoter and 5′ sequence contacted three regions (Figure 4.12 C, D; Figure 4.1 A): 1) an enhancer element lying upstream of *Pou3f3* within the *linc-*

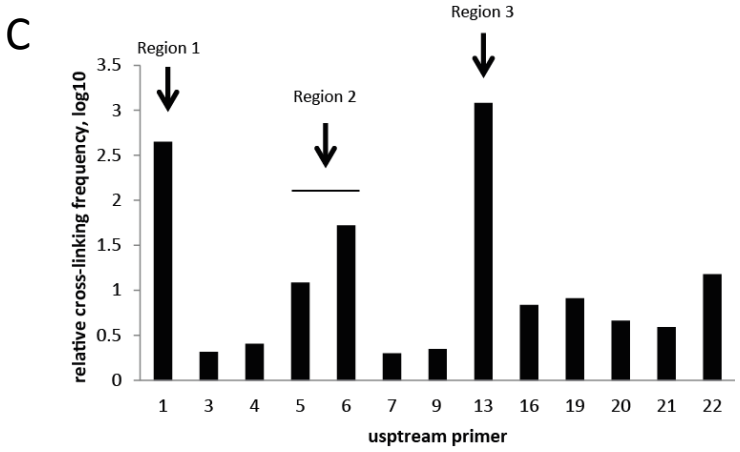
Brn1a locus, 2) a region overlapping the 3′ UTR of *Pou3f3* and full-length *AK53590* (which are both regulated by *Dali*), as well as parts of *TCONS_00000039* and *linc-Brn1b*, and including a differentially methylated region reported to be important in regulating *Pou3f3* expression (Mutai et al., 2009), and 3) a region lying within another non-coding locus, *TCONS_00000040* (Ramos et al., 2013).

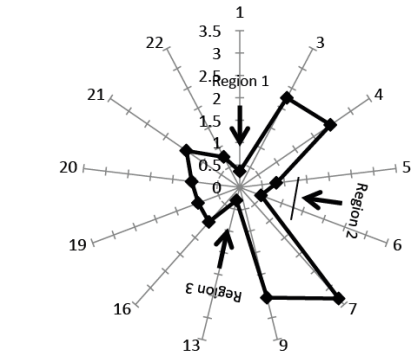
Neither *Dali* nor *Pou3f3* appears to play a role in initiating these DNA looping interactions because they were present also in E14 ES cells (Figure 4.12 E, F) where neither is expressed (Figure 4.12 B). Likewise, other lncRNAs in this region (*AK053590*, *linc-Brn1a* and *linc-Brn1b*) are unlikely to contribute to these interactions in an RNA-dependent manner, as these are silent in ES cells and up-regulated only upon neuronal differentiation ; (Figure 3.1 A) (Sauvageau et al., 2013). Nevertheless, the *Dali* locus appears to contribute to an extended structurally and transcriptionally complex region centred on the *Pou3f3* gene.

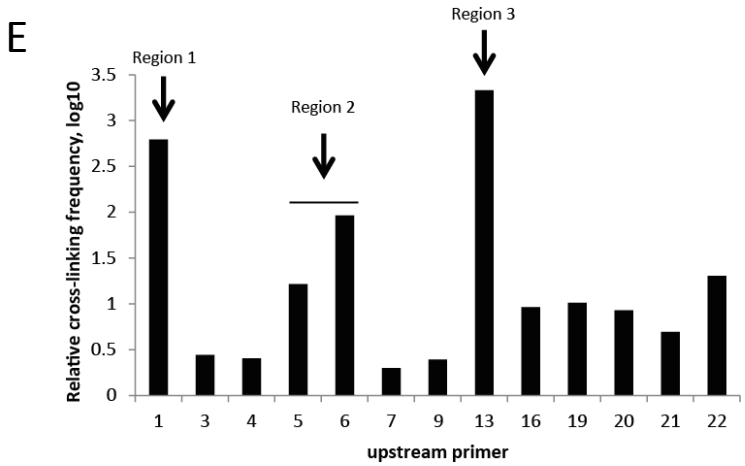
Figure 4.12 The *Pou3f3* locus occurs in a folded nuclear conformation both prior to and after the onset of the expression of its transcripts. (A) Schematic representation of the *Pou3f3* locus showing start sites of transcripts in the region and positions of recognition sites of the *Bgl*II restriction endonuclease (small vertical bars) used for the 3C experiment. Below, the restriction fragments generated by *BglII* digestion are annotated. Arrows indicate the position of the 3C qPCR primers. Numbers represent the corresponding primer (small numbers indicate primers not used in the final analysis due to technical reasons). Green bars indicate regions found to be in close proximity to the *Dali* transcription start site used as "bait". (B) Nuclear conformation of the *Pou3f3* locus was studied in ES cells where the locus is silent and ES cell derived neuronal precursors (4 days retinoic acid differentiation) where transcripts in the regions are expressed. (C and E) Quantification of genomic interactions between the *Dali* TSS and the genomic fragments indicated in (A) in ES cells (C) and ES cell derived neuronal precursors (E). The *y* axis shows relative cross-linking frequency, the *x* axis indicates the primer used in combination with primer 21. (D and F) Graphical representation of genomic distances between the *Dali* TSS, positioned at the centre of the radar, and the indicated *BglII* genomic fragments in ES cells (D) and ES cell derived neuronal precursors (F). Distance was calculated as 1/cross-linking frequency.

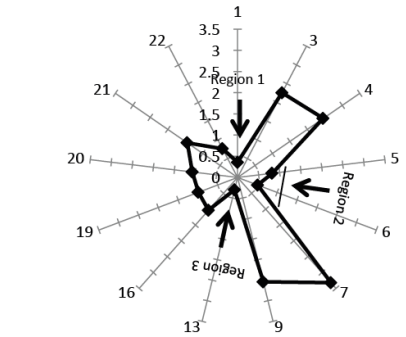

A
region 1
region 2
region 3
bait
Pou3f3
Dali
AK011913
AK053590
linc-Brn1a
linc-Brn1b
BglII
chr1 (mm9)
42,740,000
42,750,000
42,760,000
42,770,000
42,780,000
42,790,000
42,800,000
42,810,000


B
Dali
Pou3f3
Relative expression/Idh1
ES cells
4 days RA


C
Region 1
Region 2
Region 3
relative cross-linking frequency, log10
usptream primer


D
Region 1
Region 2
Region 3

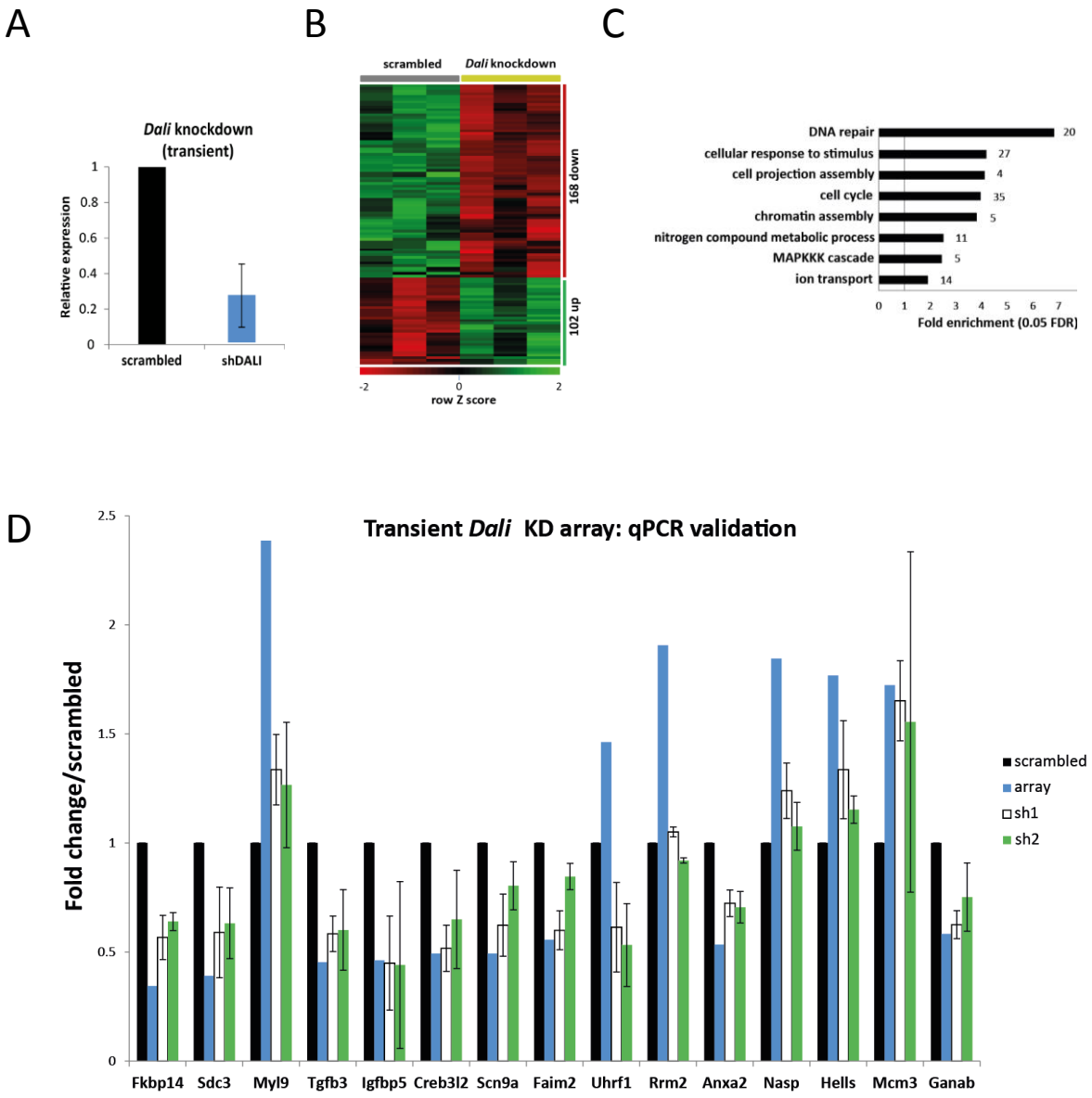

E
Region 1
Region 2
Region 3
Relative cross-linking frequency, log10
upstream primer


F
Region 1
Region 2
Region 3

4.2.8 *Dali* regulates genes *in trans*

To assess whether *Dali* can also regulate distally located genes, transcriptomes of N2A cells in which *Dali* levels were depleted using transient transfection of shRNAs were profiled using microarrays (Chapters 2.2.3.7, 2.2.7.1). Abundance of *Dali* was reduced by approximately 70% (Figure 4.13 A). This resulted in statistically significant changes in the expression of 270 genes (≥ 1.3-fold, False Discovery Rate [FDR] < 10%) compared to a non-targeting control (Appendix Table 1). 168 genes were down-regulated, 102 up-regulated (Figure 4.13 B). Gene Ontology (GO) category enrichment analysis (Chapter 2.2.7.1) showed significant enrichment of gene categories such as cell cycle, DNA repair, cellular response to stimulus, chromatin assembly, cell projection assembly, nitrogen compound metabolic process, ion transport and signalling (JNK and MAPKKK cascades, in particular) (Figure 4.13 C, Appendix Table 2). Cell cycle regulation is at the heart of neuronal differentiation which relies both on a balance between proliferation and self-renewal of progenitors and also tightly regulated cycle exit and terminal differentiation of committed cells (Politis et al., 2008). Hence, a perturbed cell cycle gene expression upon *Dali* knockdown is notable, given that *Dali* is sharply up-regulated during RA-induced neural differentiation of ES cells at a stage when differentiating cells exit the cell cycle.

Figure 4.13 *Dali* plays a role in regulating gene expression in neuronal cells. (A) N2A cells were transfected with either a non-targeting control (scrambled) or a *Dali* targeting shRNA expression vector (shDali) and *Dali* levels were measured by RT-qPCR 72 h post-transfection. Mean values +/- s.e., n = 3. (B) Transient *Dali* knockdown induces statistically significant changes in the expression of 270 genes in N2A cells (10% FDR) (Appendix Table 1). (C) Gene Ontology (GO) categories significantly enriched among *Dali* regulated genes (5% FDR, hypergeometric test, Benjamini and Hochberg correction; AppendixTable 2). (D) The results of the microarray experiment were validated in a series of independent transient knockdown experiments using two independent shRNA constructs. Mean values +/- s.e., n = 3.



Other genes differentially expressed in response to the depletion of *Dali* include MAPKKK cascade genes. MAP kinase signalling is crucial for orchestrating processes such as neuronal survival/death, differentiation, neuritogenesis, plasticity and memory (reviewed in (Miloso et al., 2008)). Among the MAPKKK cascade genes down-regulated upon *Dali* knockdown were *Cbs*, *Efna1*, *Cdc42ep5*, and *Tlr4*. *Cbs* gene is known to play a crucial role in the development and maintenance of the CNS (Enokido et al., 2005), and was recently

shown to stimulate proliferation and promote differentiation of neural stem cells to neurons and astroglia (Wang et al., 2013c). *Efna1* (*Ephrin-A1*) was found to facilitate commitment to neuronal fates (Aoki et al., 2004), while *Cdc42ep5* (*Borg3*) plays a role in the regulation of neuronal cytoskeleton and dendritic spine formation (Brand et al., 2012). *Tlr4* is expressed in neuronal progenitor cells where it was shown to be important for both cell proliferation and differentiation (Rolls et al., 2007). Also among those genes that were down-regulated most strongly after *Dali* knockdown are *Tgfb3* and *Igfbp5*, both of which were shown to promote neuronal survival and differentiation (Hu and Zuckerman, 2001; Tanno et al., 2005; Urano et al., 1999). *Tgfb3* was also shown to exert anti-proliferative effect during cortical development (Siegenthaler and Miller, 2005). The largest fold down-regulation was detected for *Fkbp14*, a regulator of Notch signalling pathway essential for neuronal differentiation (van de Hoef et al., 2013). Therefore, *Dali* may function as a pro-differentiation factor by regulating the balance between cell renewal and differentiation.

17 differentially expressed genes were selected for RT-qPCR validation of the microarray results using an additional independent shRNA expression construct targeting *Dali*. The results of the RT-qPCR experiments were consistent with those of the microarrays. Moreover, the fold changes measured by arrays were in good agreement with the degree of *Dali* knockdown achieved using different shRNA vectors (Figure 4.13 D). The results of the validation experiments indicate that the gene expression changes observed previously are specific to *Dali* knockdown and are unlikely to have resulted from off-target effects of the shRNA used. Therefore, *Dali* may function as a transcriptional regulator of cell cycle control genes in neural cells.

4.2.9 *Dali* and *Pou3f3* have both common and distinct transcriptional targets

As I previously showed that *Dali* can regulate *Pou3f3* expression, it was important to distinguish between gene expression changes resulting from the down-regulation of *Pou3f3* upon *Dali* knockdown from those that may be independent of *Pou3f3*; the letter would indicate that *Dali* can regulate not only proximal but also distal genes in the genome. To this end, I reduced *Pou3f3* transcript levels in N2A cells by 35% using transient transfection of a *Pou3f3* targeting shRNA vector (Figure 4.14 A) and used microarrays to measure gene expression levels genome-wide (Chapters 2.2.3.7 and 2.2.7.1). Statistically significant expression changes were observed in 1041 genes (≥ 1.3-fold, FDR < 10%; Figure 4.13 B, Appendix Table 3). 364 genes were down-regulated, and 677 up-regulated. As previously discussed, depletion of *Pou3f3* did not affect *Dali* expression (Figure 4.10 C).

GO analysis showed that genes differentially expressed after *Pou3f3* knockdown are enriched in categories related to cell division and cell cycle, as well as oxidation reduction (Figure 4.14 C; Appendix Table 4). The importance of cell cycle regulation in neuronal differentiation was discussed above. *Pou3f3* knockdown resulted in up-regulation of a number of positive regulators of cell cycle (*Cdc25b* (Peco et al., 2012), *Cdc25c*, (Chang et al., 2012; Chang et al., 2013), *Cdca2, Cdca7* (Gill et al., 2013), *Cdc7*, *Cdca3* (Zhao et al., 2013), and others). Activation of positive cell cycle regulators in post-mitotic neurons and re-entry into cells cycle are pro-apoptotic and cause neuronal death. *Cend1*, a key gene coordinating cell cycle exit and neuronal differentiation (Georgopoulou et al., 2006; Mamalaki et al., 1995), is down-regulated in *Pou3f3* knockdown cells.

Cellular differentiation is often accompanied by metabolic changes. Neuronal differentiation is associated with a switch from glucose oxidation reduction to fermentation (Fathi et al., 2014; Fornazari et al., 2011). Oxidative reduction genes are up-regulated upon *Pou3f3* depletion, which may indicate a less differentiated state adopted by these cells. In fact, the two processes, increased oxidation reduction and cell cycle, can be functionally coupled, as oxidative stress induced by increased oxidative reduction (category also significantly enriched in the *Pou3f3* knockdown dataset) has been shown to induce cell cycle re-entry and apoptosis of differentiated neurons (Folch et al., 2012).

The intersection between the *Dali* and *Pou3f3* knockdown sets of differentially expressed genes contained 84 genes which represented a large 6.2-fold enrichment over random expectation (p-value = 2.79×10^{-9}). The genes in the overlap constituted 31% of all genes differentially expressed upon *Dali* depletion (Figure 4.14 D). Approximately equal numbers of genes shared between the two datasets were down- (43 genes) and up-regulated (41 genes) in both experiments (Appendix Table 5). A good positive correlation between direction and fold-change of differentially expressed genes was observed between *Dali* and *Pou3f3* knockdown experiments (R= 0.74; Figure 4.14 E). I performed GO category enrichment analysis of the set of genes shared between *Dali* and *Pou3f3* knockdown datasets, as well as non-overlapping genes from both experiments. Cytoskeleton organization and ion transport genes were regulated by *Dali* in a largely *Pou3f3*-dependent manner, while chromatin assembly and MAPKKK signalling genes were regulated by *Dali* independently of *Pou3f3*. Cell cycle, DNA repair and cellular response to stimulus genes were regulated by *Dali* in both *Pou3f3*-dependent and independent manner. Dendrite development, axon guidance and sterol metabolism genes, among others, were regulated

by *Pou3f3*, but not *Dali* (Figure 4.14 F). Taken together, these results indicate that *Dali* and *Pou3f3* interact either genetically and/or molecularly to regulate a subset of common targets involved in neuronal differentiation. Also, both *Dali* and *Pou3f3* have targets regulated independently of each other, with the consequence that *Dali* can regulate genes not only *in cis*, but also *in trans*.

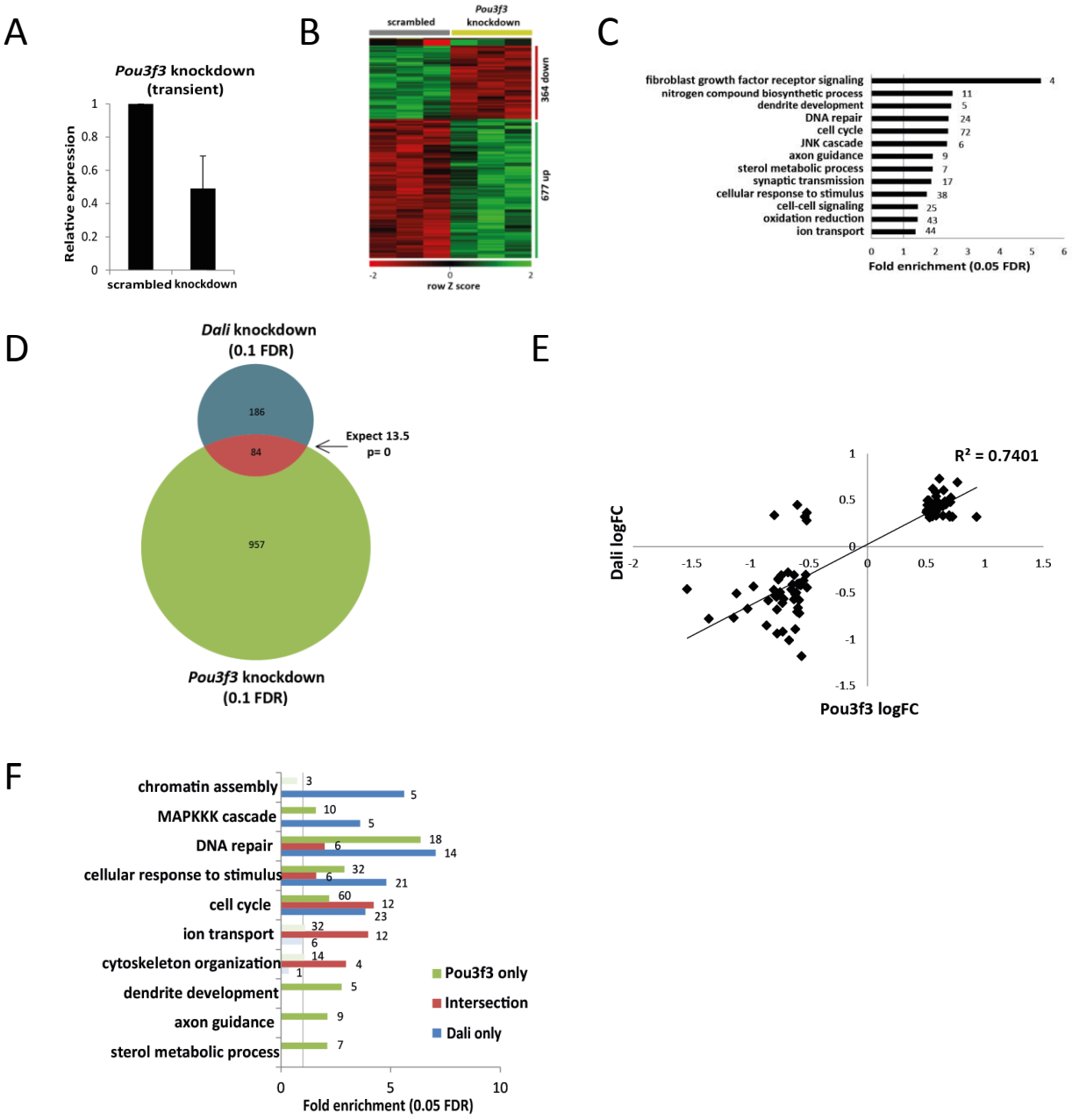


Figure 4.14 *Dali* regulates transcription in both *Pou3f3*-dependent and -independent manners. (A) N2A cells were transfected with either a non-targeting control (scrambled) or a *Pou3f3*-targeting shRNA expression vector (knockdown). *Pou3f3* and *Dali* levels were measured by qRT-PCR 72 h post-transfection. (B) *Pou3f3* knockdown resulted in statistically significant changes in the expression of 1041 genes in N2A cells ((10% FDR, Appendix Table 3). (C) GO-analysis of genes differentially expressed upon *Pou3f3* knockdown (5% FDR, hypergeometric test, Benjamini and Hochberg correction; Appendix Table 4). (D) Intersection of *Pou3f3* and *Dali* targets shows a significant (Fisher's exact test) overlap approximately 6.2 times as large as expected by chance alone. (E) Target genes common between *Dali* and *Pou3f3* show correlated expression, with the nearly all being either positively or negatively regulated by both factors (R=0.74; Appendix Table 5). (F) Enrichments of GO categories of *Pou3f3*-dependent or -independent *Dali* targets (5% FDR, hypergeometric test, Benjamini and Hochberg correction).

4.2.10 Induction of the endogenous *Dali* transcript in mouse ES cells regulates *Pou3f3* locally and *E2f2* distally

As depletion of *Dali* resulted in differential expression of genes situated both in the genomic vicinity of the *Dali* locus as well as elsewhere in the genome, I next tested whether *de novo* expressed *Dali* transcript can function as a transcriptional regulator of these genes. To achieve this, I induced *Dali* expression from its endogenous locus in E14 mouse ES cells, which do not express *Dali* or *Pou3f3* to detectable levels, using transient transfection of an artificial Transcription Activator-Like effector (TALE) transcription factor designed to bind to a region just upstream of the transcriptional start site of *Dali* (see Chapter 2.2.5.1). A TALE constitutes a fusion between a sequence-specific DNA-binding domain and a VP64 activation domain. The DNA binding domain of a TALE is composed of a series of 34-amino acid repeat modules arranged so as to recognize a given DNA sequence (Figure 4.15 A). An expression vector carrying the *Dali*-targeting TALE-TF and

an empty control vector were transfected into *Dali* non-expressing E14 mouse ES cells. After 72 h, these cells exhibited approximately 14-fold higher expression of the *Dali* endogenous locus than control cells (Figure 4.15 B) and a 3-fold increase in *Pou3f3* expression (Figure 4.15 C). *Dali* expression is thus sufficient to induce the expression of its genomically neighbouring *Pou3f3* gene. I next investigated the expression levels of *E2f2*, a gene that was found to be negatively regulated by *Dali* using shRNA mediated knockdown (Appendix Table 1). As expected, *Dali* up-regulation reduced *E2f2* transcript levels by approximately 40% (Figure 4.15 C). Taken together, these results indicate that *Dali* can regulate local and distal target genes both when its expression is induced from its endogenous locus as well as when its transcript is depleted using shRNAs.

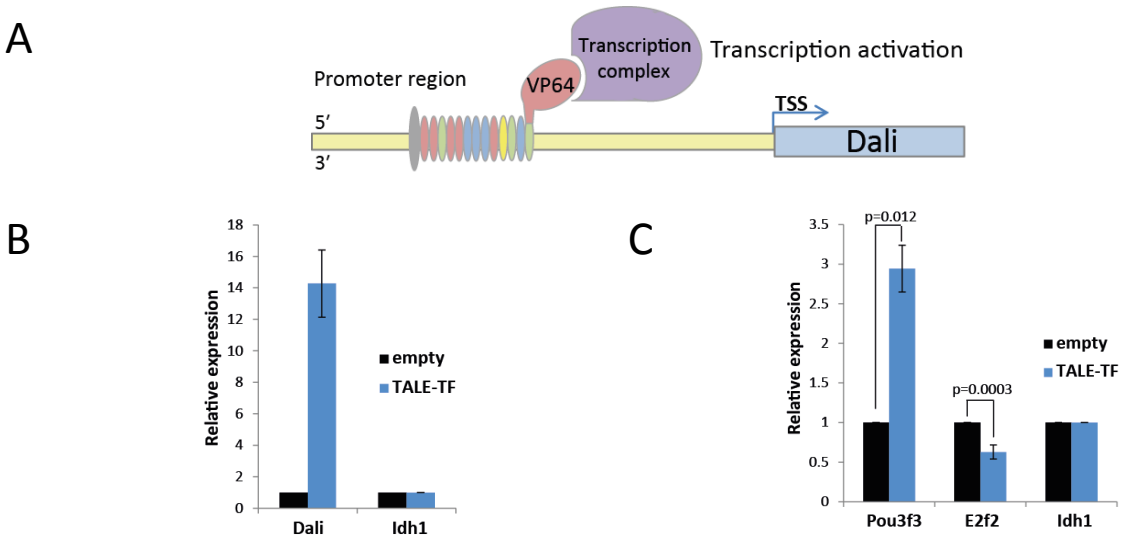


Figure 4.15 *Dali* transcript regulates *Pou3f3* locally and *E2f2* distally in mouse ES cells. (A) A custom transcription factor (TALE-TF) consists of a DNA-binding motif designed to bind DNA sequence upstream of the *Dali* TSS fused to a VP64 transcription activation domain. (B) *Dali* is expressed from its endogenous locus in non-expressing mouse E14 ES cells using custom-designed TALE-TF. (C) *De novo* induction of the endogenous *Dali* locus is sufficient to up-regulate the neighbouring *Pou3f3* gene positively regulated by *Dali* and down-regulate distally located *E2f2* negatively regulated by *Dali*.

4.2.11 *Dali* regulates gene expression during neural differentiation of N2A cells

As discussed previously, *Dali* is up-regulated during neuronal differentiation of ES cells from neuronal precursor stage onwards (Figure 3.1 A). Also, its expression begins in embryonic brain development from day E10.5 when the first few neurons arise from progenitors in cortical areas of neurogenesis (Figure 4.2). Moreover, stable *Dali* knockdown N2A lines exhibit impaired neuritic outgrowth upon differentiation (Figure 4.7 C). Furthermore, *Dali* regulates both local gene expression and the expression of numerous genes involved in neuronal differentiation and function genome-wide (Figure 4.13). Together, these results suggest that *Dali* is involved in neuronal differentiation through regulating its gene expression programmes. Therefore, I decided to study gene expression changes occurring during RA-induced differentiation in both control and stable *Dali* knockdown cells. I treated both cell types with RA under reduced serum conditions for 72 hours (Chapter 2.2.1.3). I then used microarrays to study transcript levels in untreated (proliferating) as well as RA-differentiated (terminally differentiated quiescent) cells of both control and *Dali* knockdown lines (Chapter 2.2.3.7 and 2.2.7.1). I used one stable *Dali* knockdown line for the microarray experiment, and another independent clone to validate the results by qPCR. Prior to differentiation, proliferating control and *Dali* knockdown cells exhibited significant expression differences for 733 genes. 408 genes were up-regulated, while 325 genes down-regulated (Figure 4.16 A, Appendix Table 6). The most populous among these differentially expressed genes were gene ontology categories related to cell differentiation, generation of neurons, apoptosis, cell cycle, cell adhesion, cell projection organization, oxidation reduction, cell signalling, nerve impulse transmission, and cation transport (Figure 4.16 B, Appendix Table 7). Down-regulated genes upon stable *Dali* depletion included *Tgfb2* and *Tgfb3*, both known to be important

for neuronal differentiation and survival (Elvers et al., 2005; Roussa et al., 2013; Vogel et al., 2010). Expression of *Grin1*, previously shown to stimulate neurite extension in N2A cells and play significant role in neuronal differentiation and migration (Masuho et al., 2008), was also significantly reduced. Among the up-regulated genes in stable *Dali* knockdown cells were *Atm*, known to be down-regulated during normal neuronal differentiation (Allen et al., 2001; Carlessi et al., 2013), *Chek1*, known to prevent cell cycle exit associated with terminal neuronal differentiation (Ullah et al., 2011) and *Kif11*, which functions to negatively regulate axonal growth and neuronal migration (Myers and Baas, 2007; Nadar et al., 2012).

After RA was added, 958 genes were differentially expressed between proliferating and differentiated control cells (570 genes were up-regulated, 388 down-regulated; Figure 4.16 C, Appendix Table 8). For *Dali* knockdown cells, a similar number of genes, 1016, were differentially expressed between proliferating and differentiated samples (547 up-regulated, 469 down-regulated; Figure 4.16 E, Appendix Table 9). Differentiation of both control and knockdown cells was associated with differential expression of cell cycle, cell differentiation, nervous system development, cytoskeleton organization, apoptosis, energy metabolism and sterol metabolism genes (Figure 4.16 D, F, Appendix Tables 10 and 11). This suggests that the differentiation programme of the two cell lines were broadly similar. However, terminally differentiated control and *Dali* knockdown cells exhibited 804 differentially expressed genes (385 up-regulated, 419 down-regulated; Figure 4.16 G, Appendix Table 12). Among these, 376 (47%) genes were already differentially expressed between *Dali* knockdown and control cells prior to differentiation (Figure 4.16 I). GO analysis of the 428 differentially expressed genes unique to differentiated cells revealed

enrichment of categories related to energy metabolism (oxidation reduction, carbohydrate metabolic process, cellular alcohol metabolic process), as well as response to chemical stimulus, cell cycle, adhesion, axon extension, small GTPase signalling and ion transport (Figure 4.16 H, Appendix Table 13). Sterol biosynthesis genes, all but one of which were down-regulated in *Dali* knockdown cells, comprised the highest enriched category. Neuritogenesis and neurite outgrowth accompanying neuronal differentiation are known to critically rely on membrane biosynthesis, and therefore, expression of genes involved in sterol biosynthesis (Paoletti et al., 2011). This observation is in agreement with impaired neurite outgrowth which was observed in stable *Dali* knockdown cells differentiated with RA. Perturbed expression of cell adhesion and cell projection organization genes is in agreement with the previously described cellular phenotype of *Dali* knockdown cells. Also in agreement with the observed phenotype is the significant down-regulation of *Rit2*, a neuron specific small GTPase that promotes neurite outgrowth through activation of Rac/Cdc42, association with calmodulin and downstream Plexin B3 signaling (Hartwig et al., 2005; Hoshino and Nakamura, 2003), in differentiated *Dali* knockdown cells.

Several molecules that are important for neuronal differentiation were differentially expressed between *Dali* knockdown and control cells both prior to and following the onset of differentiation. Among those down-regulated in *Dali* knockdown cells were factors crucial for neurite outgrowth, such as *Nrcam*, *Dscam* (Agarwala et al., 2001; Sakurai, 2012), as well as several neuronal transcription factors, such as *Dlx1* and *Pax3*, which are involved in neuronal fate specification and neuronal differentiation (Andrews et al., 2003; Milet et al., 2013; Petryniak et al., 2007; Reeves et al., 1999; Yu et al., 2001). In

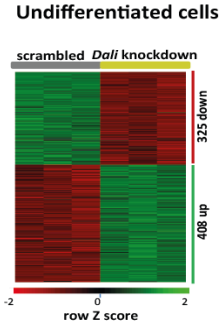
summary, the data suggests that stable *Dali* knockdown cells exist in a different cellular state both prior to and following their terminal differentiation.

To pinpoint potential direct regulatory targets of *Dali* that are involved in the differentiation process, I performed multifactorial analysis of the stable array data sets (Chapter 2.2.7.1) and identified a set of 174 genes that respond to RA differently depending on their *Dali* status (FDR 10%; Appendix Table 14). These genes were enriched in categories relevant to neuronal development, such as cell differentiation, nervous system development, response to chemical stimulus, cell proliferation, neurogenesis, carbohydrate metabolism, cell communication and regulation of transcription (Figure 4.16 K).

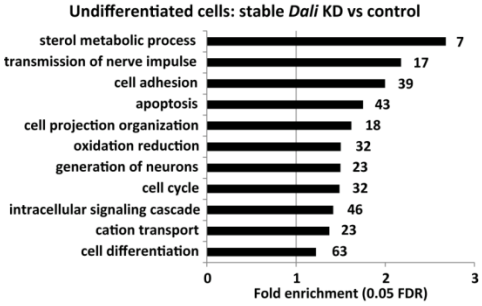
The results of stable *Dali* knockdown microarray experiments were validated in an independent stable cell line carrying an shRNA construct that targets a different region of the *Dali* transcript (Figure 4.17) and are unlikely to represent off-target effects of a specific shRNA or genomic integration event.

In summary, compared to control cells, stable *Dali* knockdown cells exhibit altered response to the differentiation signal. This manifests in mis-expression of a number of genes, including key regulators of neuronal differentiation, as well as the observed phenotype of severely impaired neurite outgrowth development. Therefore, *Dali* regulates not only steady state gene expression levels of some of its targets, but also a dynamic transcriptional response to a differentiation stimulus.

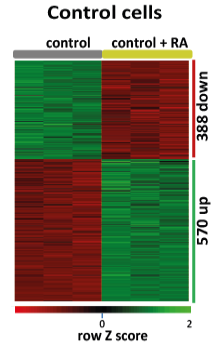
A


Undifferentiated cells
scrambled
Dali knockdown
325 down
408 up
-2
0
2
row Z score

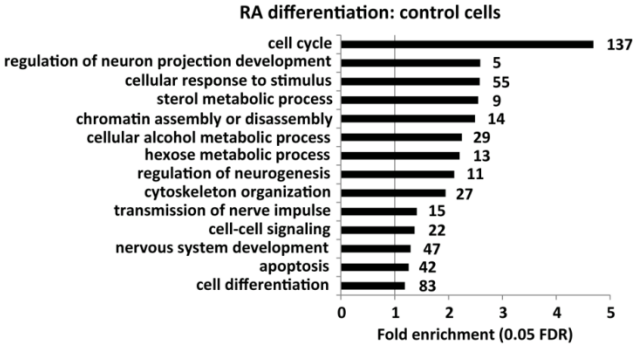
B


Undifferentiated cells: stable Dali KD vs control
sterol metabolic process 7
transmission of nerve impulse 17
cell adhesion 39
apoptosis 43
cell projection organization 18
oxidation reduction 32
generation of neurons 23
cell cycle 32
intracellular signaling cascade 46
cation transport 23
cell differentiation 63
0
1
2
3
Fold enrichment (0.05 FDR)

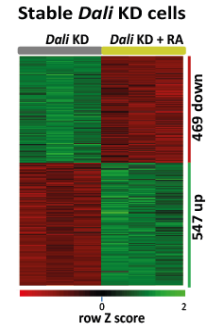
C


Control cells
control
control + RA
388 down
570 up
0
2
row Z score

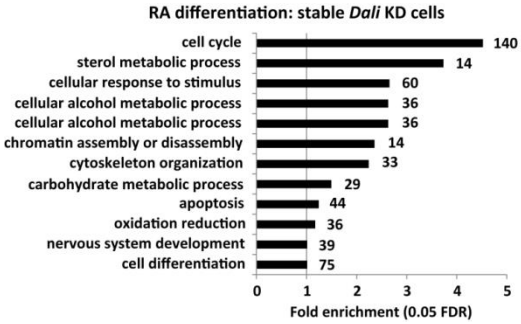
D


RA differentiation: control cells
cell cycle 137
regulation of neuron projection development 5
cellular response to stimulus 55
sterol metabolic process 9
chromatin assembly or disassembly 14
cellular alcohol metabolic process 29
hexose metabolic process 13
regulation of neurogenesis 11
cytoskeleton organization 27
transmission of nerve impulse 15
cell-cell signaling 22
nervous system development 47
apoptosis 42
cell differentiation 83
0
1
2
3
4
5
Fold enrichment (0.05 FDR)

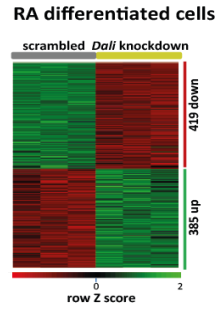
E


Stable Dali KD cells
Dali KD
Dali KD + RA
469 down
547 up
0
2
row Z score

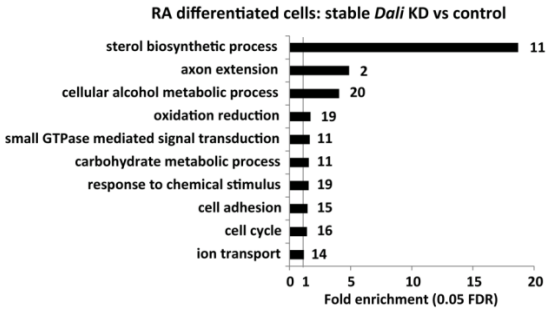
F


RA differentiation: stable Dali KD cells
cell cycle 140
sterol metabolic process 14
cellular response to stimulus 60
cellular alcohol metabolic process 36
cellular alcohol metabolic process 36
chromatin assembly or disassembly 14
cytoskeleton organization 33
carbohydrate metabolic process 29
apoptosis 44
oxidation reduction 36
nervous system development 39
cell differentiation 75
0
1
2
3
4
5
Fold enrichment (0.05 FDR)

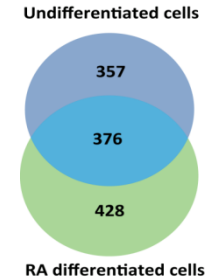
G


RA differentiated cells
scrambled
Dali knockdown
419 down
385 up
0
2
row Z score

H


RA differentiated cells: stable Dali KD vs control
sterol biosynthetic process 11
axon extension 2
cellular alcohol metabolic process 20
oxidation reduction 19
small GTPase mediated signal transduction 11
carbohydrate metabolic process 11
response to chemical stimulus 19
cell adhesion 15
cell cycle 16
ion transport 14
0
1
5
10
15
20
Fold enrichment (0.05 FDR)

I


Undifferentiated cells
357
376
428
RA differentiated cells

K

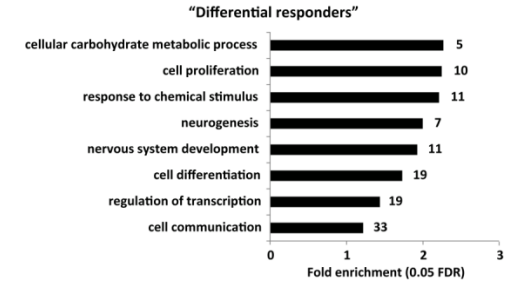

"Differential responders"
cellular carbohydrate metabolic process 5
cell proliferation 10
response to chemical stimulus 11
neurogenesis 7
nervous system development 11
cell differentiation 19
regulation of transcription 19
cell communication 33
0
1
2
3
Fold enrichment (0.05 FDR)

Figure 4.16 Gene expression analyses of stable *Dali* knockdown cells. (A) Stable *Dali* knockdown resulted in statistically significant changes in the expression of 733 genes in N2A cells (1.3-fold, 5% FDR, Appendix Table 6). 325 genes were up-regulated, 408 down-regulated. (B) GO-analysis of genes differentially expressed upon stable *Dali* depletion (5% FDR, hypergeometric test, Benjamini and Hochberg correction). (C-F) Stable *Dali* knockdown and control cells were differentiated with retinoic acid for 72 h. (C) 958 genes were differentially expressed between differentiated and proliferating control cells ((≥1.3-fold, 5% FDR, Appendix Table 8). 570 genes were up-regulated, 388 down-regulated. (D) GO-analysis of genes differentially expressed upon RA-differentiation of control cells (5% FDR, hypergeometric test, Benjamini and Hochberg correction). (E) 1016 genes were differentially expressed between differentiated and proliferating *Dali* knockdown cells ((≥1.3-fold, 5% FDR, Appendix Table 9). 547 genes were up-regulated, 469 down-regulated. (F) GO-analysis of genes differentially expressed upon RA-differentiation of *Dali* knockdown cells (5% FDR, hypergeometric test, Benjamini and Hochberg correction). (G) 804 genes were differentially expressed between differentiated knockdown and control lines ((≥1.3-fold, 5% FDR, Appendix Table 12). 385 genes were up-regulated, 419 down-regulated. (H) GO-analysis of genes differentially expressed only between differentiated stable *Dali* knockdown and control cells (5% FDR, hypergeometric test, Benjamini and Hochberg correction). (I) Intersection of gene sets differentially expressed between stable *Dali* knockdown and control cells prior to (undifferentiated) and after retinoic acid addition (differentiated). (K) GO-analysis of genes responding to retinoic acid treatment differently between stable *Dali* knockdown and control cells (5% FDR, hypergeometric test, Benjamini and Hochberg correction). "Differential responder" genes were identified using multifactorial analysis of the stable *Dali* knockdown arrays using limma (Appendix Table 14) (Smyth, 2004).

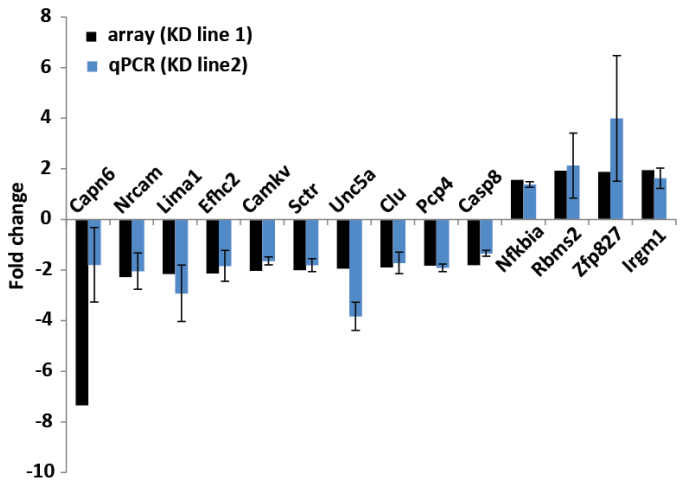


Figure 4.17 Results of the stable *Dali* knockdown microarray experiments were validated by RT-qPCR. Two independent stable *Dali* knockdown cell lines (including the one used for the original microarray study) were used to validate the microarray results. Mean values +/- s.e., n = 3.

4.2.12 Knockdown and over-expression of *Dali* generate distinct transcriptional responses in N2A cells

I also over-expressed *Dali* from its endogenous locus in N2A cells using the same TALE-TF expressed under the control of CMV promoter. Despite *Dali* being over-expressed by approximately 12-fold compared to its original levels in N2A cells, I did not observe changes in expression levels of both *Pou3f3* and *E2f2* in these cells (Figure 4.18 A). This is in contrast to the analysis in ES cells where *Dali* expression is not detectable (Figure 3.1 A). To examine the transcriptional effect of *Dali* genome-wide, transcript levels in *Dali* over-expressing and control cells transfected with an empty vector were measured using microarrays. 237 genes exhibited significant expression changes (FDR < 10%), with 169 genes being up-regulated and 68 genes down-regulated (Figure 4.18 B, Appendix Table 15). The comparison of this gene set with the previously described transient *Dali* knockdown set (Figure 4.13 B) identified 29 genes recurring on both gene lists (Figure 4.18 C). However, no overall correlation between directions and fold-changes was observed between the *Dali* knockdown and over-expression experiments (R=0.27), with only 6 of these genes exhibiting changes in opposite directions as might have been expected (Figure 4.18 D). 36 genes were found to be shared between the *Dali* over-expression and *Pou3f3* knockdown experiments (Figure 4.18 C, Appendix Table 16), which represented a 3.0-fold enrichment over random expectation. However, here also no overall correlation between directions and fold-changes were also observed between the *Dali* over-expression and *Pou3f3* knockdown experiments (R=0.03; Figure 4.18 E). 8 genes were found to be shared between all three data sets, although only 2 displayed

expression changes in opposite direction in the over-expression and knockdown experiments (not shown).

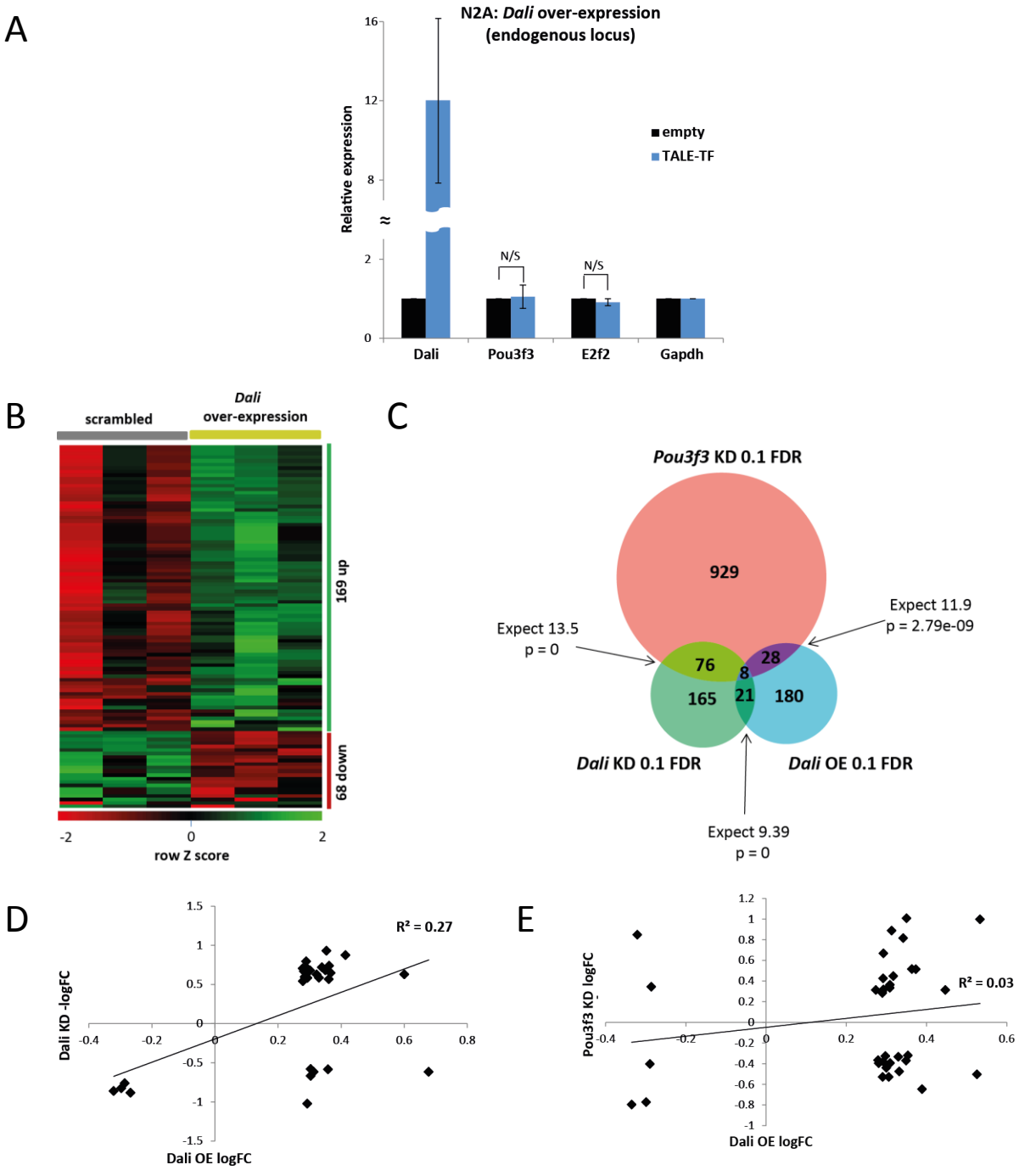


Figure 4.18 Transcriptional response to *Dali* over-expression is different from that of its knockdown. (A) The levels of *Pou3f3* and *E2f2* are unchanged when *Dali* is over-expressed from its endogenous locus in N2A cells using a custom transcription factor (TALE-TF). (B) 237 genes were differentially expressed upon *Dali* over-expression in N2A cells (10% FDR). 169 genes were up-regulated, 68 down-regulated. (C) Intersection between *Dali* knockdown, *Dali* over-expression, and *Pou3f3* knockdown microarray datasets. (D, E) Target genes common between *Dali* knockdown and *Dali* over-expression (D) or *Pou3f3* knockdown (E) show little correlated expression (R=0.27 and R=0.03, respectively).

Overall, the gene expression changes in response to *Dali* over-expression in N2A cells were of a smaller magnitude (with the strongest up-regulated gene changing by 1.6-fold, and the strongest down-regulated gene changing by 1.3-fold). Moreover, the observed changes were largely non-overlapping with those resulting from *Dali* knockdown or *Pou3f3* knockdown. Taken together, these results may indicate that *Dali* is already present in these cells at levels sufficient to fulfil its roles in gene expression regulation, and adding more copies has no additional effect on the expression of its targets. Also, equivalent, rather than opposite, expression changes of some genes upon either knockdown or overexpression could be explained if *Dali* overexpression has a dominant negative effect by titrating away a factor that is required for endogenous *Dali* function. In addition, the required stoichiometry of co-factors may not be achieved for the over-expressed transcript to function effectively at all of its targets.

4.3 Discussion

Cis- and *trans*-acting modes of action of lncRNAs have been illustrated with a growing number of examples (Fatica and Bozzoni, 2014). Also, mammalian expressed lncRNAs have been found to be particularly enriched in the brain (Mercer et al., 2008; Necsulea et al., 2014). A large proportion of these are specifically expressed during brain development and neural differentiation (Ng et al., 2012; Ramos et al., 2013), although lncRNAs involved in the normal function of nervous system, such as regulation of synaptic activity, as well as neurological disease, have also been reported (Bond and Fox, 2009; Clark and Blackshaw, 2014; Gitai et al., 2011; Zhong et al., 2009).

In this Chapter, I investigated the role of a chromatin-associated CNS-specific lncRNA *Dali* in gene expression regulation and neural differentiation. I employed shRNA mediated loss-of-function and TALE-TF mediated gain-of-function experiments in combination with transcriptome profiling using microarrays to identify the nature and extent of the *Dali* genome-wide transcriptional response. I also used the 3C approach to study the spatial conformation of the *Pou3f3* genomic locus in both silent and active states.

I discovered that *Dali* depletion disrupts terminal stages of neural differentiation, more particularly neurite outgrowth development. *Dali* works in a transcript-dependent manner to perform both local and distal gene regulatory roles. Locally, *Dali* regulates expression of its nearest protein-coding gene *Pou3f3* and at least one of the transcripts in its vicinity, *AK053590* (*linc8b*). Distally, it regulates both *Pou3f3*-dependent and -independent target genes. *Dali* provides a second example of such mode of action for an intergenic lncRNA positioned in the vicinity of a neuronal transcription factor. Both *Dali* and the previously

reported *Paupar* lncRNAs come from a set of loci identified by Ponjavic et al. (Ponjavic et al., 2009) as CNS-expressed evolutionary constrained transcripts co-expressed with their neighbouring transcription factor genes both spatially and temporally. Therefore, a combination of local and distal effects may be a common feature of such loci and merits further investigation.

An increasing body of evidence suggests that lncRNAs play an important role in the development of the nervous system and neurological function. Therefore, the function of these molecules may have profound implications for neurodevelopmental disorders. I have shown that *Dali* regulates genes involved in neuronal development and function and its depletion disrupts terminal stages of neural differentiation, in particular, the development of neuronal projection.

I also showed that in ES cell-derived neuronal precursor cells, where both *Pou3f3* and *Dali* are expressed, the *Dali* promoter region makes local genomic contacts with three regions, including a *Pou3f3* enhancer and a differentially methylated region involved in the *Pou3f3* expression regulation. However, *Dali* appears not to play a role in the initiation of these DNA looping interactions because these interactions were also observed in E14 ES cells where neither *Dali* nor *Pou3f3* are expressed. Therefore, in ES cells the *Pou3f3* locus appears to be poised for induction of expression which occurs upon their neuronal differentiation. Interestingly, the expression of *Dali* itself seems to be regulated in a reciprocal manner by *AK053590*, which also regulates *Pou3f3* and *linc-Brn1a* situated to the 5′ of and sharing the promoter region with it. These findings highlight the complexity of interaction involving long non-coding transcripts that regulate the *Pou3f3* transcription

factor gene. Similarly complex regulatory networks may be involved in the expression of multiple other genes across the genome that have lncRNAs transcribed nearby.

Single-exon intergenic lncRNAs are under-represented among transcripts that have been functionally studied. This is because they have often been regarded as likely non-functional products of noisy transcription. Furthermore, few of the single-exon lncRNAs studied in depth to date, except for *Malat1* and *Neat1*, are expressed at high levels (Hutchinson et al., 2007). The single-exon intergenic lncRNA *Dali*, despite its low expression in N2A cells (but considerably higher expression *in vivo*), regulates hundreds of genes genome-wide. Consequently, such single-exon transcripts may be commonly involved in gene expression regulation and merit further functional investigation.

Chapter 5 *Dali* interacts with chromatin modifying proteins and transcription factors and binds to chromatin at promoter-proximal and other regulatory regions to regulate its target genes

5.1 Introduction

The ability of nuclear localised lncRNAs to act *in trans* at distal genomic locations to regulate gene expression programs is currently poorly understood. This is in large part because the direct transcriptional targets of only a small number of such transcripts (for example, *Paupar* (mouse), *HOTAIR*, *NEAT1*, *TERC*, *RMST* (all human), and *rox2* (*Drosophila*)) have been identified thus far (Chu et al., 2011; Ng et al., 2013; Simon et al., 2011; Vance et al., 2014). Consequently, it has been unclear how these transcripts are targeted to distal functional elements and whether thereafter they alter chromatin structure *in situ*. It has been proposed that lncRNAs can interact with chromatin either directly by canonical base-pairing with the underlying DNA or nascent RNA (Buske et al., 2012; Schmitz et al., 2010), or indirectly by chromatin-associated proteins ((Feng et al., 2006; Yang et al., 2013)).

In this Chapter, I investigated how *Dali* regulates gene expression away from its own genomic locus. I took advantage of the Capture Hybridisation Analysis of RNA Targets (CHART)-Seq approach and genes that I previously found to be differentially expressed upon *Dali* depletion (Chapter 4) to identify direct transcriptional targets of *Dali* that are both bound and regulated by it. I also used RNA-affinity chromatography and UV cross-linked RNA Immunoprecipitation (UV-RIP) techniques to reveal proteins that are directly associated with the *Dali* transcript. Together with the genome-wide binding profiling, these

approaches were used to gain insights into the mechanisms of genomic targeting and RNA-dependent mode of action of *Dali*.

I discovered that in proliferating N2A cells *Dali* binds to > 1400 focal loci genome-wide. Overall, *Dali* preferentially (approximately a third of CHART-Seq peaks) binds in proximity of transcriptional start sites (TSSs) of transcriptionally active genes. No evidence was found in the CHART-Seq data that *Dali* RNA is recruited to chromatin via direct interactions with the underlying DNA, either through canonical base-pairing or non-canonical structures, such as RNA:DNA-DNA triplexes. Instead, *Dali* interacts with DNA-binding proteins, including chromatin modifiers and sequence-specific transcription factors. Such interactions are likely to mediate the recruitment of *Dali* to chromatin. Interestingly, *Dali* physically interacts with POU3F3, the protein product of its neighbouring gene, which it also regulates transcriptionally. Together, *Dali* transcript and POU3F3 protein occupy a set of shared loci associated with the genes previously found to be regulated by both factors. Therefore, *Pou3f3*-dependent target genes are regulated by *Dali* both indirectly, via its transcriptional regulatory effect on the *Pou3f3* gene, and directly via its physical interaction with the POU3F3 protein and their co-occupancy at regulatory regions of target genes.

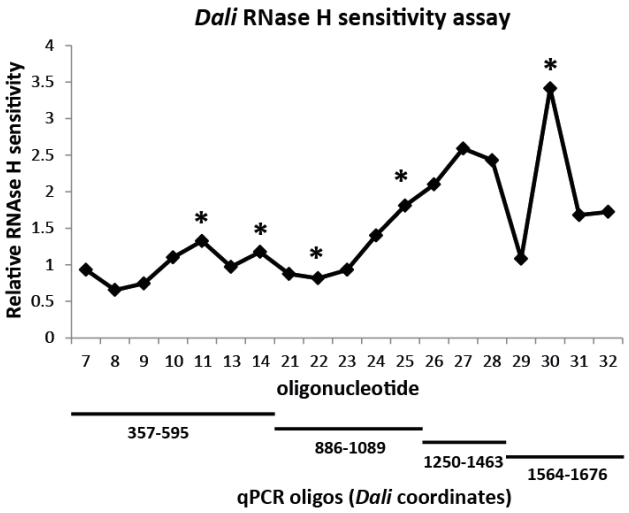
I also showed that in both mouse and human *Dali* interacts with the DNMT1 DNA methyltransferase, hence this transcript's name (*D*NMT1-*A*ssociated *L*ong *I*ntergenic). This finding indicates that the function of this lncRNA may be conserved between the two species, adding to its previously observed sequence conservation and syntenic transcription.

5.2 Results

5.2.1 *Dali* binds to chromatin genome-wide in focal peaks

To identify direct regulatory targets of *Dali* and to gain further insight into how it can regulate multiple genes across the genome, its genome-wide binding profile was determined in N2A cells using the CHART-Seq approach (Simon, 2013; Simon et al., 2011) (Chapter 2.2.3.8). CHART-Seq employs immobilised antisense oligonucleotides to isolate transcripts of interest together with their associated chromatin. Therefore, CHART-Seq identifies direct as well as indirect genomic associations of the RNA of interest. I assayed the accessibility of regions along the first 1700 nt of the *Dali* transcript (a portion of the transcript where efficient qPCR primers could be designed) for binding to antisense oligos using RNase H sensitivity mapping (Figure 5.1 A; Chapter 2.2.3.8). I then designed 5 biotinylated DNA oligos complementary to the more accessible regions of the transcript (cocktail 1). I also designed 5 additional oligos complementary to the most evolutionarily conserved parts of *Dali* sequence (cocktail 2) which lay outside of the region surveyed by the RNase H sensitivity assay. (Figure 5.1 B). Together, the 10 C-oligos designed spanned nearly the full length of the *Dali* sequence. These 10 oligos were used in three capture (C)-oligo cocktails: 2 cocktails of 5 non-overlapping oligos (cocktails 1 and 2), and a pool of all 10 oligos (cocktail 3). As controls, I used an oligo designed to target the antisense *Dali* sequence (not expressed in N2A cells) and *E. coli lacZ* sequence (absent from the mouse genome). Compared to these controls, all three cocktails of C-oligos showed significant enrichment of the *Dali* transcript (10-fold compared to *lacZ*), but no enrichment of the abundant cellular mRNA *Gapdh* (Figure 5.2 A).

A



B

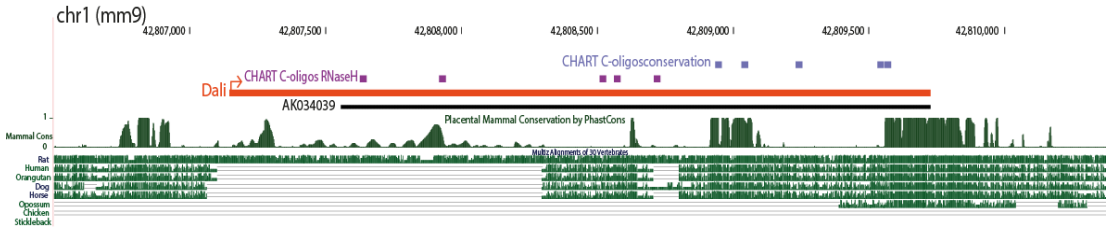


Figure 5.1 *Dali* CHART-Seq C-oligo design. (A) Mapping regions of RNase H sensitivity along the first half of the *Dali* transcript. Individual oligos were added to cross-linked and sheared chromatin extracts prepared from N2A cells. Samples were hybridized, treated with RNaseH to digest the RNA strand in the RNA-DNA hybrids and incubated with DNase I. RNase H cleavage of the *Dali* transcript was assayed by RT-qPCR. RNAse H sensitivity is shown as a ratio between cleaved and uncleaved transcript determined using a no oligo control. qPCR oligos are numbered according to positions on the *Dali* transcript. Asterisks mark oligos used to design CHART-Seq C-oligos. (B) Two sets of C-oligos were designed for the *Dali* CHART-Seq: 5 oligos complementary to RNAseH sensitive (violet; see panel A) and 5 oligos complementary to the evolutionarily conserved (blue) regions of *Dali*.

In the absence of any prior information about *Dali* genomic binding, I used its endogenous site of synthesis to assess the enrichment of transcript-associated DNA loci using qPCR. I observed specific enrichment of *Dali* at its endogenous locus (Figure 5.2 B). Performance of the C-oligo cocktails reflected which parts of the transcripts they targeted, as was to be expected if *Dali* is tethered to the DNA at its locus while still being transcribed. Cocktail 1 composed solely of the C-oligos against the first half of the transcript (Figure 5.1 B) enriched for the DNA of the corresponding part of the locus to the highest degree (approximately 15-fold enrichment over the no oligo control). Cocktail 3 composed of equal amounts of oligos targeting both halves of *Dali* enriched for this portion of the *Dali* locus less efficiently (approximately 6-fold enrichment), while cocktail 2 composed of oligos targeting only the second half of the *Dali* transcript, performed worse than the other two cocktails (approximately 3-fold enrichment) (Figure 5.2 B).

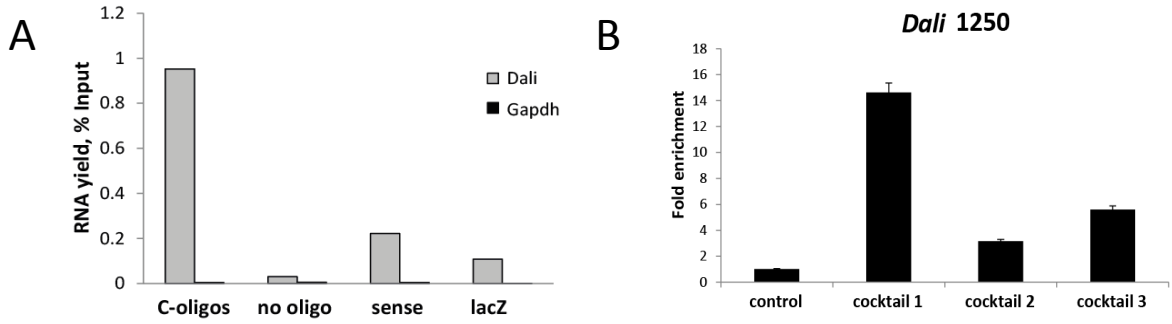


Figure 5.2 CHART-Seq specifically purifies *Dali* RNA and DNA of its endogenous locus. (A) Specific purification of *Dali* RNA using C-oligos compared to controls was assayed by RT-qPCR. (B) The specific enrichment of the *Dali* genomic locus (at position 1250) using C-oligos compared to no oligo control was assayed by qPCR.

Guided by the CHART-Seq protocol, I used RNase H treatment to elute genomic DNA associated with endogenous *Dali* transcripts captured on the oligo-coated beads. The experimental replicates obtained with three *Dali* oligo cocktails (described above) and a matched input DNA sample were sequenced on the Illumina HiSeq platform (50 bp paired-end reads). Two control samples obtained using *lacZ* oligos (GSE52571) were prepared and sequenced separately (Vance et al., 2014). To identify *Dali* binding loci compared to both *lacZ* and input controls, a paired-end peak caller MACS2 was employed (Chapter 2.2.7.2) (Zhang et al., 2008). 4155 peaks were called in all three replicates against the input control. 1928 peaks were identified in all three replicates compared to the two *lacZ* controls. 1427 peaks were shared between the two sets, and therefore were present in all there replicates compared to both types of controls (Figure 5.3 A; Appendix Tables 17, 18). Specificity of CHART pull down was further assessed at nine representative loci using the three *Dali* targeting C-oligo cocktails as well as multiple different non-targeting controls. All nine of the selected loci were successfully validated by an independent CHART-qPCR experiment using the same three oligo cocktails, as well as control *lacZ* and antisense *Dali* oligos (Figure 5.3 B).

The downstream computational analyses were performed on the more stringent set of 1427 peaks called against both input and *lacZ* controls (Chapters 2.2.7.2 and 2.2.7.3). Vast majority of the peaks identified were relatively narrow (Figure 5.4 A) with typical widths between 500-1000 bp in size (Figure 5.4 B), although a small number of peaks spanning several kilobases (up to approximately 7 kb being the longest) were also observed. Therefore, *Dali* peaks were similar to those previously reported for other lncRNAs, such as *Hotair* and *Terc* (Chu et al., 2011), but unlike the recently published CHART-Seq dataset

for *Paupar*, *Malat1* and *Neat1*, where peaks spanning several kb were reported (Vance et al., 2014; West et al., 2014). *Dali* peaks were distributed across the genome with no apparent chromosomal biases, except for the depletion on the X chromosome, which may reflect the inactivation of one X chromosome copy in these female N2A cells (Figure 5.4 C).

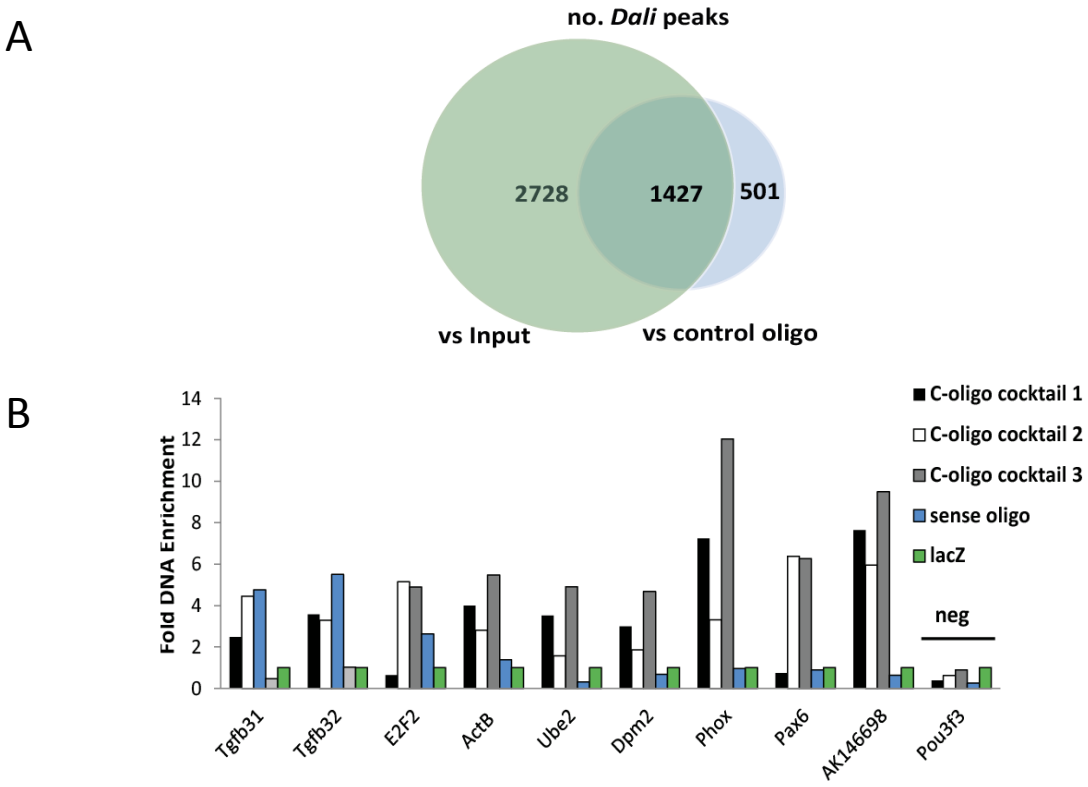


Figure 5.3 CHART-Seq analysis of *Dali* genomic binding sites. (A) Peaks were called by comparing *Dali* CHART-Seq samples with sequences both from control CHART-Seq experiments and from input DNA. For further analysis, only the 1427 peaks common to both comparisons were considered (Appendix Table 18). (B) CHART-seq results were validated by performing an independent experiment with the same three cocktails of oligonucleotides, a control sense oligo complementary to the opposite strand of *Dali* locus and a non-targeting *lacZ* control oligo. Enrichment of genomic regions identified as peaks was assayed by qPCR.

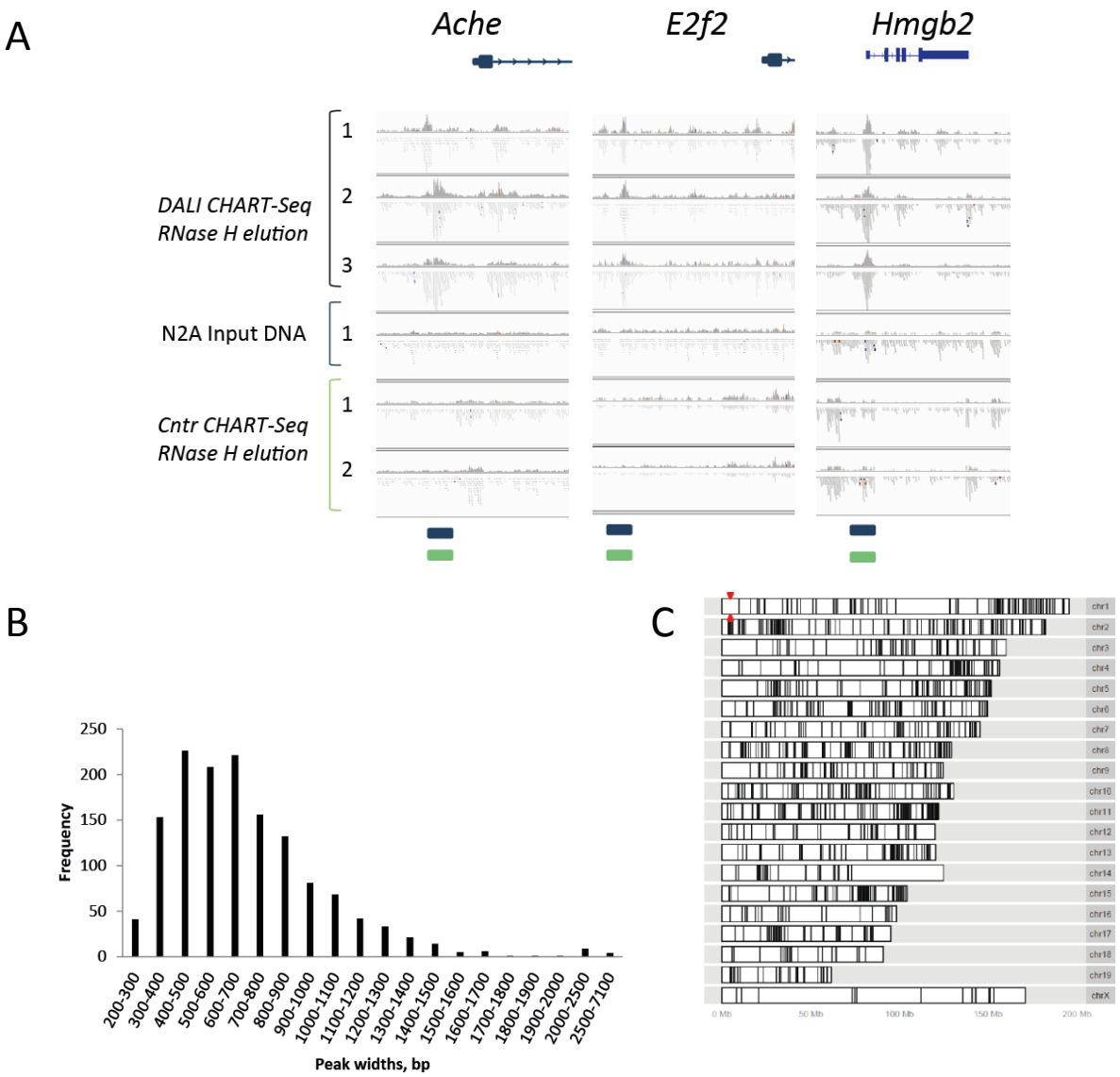


Figure 5.4 CHART-Seq analysis of *Dali* genomic binding sites. (A) Sequencing of *Dali* bound DNA (RNase H elution) reveals focal peaks of *Dali* binding, including those at the promoter of *Ache*, *E2f2* and *Hmgb2*. (B) *Dali* binds to chromatin genome-wide in a focal manner, with most peaks being < 1000 bp wide. (C) *Dali* peaks are broadly distributed across the mouse genome with no apparent chromosomal bias apart from the depletion from the X chromosome, which may reflect the inactivation of one X chromosome copy in these female N2A cells. Red arrowheads indicate the position of the *Dali* locus.

5.2.2 *Dali* binds to chromatin at promoter-proximal and other regulatory regions associated with transcriptionally active genes in N2A cells

To reveal the mechanism whereby *Dali* regulates gene expression, it was important to determine the distribution of *Dali* binding loci with respect to protein-coding gene models. To do so, genome segment analysis of the *Dali* CHART-Seq peak set was performed (Chapter 2.2.7.3). This showed that *Dali* preferentially binds to 5′ flanking regions and 5′ UTRs of protein-coding genes (Figure 5.5 A). Indeed, the Genomic Regions Enrichment of Annotations Tool (GREAT) showed that 30.5% of peaks lay within 5 kb of a transcriptional start site (TSS) of a gene, and 70% within 50 kb of a TSS (Figure 5.5 B), suggesting that *Dali* may modulate the activity of promoter-proximal regulatory regions *in trans* across multiple chromosomes.

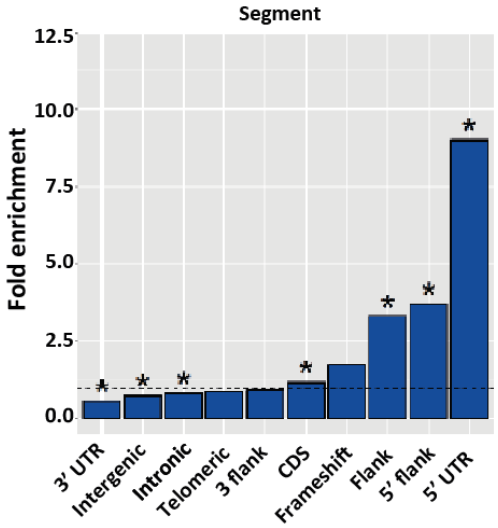
To further characterize regions bound by *Dali*, Genomic Association Test (GAT) was used to perform a chromatin marks enrichment analysis (Chapter 2.2.7.3) using DNase I hypersensitivity (HS) data generated by the Stamatoyannopoulos lab at the University of Washington and chromatin features identified by the Ren lab at the Ludwig Institute for Cancer Research (ENCODE, 2011; Shen et al., 2012). The analysis of chromatin marks showed a significant enrichment of active chromatin marks (H3K4me3, H3K4me1, H3K27ac) and Pol II occupancy sites, and a corresponding significant depletion of repressive marks (H3K9me3, H3K27me3) from the CHART-Seq peaks (Figure 5.5 C). In agreement with the genome segment analysis results, H3K36me, which is usually associated with gene bodies, was also depleted from the *Dali*-bound regions (Figure 5.5 C). Moreover, *Dali* peaks were enriched in DNase I hypersensitive sites which, together with the active mark enrichment, indicates that genes associated with *Dali*-bound loci are

transcriptionally active in N2A cells. *Dali*-bound loci were also significantly more often associated with CpG island (CGI) and enhancer annotations than expected by chance (approximately 10- and 3-fold enrichment, respectively; Figure 5.5 C). To make sure that *Dali* specifically binds to CGI-associated regions as opposed to the CHART-Seq technique non-specifically isolating regions of high GC content, *Dali* CHART-Seq peaks were split into those overlapping a CGI annotation and those not overlapping such an annotation. When GC content of both groups was plotted, no substantial difference between the two categories was observed (Figure 5.5 D). This result suggests that this *Dali* CHART-Seq experiment is unlikely to non-specifically enrich for GC-rich regions, and the enrichment of CGI annotations among the *Dali*-bound regions likely reflects an aspect of *Dali*'s biology.

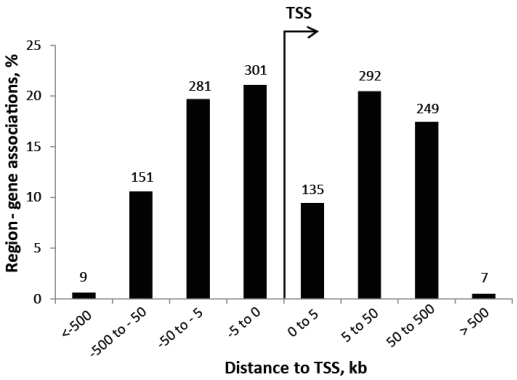
Taken together, the results presented above suggest that *Dali* binds to transcriptional regulatory elements, such as promoters and more distally located enhancers, of transcriptionally active genes. In particular, *Dali* often binds in proximity to CGI-associated promoters.

Figure 5.5 CHART-Seq analysis of *Dali* genomic binding sites. (A) *Dali* binding sites are particularly enriched in 5′ UTRs and gene promoters. (B) A third of *Dali* peaks are situated within 5 kb of a transcriptional start site. (C) *Dali*-bound loci are enriched in active chromatin marks (H3K4me3, H3K27ac, PolII), DNase I hypersensitivity regions, regions annotated as enhancers and CpG islands (CGI), and CTCF-bound regions. At the same time, *Dali* peaks are depleted of gene body marks (H3K36me3) and repressive chromatin marks (H3K9me3 and H3K27me3). (D) *Dali* peaks called against both *lacZ* and input controls that overlap CGI annotations do not have significantly higher GC content than those that do not overlap such annotations.

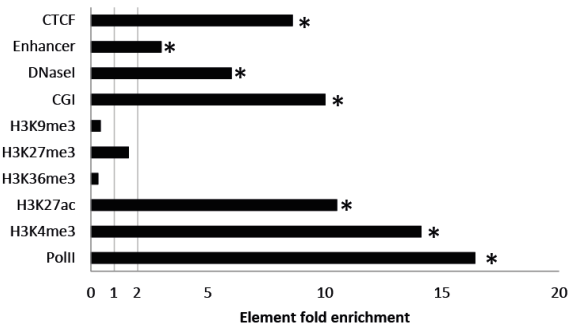
A


Segment
Fold enrichment
12.5
10.0
7.5
5.0
2.5
0.0
3' UTR
Intergenic
Intronic
Telomeric
3 flank
CDS
Frameshift
Flank
5' flank
5' UTR

B


TSS
Region - gene associations, %
25
20
15
10
5
0
9
151
281
301
135
292
249
7
<-500
-500 to -50
-50 to -5
-5 to 0
0 to 5
5 to 50
50 to 500
> 500
Distance to TSS, kb

C


CTCF
Enhancer
DNaseI
CGI
H3K9me3
H3K27me3
H3K36me3
H3K27ac
H3K4me3
PolII
0 1 2 5 10 15 20
Element fold enrichment

D

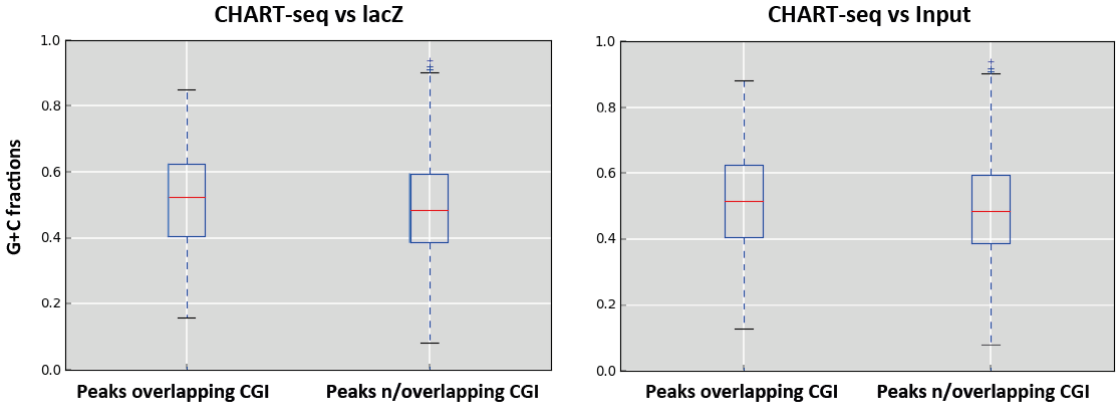


CHART-seq vs lacZ
CHART-seq vs Input
G+C fractions
1.0
0.8
0.6
0.4
0.2
0.0
Peaks overlapping CGI
Peaks n/overlapping CGI

5.2.3 *Dali* binds to chromatin in proximity of its target genes *in trans*

To investigate how lncRNAs regulate genes located away from their site of synthesis, it is crucial to identify direct targets that are the subjects of such regulation. Direct transcriptional targets of *Dali* are expected to be both bound by *Dali* and differentially expressed upon its depletion. In order to identify such targets, I overlapped the list of genes associated with peaks from the stringent *Dali* CHART-Seq set (Appendix Table 19) with a non-redundant list of genes regulated by *Dali* (Appendix Table 20). The genes I assumed to be regulated by *Dali* included the following categories: 1) genes differentially expressed upon transient knockdown of *Dali*; 2) genes differentially expressed in proliferating cells upon stable knockdown of *Dali*; and 3) genes differentially responding to RA treatment in stable *Dali* knockdown cells compared to control. In total, 162 genes were found to be both bound and regulated by *Dali* (Figure 5.6). This constituted 8.7% of all genes associated with *Dali*-bound regions. This finding is similar to the data reported for transcription factor proteins which were also found to regulate a median of 8.1% of the bound genes (Cusanovich et al., 2014). On the other hand, almost 15.1% of genes that have evidence of being regulated are also associated with a *Dali*-bound region (Figure 5.6).

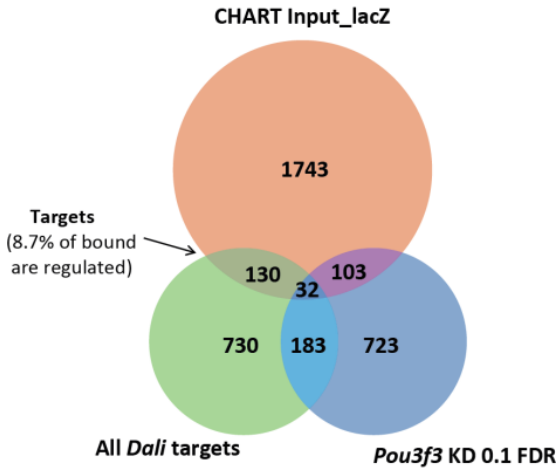


Figure 5.6 *Dali* both binds and regulates a subset of genes. The intersection of genes proximal (<1Mb) to *Dali* peaks, genes regulated by *Dali* and genes changing expression upon *Pou3f3* (10% FDR) knockdown identifies a set of 162 genes both bound and regulated by *Dali*, as well as 32 genes regulated by both *Dali* and *Pou3f3* and directly bound by *Dali*.

There can be multiple reasons why only a minor fraction of binding sites detected by CHART-Seq appear to be functional. These reasons have to do with the limitations of both the CHART-Seq technique itself, as well as the approaches used here to identify regulated and bound genes. Firstly, previous studies looking at transcription factors have indicated that the redundancy inherent in gene regulatory mechanisms makes it difficult to identify the specific effects of knockdowns (Biggin, 2011; Farnham, 2009). Secondly, in my experiments, microarrays were used instead of a more sensitive and comprehensive RNA-Seq approach to study gene expression. Consequently, only genes represented by probes on the arrays were measured. Transcription outside of genes represented on the array and alternative events (such as alternative splicing and alternative promoter usage) affecting genes on the arrays went undetected. Furthermore, microarrays are not sufficiently sensitive to detect differential expression of genes at the extremes of the expression levels range (although, arguably, only a small proportion of protein-coding genes would be affected) (Wang et al., 2009). Therefore, my list of putative *Dali* target genes will contain genes that are not direct targets (false positives) and omit some genes that are true targets (false negatives). Thirdly, peaks were attributed to genes based on genomic proximity along the chromosome which is certain to produce some spurious associations, by wrongly assigning binding in some cases, while in others missing true regulatory interactions. Peak-gene association could potentially be improved by assigning intergenic regulatory regions to genes independently of genomic proximity as suggested, for example, by Thurman et al. (Thurman et al., 2012). These authors used correlations in DNase I hypersensitivity sites between intergenic and promoter-associated hypersensitive sites across diverse tissues to assign intergenic regulatory territories to specific genes.

Also important is the fact that I only profiled *Dali* binding in one cell line (N2A) under one particular set of conditions (normal culture). It can be argued that some of the bound loci may require the presence of specific co-factors available only after a certain stimulus become functional, or may function only under a different set of conditions. Additionally, the CHART-Seq technique relies on formaldehyde cross-linking. Therefore, it is expected to identify both directly bound loci, as well as loci spatially proximal to those in nuclear space.

The fraction of functional binding sites is likely to be underestimated. A potentially large number of CHART-Seq peaks may represent technical artefacts because of all of the reasons listed above. However, calling peaks only when present in all three biological replicates obtained with oligo cocktails (two of which do not share common oligos), as well as against two independent controls (input and *lacZ*), is likely to reduce the number of such artefacts.

Integrating the *Dali* binding data with the results of knockdown studies should allow me to focus on the most likely direct targets. However, more studies of this kind reporting genes both regulated and bound by a chromatin-associated lncRNAs will be needed to evaluate whether what has been found for *Dali* is applicable to more transcripts from this class.

More studies are also needed to discern general principles of the stoichiometry of functional lncRNA-chromatin interactions. Here, I showed that *Dali*, which is expressed at a population-average copy number of 2, directly binds and regulates 162 genes. This can potentially be explained by *Dali* being expressed at higher copy number, but only in a subset of cells in a population. Also, *Dali* may contact more than one region discontinuous in the genome simultaneously, if these exist in spatial proximity in the nucleus. Moreover,

Dali molecules can be re-used to regulate several different genes during their lifetime (the half-life of *Dali* is approximately 30 min, Chapter 4.2.5). CHART-Seq is performed in smaples composed of a large number of cells and, therefore, reflects interactions occurring in a cell population as a whole.

GO category enrichment analysis of genes closest to the CHART-Seq peaks (< 1000 kb) revealed enrichments of genes involved, among others, in cell cycle, cytoskeleton organization, signalling, gene expression, and synaptic transmission (Figure 5.7, Appendix Table 21). These results are in agreement with the previously observed neurite outgrowth defect in stable *Dali* knockdown cells (cytoskeleton organization), as well as the general theme of *Dali* regulating genes important for neuronal function and differentiation (discussed in Chapters 4.2.8, 4.2.9 and 4.2.11).

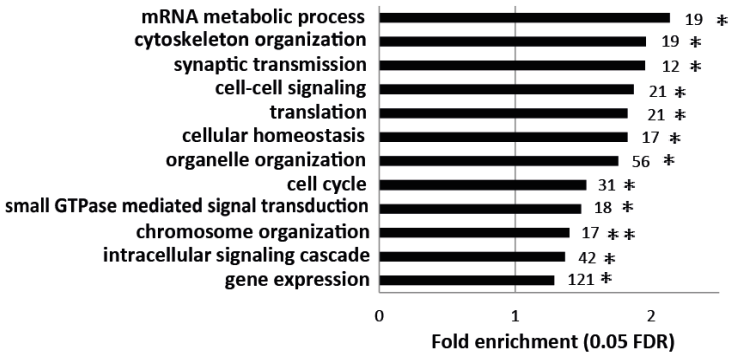


Figure 5.7 Gene Ontology (GO) analysis of genes associated with *Dali* binding sites. Categories of genes within 1Mb of *Dali* binding site are shown (5% FDR, hypergeometric test, Benjamini and Hochberg correction). Representative categories include gene expression, cell cycle, signaling, synaptic transmission and cytoskeleton organization among others. Categories marked with an asterisk (*) are significantly enriched also among genes associated with peaks within 10 kb of a TSS. A category marked with two asterisks (**) is significantly enriched in the gene set associated with peaks within 100 kb (Appendix Table 21).

5.2.4 *Dali* binds in proximity of a subset of its target genes that are also regulated by *Pou3f3*

Previously, I showed that *Dali* and *Pou3f3* shared common regulatory targets that all were differentially expressed upon the knockdown of either factor (Chapter 4.2.9). I also proposed that these common targets may be regulated by *Dali* indirectly via its effect on *Pou3f3* transcription. However, as shown in Figure 5.6, 135 genes that were differentially expressed upon transient *Pou3f3* knockdown were also bound by *Dali*. 32 out of these 135 genes were also regulated by *Dali* (Figure 5.6; see Chapter 5.2.3). This suggests that *Dali* may not only regulate some genes indirectly via its effect on *Pou3f3* expression, but that the *Dali* transcript and POU3F3 protein may regulate some of the shared targets directly by both binding to and regulating transcription of these genes. This observation is similar to what was reported for the intergenic lncRNA *Paupar* which is encoded in the genomic vicinity of another neuronal transcription factor, *Pax6* (Vance et al., 2014). Similar to *Dali*, *Paupar* was found to both regulate the transcription of the *Pax6* gene, as well as jointly regulate a subset of target genes by binding to their regulatory regions together with the PAX6 protein.

The fact that more genes are both bound by *Dali* and regulated by *Pou3f3* than bound by *Dali* and regulated by both factors may arise from one or a combination of the following factors: 1) insufficient depletion of *Dali* in the knockdown studies to cause expression changes of some of these targets; 2) insensitivity of microarray technology to small gene expression changes that may have occurred at these genes; and, 3) redundant regulatory mechanisms compensating for the depletion of *Dali*.

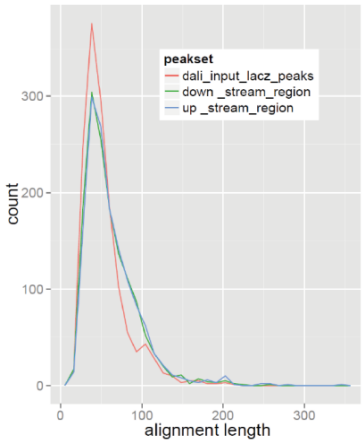
5.2.5 *Dali* is unlikely to be recruited to chromatin via direct RNA-DNA interaction

How chromatin-associated lncRNAs are recruited to specific genomic targets is one of the most important and largely unanswered questions in the field. Several potential mechanisms of how such targeting can be achieved have been proposed. In principle, a lncRNA molecule can associate with a region of DNA other than its own site of synthesis via direct RNA-DNA interaction, by base-pairing with a nascent transcript emanating from the target locus, or indirectly via a DNA-binding protein(s). Direct interaction between RNA and DNA can be governed either by the canonical rules of Watson-Crick base pairing or, alternatively, by the rules governing the formation of non-canonical structures such as RNA-DNA:DNA triplexes (Buske et al., 2012). For example, RNA-DNA:DNA triplex formation is thought to mediate the interaction between promoter-associated RNA (pRNA) transcribed upstream of the rDNA cluster and the rDNA promoter it regulates (Schmitz et al., 2010).

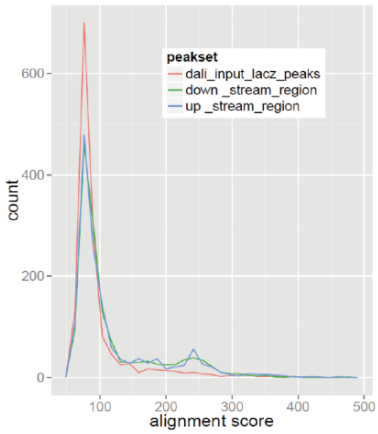
To gain insight into *Dali*'s genomic targeting, *Dali* CHART-Seq data was used to assess whether it may interact with DNA directly. To examine if *Dali* binds to genomic loci containing regions complementary to its sequence, *Dali* sequence was aligned to sequences underlying the CHART-Seq peaks and a matching set of control unbound regions using local alignment tools (Chapter 2.2.7.3). No significant differences were detected between the CHART-Seq and control regions in terms of either alignment length or alignment quality (Figure 5.8 A, B), suggesting that *Dali* is unlikely to interact with its binding locations via canonical Watson-Crick base pairing. Also, the Triplexator suite of software was used to assess whether *Dali* is capable of forming RNA-DNA:DNA triplexes with the regions it binds to (Chapter 2.2.7.3). This software predicts regions within a transcript of

interest that fulfil criteria for being able to form triplexes with DNA. It then searches a set of target DNA regions for fragments capable of partnering with the predicted regions of the transcript of interest to form triplexes. When *Dali* CHART-Seq regions were compared to a set of matched flanking control regions, more matches were found within these control regions. Also, only approximately 6% of peaks contained regions capable of forming triplexes with the *Dali* transcript (Figure 5.8 C). Taken together, these results suggest that there is no evidence that *Dali* is predominantly targeted to chromatin via RNA-DNA:DNA triplex formation.

A



B



C

No. CHART peaks	No. Triplexator-predicted peaks	% predicted RNA-DNA:DNA
1427	86	6.026629292

No. Triplexator-predicted peaks	No. predicted RNA-DNA:DNA	Frequency of hits
86	108	1.255813953

No. Flanking regions	No. predicted RNA-DNA:DNA	Frequency of hits
238	544	2.285714286

Figure 5.8 *Dali* is not recruited to chromatin via direct RNA-DNA interactions. (A, B) Sequence analysis of CHART-seq peak set and *Dali* showed that DNA sequences under peaks are not more complementary to *Dali* sequence than control flanking regions, as judged by either length of aligned regions (A) or alignment quality score (B). (C) DNA sequences under peaks are also not predicted to form more RNA:DNA-DNA triplexes with the *Dali* transcript than control flanking regions.

In summary, *Dali* is unlikely to be recruited to chromatin via direct interaction of its transcript with the underlying DNA, both through canonical base pairing and non-canonical RNA-DNA structures, such as triplexes. By extension, it is also unlikely that *Dali* interacts with chromatin by base pairing with complementary nascent transcripts, if such were produced from the loci bound by it. However, there may be gene specific examples of RNA-DNA targeting of *Dali*.

5.2.6 *Dali* might be recruited to chromatin via interaction with chromatin-associated proteins

As shown above, *Dali* is unlikely to interact with DNA directly. Therefore, I next tested if *Dali* can interact with proteins, which may mediate both its chromatin association and molecular function. I took a two-step approach to identify *Dali*-associated proteins. First, I performed a pull down assay using *in vitro* transcribed and terminally biotinylated *Dali* (Chapter 2.2.3.9) and nuclear protein extract from ES cell-derived neuronal progenitors (RA-differentiated ES cells at day 4). Proteins that remained associated with *Dali* immobilised on magnetic beads after extensive washes were submitted for mass spectrometry analysis (Chapter 2.2.5.2). As controls, I performed analogous experiments with antisense *Dali* and a size-matched unrelated lncRNA. I removed from the resulting list of candidate *Dali*-interacting proteins those that were also found in my control samples or associated with other lncRNA studies in the lab (such as *Paupar* (Vance et al., 2014) and *linc-Sox8* (unpublished)) (Table 5.1).

Because the pull down experiment as performed returned proteins which could interact with *Dali* either directly or indirectly (as a part of larger *Dali*-associated ribonucleoprotein complexes), as well as potentially a large number of false positive interactions (Mili and Steitz, 2004; Riley and Steitz, 2013; Riley et al., 2012), I then proceeded to validate selected candidates using the UV cross-linked RNA Immunoprecipitation (UV-RIP) approach. Because UV irradiation creates a zero-length link between proteins and RNAs, UV-RIP is used for identifying proteins directly interacting with the RNA of interest (Reed and Chiara, 1999). I validated the association of four of the mouse proteins identified *in vitro*, i.e. DNMT1, BRG1, P66b, and SIN3A, with endogenous *Dali* transcript in N2A cells (Figure 5.9 A, C). *Dali*'s interaction with two other candidate proteins, ORC3l and BAF53a, observed in the *in vitro* pull down experiment were not confirmed using UV-RIP in N2A cells (Figure 5.9 B). Therefore, these proteins either interact with *Dali* indirectly, or most likely represent false positives.

DNA methyltransferase DNMT1 is the enzyme that methylates CpG residues preferentially in a context of hemimethylated DNA and is the primary enzyme responsible for the propagation of DNA methylation patterns during S phase of cell division (Li et al., 1992). It also associates with chromatin in G2 and M phases to maintain DNA methylation patterns independently of DNA replication (Easwaran et al., 2004). Furthermore, a growing body of evidence suggests that the activity of DNMT1 is not restricted to DNA methylation maintenance. For example, DNMT1 has been proposed to possess non-catalytic roles in transcriptional regulation of multiple genes genome-wide (Espada et al., 2011), including *E-cadherin* and other examples (Clements et al., 2012).

DNMT1 has been implicated in both maintenance and differentiation of progenitor cells in various tissues. *Dnmt1*-null mouse ES cells show impaired differentiation and are prone to apoptosis, yet have normal proliferative capacity in the undifferentiated state (Lei et al., 1996; Panning and Jaenisch, 1996). In human adult epidermal stem cells DNMT1 promotes self-renewal by repressing genes involved in cell cycle exit and differentiation (Sen et al., 2010). In neurogenesis, DNMT1 blocks precocious astrogliogenesis of neural precursors, although loss of DNMT1 in telencephalic precursors causes cell death and forebrain degeneration (Fan et al., 2005; Hutnick et al., 2009). Overall, the studies to date indicate that the roles of DNMT1 are context-dependent (Georgia et al., 2013).

BRG1 (Brahma-related gene 1) is a member of the SWI/SNF family of proteins whose members have both helicase and ATPase activities and regulate transcription via chromatin remodelling. BRG1 participates in the large ATP-dependent chromatin remodeling complex BAF, which works to either activate or repress transcription depending on the context (Narayanan and Tuoc, 2014). During neural development, the transition from proliferating progenitor cells to post-mitotic neurons is accompanied by a change in subunit composition of the neuronal progenitor-specific (npBAF) and neuron-specific (nBAF) versions of the BAF complex (Lessard et al., 2007). The different BAF complexes have specialised roles at different stages of neural differentiation that arise from their distinct subunit compositions. Different BAF complexes are involved in the maintenance of neuronal stem cells (NSCs), as well as generation of distinct neural and glial cell types (Matsumoto et al., 2006; Narayanan and Tuoc, 2014). Of note, nBAF together with CREST are involved in regulating genes essential for dendrite growth (Wu et al., 2007).

Brg1-null mice die early in embryogenesis and manifest severe neural tube defects (Bultman et al., 2000).

P66beta (as well as the closely related P66alpha which is also found to be associated with *Dali in vitro*) is a transcriptional repressor involved in the MeCP1 histone deacetylase complex which imparts DNA methylation-dependent transcriptional repression (Brackertz et al., 2002).

SIN3A is also a transcriptional repressor. It is a functional component of the REST-CoREST repressor complex. SIN3A-mediated repression may occur via either histone deacetylation and/or direct inhibition of the transcriptional machinery (Grimes et al., 2000). However, it is becoming increasingly apparent that the REST-CoREST complex as well as SIN3A itself can interact and recruit to specific genomic loci a large variety of additional regulatory factors, including DNA methyltransferases and chromatin remodelling complexes (Qureshi et al., 2010). Functionally, REST-CoREST repressor complex is crucial for suppressing neuronal fate outside of the developing nervous system (Grimes et al., 2000) and transiently in neuronal precursors (Abrajano et al., 2010).

Interactions of *Dali* with DNMT1 and BRG1 were further tested using UV-RIP in human neuroblastoma SH-SY5Y cells. For this, I first identified the full-length human ortholog of *Dali* using an RNA sample from human embryonic brain using RACE (Figure 3.5 A) and performed UV-RIP in *Dali*-expressing human neuroblastoma SH-SY5Y cells. DNMT1, but not BRG1, was found to be associated with human *DALI* (Figure 5.9 C, D). Consequently, in further experiments, I focused on the evolutionarily conserved interaction with DNMT1, but not BRG1.

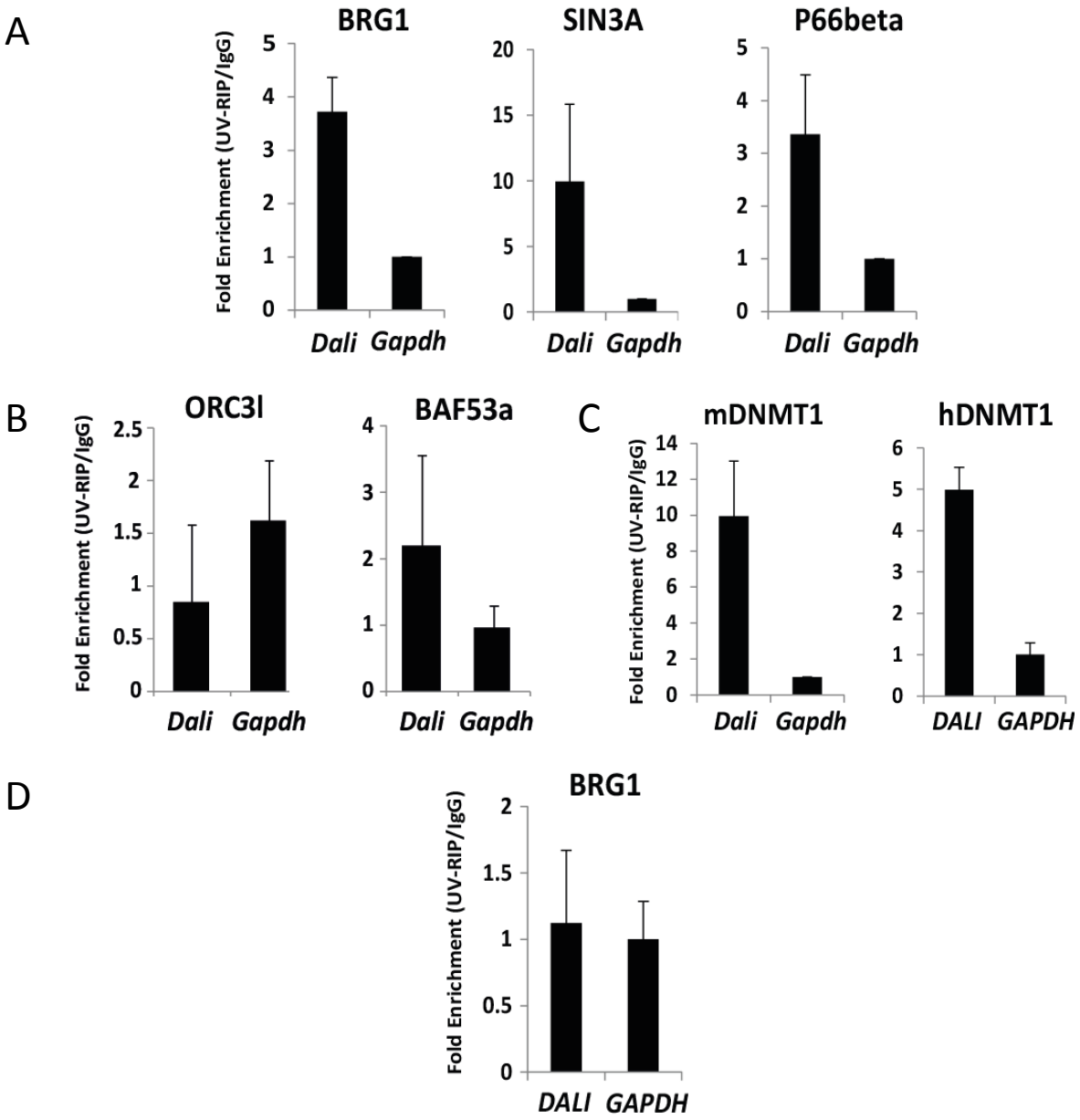


Figure 5.9 *Dali* directly associates with chromatin and transcriptional regulatory proteins. (A) *Dali* interacts with BRG1, SIN3A and P66beta proteins (A), but not ORC3L or BAF53a (B) in mouse N2A cells (A). It also associates with DNMT1 (C), but not BRG1 protein (D) in both mouse N2A and human SH-SY5Y cells. Nuclear extracts were prepared from UV cross-linked cells and immuno-precipitated using either anti-DNMT1 or control IgG antibodies. Associated RNAs were purified and the levels of *Dali* and control *Gapdh* mRNA were detected in each UV-RIP using qRT-PCR. Results are expressed as fold enrichment relative to an isotype IgG control antibody. Results are presented as mean value +/- s.e. of three independent experiments.

In summary, *Dali* interacts with several chromatin-associated proteins which regulate chromatin state and gene activity. Therefore, as *Dali* was previously found unlikely to be recruited to chromatin via direct RNA-DNA interactions, it may to be recruited to DNA via interactions with DNA-binding proteins. Also, *Dali*-interacting proteins may mediate its regulatory effect on its target genes via, as-yet-unknown, mechanisms. How this happens remains to be discovered. *Dali* may also play a role in modulating the activity of its associated proteins.

Table 5.1 Summary of mass spectrometry analysis of protein associated with *Dali in vitro*.

sp|O88477|IF2B1_MOUSE	Insulin-like growth factor 2 mRNA-binding protein 1 OS=Mus musculus GN=Igf2bp1 PE=1 SV=1	Igf2bp1
sp|P09405|NUCL_MOUSE	Nucleolin OS=Mus musculus GN=Ncl PE=1 SV=2	Ncl
sp|Q04750|TOP1_MOUSE	DNA topoisomerase 1 OS=Mus musculus GN=Top1 PE=1 SV=2	Top1
sp|Q7TNV0|DEK_MOUSE	Protein DEK OS=Mus musculus GN=Dek PE=1 SV=1	Dek
sp|Q920B9|SP16H_MOUSE	FACT complex subunit SPT16 OS=Mus musculus GN=Supt16h PE=1 SV=2	Supt16h
tr|Q3UJB0|Q3UJB0_MOUSE	Protein Sf3b2 OS=Mus musculus GN=Sf3b2 PE=2 SV=1	Sf3b2
sp|P25976-2|UBF1_MOUSE	Isoform UBF2 of Nucleolar transcription factor 1 OS=Mus musculus GN=Ubtf	Ubtf
sp|Q8BK67|RCC2_MOUSE	Protein RCC2 OS=Mus musculus GN=Rcc2 PE=2 SV=1	Rcc2
sp|Q8VIJ6|SFPQ_MOUSE	Splicing factor, proline- and glutamine-rich OS=Mus musculus GN=Sfpq PE=1 SV=1	Sfpq
sp|Q99K48|NONO_MOUSE	Non-POU domain-containing octamer-binding protein OS=Mus musculus GN=Nono PE=1 SV=3	Nono
sp|Q91VM5|RMXL1_MOUSE	RNA binding motif protein, X-linked-like-1 OS=Mus musculus GN=Rbmxl1 PE=1 SV=1	Rbmxl1
tr|E9Q0F0|E9Q0F0_MOUSE	Protein Krt78 OS=Mus musculus GN=Krt78 PE=4 SV=1	Krt78
tr|Q921K2|Q921K2_MOUSE	Poly (ADP-ribose) polymerase family, member 1 OS=Mus musculus GN=Parp1 PE=2 SV=1	Parp1
sp|P63038|CH60_MOUSE	60 kDa heat shock protein, mitochondrial OS=Mus musculus GN=Hspd1 PE=1 SV=1	Hspd1
sp|P80314|TCPB_MOUSE	T-complex protein 1 subunit beta OS=Mus musculus GN=Cct2 PE=1 SV=4	Cct2
sp|Q3URS2|ZSC4F_MOUSE	Zinc finger and SCAN domain containing protein 4F OS=Mus musculus GN=Zscan4f PE=2 SV=2	Zscan4f
sp|P13864-2|DNMT1_MOUSE	Isoform 2 of DNA (cytosine-5)-methyltransferase 1 OS=Mus musculus GN=Dnmt1	Dnmt1
sp|P17182|ENOA_MOUSE	Alpha-enolase OS=Mus musculus GN=Eno1 PE=1 SV=3	Eno1
sp|P80315|TCPD_MOUSE	T-complex protein 1 subunit delta OS=Mus musculus GN=Cct4 PE=1 SV=3	Cct4
sp|P97496-2|SMRC1_MOUSE	Isoform 2 of SWI/SNF complex subunit SMARCC1 OS=Mus musculus GN=Smarcc1	Smarcc1
sp|Q60520-1|SIN3A_MOUSE	Isoform 2 of Paired amphipathic helix protein Sin3a OS=Mus musculus GN=Sin3a	Sin3a
sp|Q8K4Z5|SF3A1_MOUSE	Splicing factor 3A subunit 1 OS=Mus musculus GN=Sf3a1 PE=1 SV=1	Sf3a1
sp|Q8VHR5-2|P66B_MOUSE	Isoform 2 of Transcriptional repressor p66-beta OS=Mus musculus GN=Gatad2b	Gatad2b
sp|Q9Z2N8|ACL6A_MOUSE	Actin-like protein 6A OS=Mus musculus GN=Actl6a PE=1 SV=2	Actl6a
sp|O54879|HMGB3_MOUSE	High mobility group protein B3 OS=Mus musculus GN=Hmgb3 PE=2 SV=3	Hmgb3

sp|O54941|SMCE1_MOUSE	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1 OS=Mus musculus GN=Smarce1 PE=1 SV=1	Smarce1
sp|P14733|LMNB1_MOUSE	Lamin-B1 OS=Mus musculus GN=Lmnb1 PE=1 SV=3	Lmnb1
sp|Q61937|NPM_MOUSE	Nucleophosmin OS=Mus musculus GN=Npm1 PE=1 SV=1	Npm1
sp|Q6P542|ABCF1_MOUSE	ATP-binding cassette sub-family F member 1 OS=Mus musculus GN=Abcf1 PE=1 SV=1	Abcf1
sp|Q8CHY6|P66A_MOUSE	Transcriptional repressor p66 alpha OS=Mus musculus GN=Gatad2a PE=1 SV=2	Gatad2a
sp|P68040|GBLP_MOUSE	Guanine nucleotide-binding protein subunit beta-2-like 1 OS=Mus musculus GN=Gnb2l1 PE=1 SV=3	Gnb2l1
sp|Q7SIG6|ASAP2_MOUSE	Arf-GAP with SH3 domain, ANK repeat and PH domain-containing protein 2 OS=Mus musculus GN= PE=1 SV=3	Asap2
tr|E9PVX6|E9PVX6_MOUSE	Protein Mki67 OS=Mus musculus GN=Mki67 PE=4 SV=1	Mki67
sp|P80317|TCPZ_MOUSE	T-complex protein 1 subunit zeta OS=Mus musculus GN=Cct6a PE=1 SV=3	Cct6a
sp|Q3U9G9|LBR_MOUSE	Lamin-B receptor OS=Mus musculus GN=Lbr PE=1 SV=2	Lbr
sp|Q3TKT4-2|SMCA4_MOUSE	Isoform 2 of Transcription activator BRG1 OS=Mus musculus GN=Smarca4	Smarca4
sp|Q60972|RBBP4_MOUSE	Histone-binding protein RBBP4 OS=Mus musculus GN=Rbbp4 PE=1 SV=5	Rbbp4
sp|Q8C1B7-2|SEP11_MOUSE	Isoform 2 of Septin-11 OS=Mus musculus GN=Sept11	Setp11
sp|P11031|TCP4_MOUSE	Activated RNA polymerase II transcriptional coactivator p15 OS=Mus musculus GN=Sub1 PE=1 SV=3	Sub1
sp|P27773|PDIA3_MOUSE	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1 SV=2	Pdia3
sp|P0CG49|UBB_MOUSE	Polyubiquitin-B OS=Mus musculus GN=Ubb PE=1 SV=1	Ubb
sp|Q3TB82|PKHF1_MOUSE	Pleckstrin homology domain-containing family F member 1 OS=Mus musculus GN=Plekhf1 PE=2 SV=1	Plekhf1
tr|F6QFD1|F6QFD1_MOUSE	cAMP-specific 3',5'-cyclic phosphodiesterase 4D (Fragment) OS=Mus musculus GN=Pde4d PE=3 SV=1	Pde4d
sp|P80316|TCPE_MOUSE	T-complex protein 1 subunit epsilon OS=Mus musculus GN=Cct5 PE=1 SV=1	Cct5
sp|P19324|SERPH_MOUSE	Serpin H1 OS=Mus musculus GN=Serpinh1 PE=1 SV=3	Serpinh1
sp|P63158|HMGB1_MOUSE	High mobility group protein B1 OS=Mus musculus GN=Hmgb1 PE=1 SV=2	Hmgb1
sp|Q9ERK4|XPO2_MOUSE	Exportin-2 OS=Mus musculus GN=Cse1l PE=2 SV=1	Cse1l
sp|Q99JF8-2|PSIP1_MOUSE	Isoform 2 of PC4 and SFRS1-interacting protein OS=Mus musculus GN=Psip1	Psip1
sp|Q9JK30-2|ORC3_MOUSE	Isoform 2 of Origin recognition complex subunit 3 OS=Mus musculus GN=Orc3	Orc3
sp|Q9CWS4|INT11_MOUSE	Integrator complex subunit 11 OS=Mus musculus GN=Cpsf3l PE=2 SV=1	Cpsf3l

5.2.7 *Dali* binding loci are enriched for consensus motifs of known DNA-binding proteins

So far, results support the hypothesis that *Dali* relies on proteins for its recruitment to chromatin, and for its action *in trans*. Mass spectrometry analysis, however, is limited to identifying the more abundant proteins and neither the UV-RIP confirmed protein partners nor the full mass spectrometry list contained sequence-specific DNA-binding proteins among them, perhaps because of their tendency to be less abundant. Consequently, *Dali* CHART-Seq data were subject to *de novo* motif discovery analysis to investigate whether regions bound by *Dali* contain recurrent DNA-binding factor sequence motifs (Chapter 2.2.7.3). The resulting list of these *de novo* discovered motifs was then searched for matches to known TF binding sites (Chapter 2.2.7.3).

De novo motif discovery revealed a motif that nearly-perfectly matched a consensus CTCF binding motif in 125 out of 1427 *Dali* peaks (8.8% of peaks; Figure 5.10 A). This result was in agreement with the greater than expected overlap also observed between *Dali* CHART-Seq and publicly available CTCF ChIP-Seq peaks from neuronal tissues (Figure 5.5 C) (Shen et al., 2012). To investigate whether *Dali* and CTCF bind to shared loci, I performed ChIP-qPCR experiment at a number of *Dali*-bound and regulated promoters that had previously been found to be bound by CTCF in ChIP-Seq experiments in neuronal tissues (Shen et al., 2012). I observed enrichment of CTCF at these regions in N2A cells, but not at control regions that were not predicted to be bound by either CTCF or *Dali* (Figure 5.10 B). Therefore, CTCF and *Dali* occupy a subset of shared genomic binding sites.

The two factors might bind to closely spaced loci independently of each other, or compete for binding to the same region, or jointly bind to the co-occupied regions. The latter possibility raised a question of whether *Dali* and CTCF might directly interact with each other. However, UV-RIP analysis showed no evidence of direct physical interaction between *Dali* and CTCF (Figure 5.10 C). That said, it is plausible that *Dali* and CTCF interact indirectly within a larger ribonucleoprotein complex. For example, *Dali*'s regulatory effect on transcription of a subset of its targets might be mediated by CTCF through its association with *Dali*-interacting proteins, such as DNMT1, or through competition between *Dali*-containing complexes and CTCF for the same or closely spaced binding sites. In fact, DNMT1 is known to occur in a complex with CTCF and PARP1 also identified as associated with *Dali* in the pull down experiment (Table 5.1) (Guastafierro et al., 2008). In this complex, CTCF and auto-poly(ADP-ribosyl)ated PARP1 inhibit DNMT1 catalytic activity and prevent methylation of CpG dinucleotides in the vicinity of unmethylated CTCF target sequences. Depletion of ADP-ribose polymers was shown to result in the methylation of underlying sequences by DNMT1 and subsequent eviction of CTCF (Zampieri et al., 2012), because binding of CTCF is inhibited when its recognition sequences are methylated (Caiafa et al., 2009).

Taken together, these results indicate that *Dali* is likely to be targeted to the genome via interaction with chromatin-associated proteins, either directly or indirectly via its non DNA sequence-specific protein partners being associated with other, sequence-specific, proteins (such as CTCF).

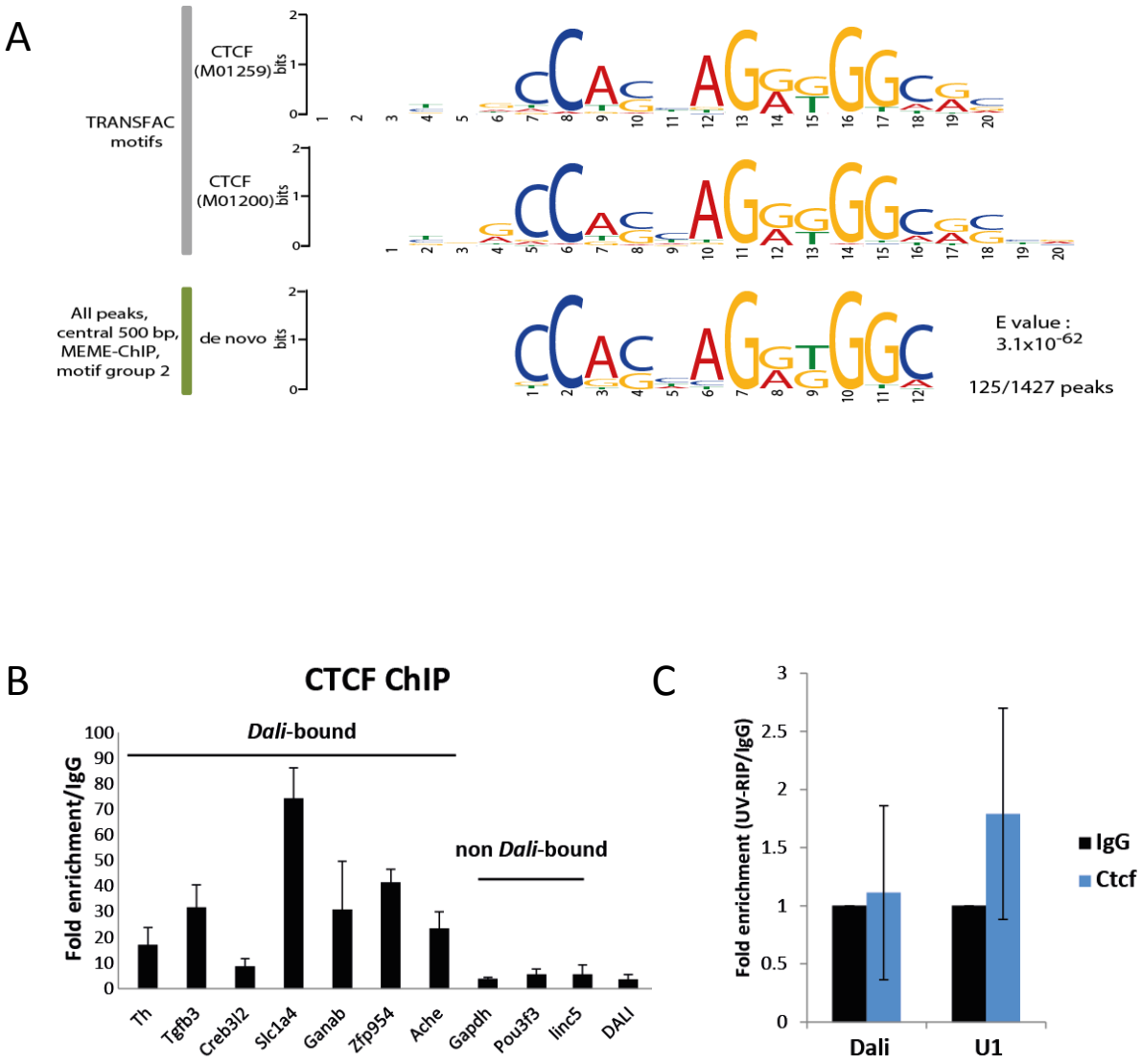
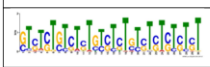
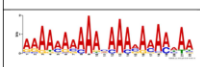
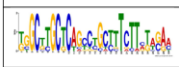
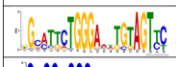
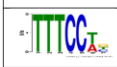
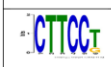
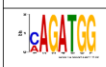
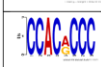
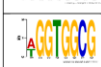
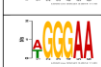


Figure 5.10 *Dali* co-occupies a subset of CTCF binding loci. (A) *De novo* motif discovery revealed a near-perfect match to a CTCF binding motif in 125 out of 1427 (8.8%) *Dali* CHART-seq peaks. (B) *Dali* occupies a set of locations shared with CTCF. Control regions are not predicted to be bound by CTCF and are not bound by *Dali*. ChIP assays were performed in N2A cells using either an antibody against CTCF or an isotype specific control. The indicated DNA fragments were amplified using qPCR. Fold enrichment was calculated as 2-ΔΔCt (IP/IgG) and is presented as mean value +/- s.e. of at least three independent experiments. (C) However, *Dali* does not interact with CTCF directly in the UV-RIP assay.

Table 5.2 Summary of *de novo* motif discovery analysis of CHART-Seq peaks. CHART-Seq peaks are enriched for known binding motifs of multiple transcription factors.

Representative motif	E value	No. peaks	Transfac motif matches
	1.9e-175	332	MAZ (M02023), Zfp281 (M01597), KROX (M00982)
	3.1e-062	125	CTCF (M01259, M01200), MATH1 (M01716)
	4.8e-122	32	SMAD3 (M00701), GAGA (M00723)
	2.3e-045	41	FOXP1 (M00987), FOXO1 (M01216), BR-C (M00092)
	5.3e-009	10	Dde (M00304), BPC1 (M01126)
	5.4e-038	23	Pax-6 (M00979), LRH1 (M01142), ZBED6 (M01598)
	5.7e-046	24	STATx (M00223), Staf (M00264), STAT1 (M01823)
	7.8e-036		Zfp281 (M01597), UF1H3BETA (M01068), MAZ (M02023)
	8.8e-028		GAGA (M00723)
	7.6e-023		ML1 (M01193), Lentiviral (M00318), PBF (M00355)
	3.1e-021		FPM315 (M01587), ETF (M00695)
	1.7e-019		NFAT3 (M01734), NF-AT (M00935), PARP (M02027)
	1.3e-016		C-ets-1 (M01870), Tel-2 (M00678), ELF1 (M01266)
	2.9e-013		Tal-1beta:ITF-2 (M00070), Tal-1beta:E47 (M00065) Tal-1alpha:E47 (M00066)

Representative motif	E value	Transfac motif matches
	7.3e-013	GT-1 (M01827), FOXO1 (M01216), HNF3alpha (M00724)
	4.5e-012	Smad3 (M01888), SMAD (M00974)
	2.0e-010	PARP (M02027), STAT5A (M01890), BPC1 (M01126)
	2.0e-010	Zfp206 (M01742), NRF-1 (M00652), AP-2 (M00189)
	7.7e-007	Pax-6 (M00979), Pax (M00808), HSF (M00641)
	2.5e-006	AP-1 (M00174, M00173, M00926)
	3.0e-006	CACD (M01113), EKLF (M01874), GKLF (M01588)
	2.9e-006	USF (M00121), TFEB (M01768), CBF1 (M01793)
	2.9e-005	RPN4 (M01925), MyoD (M00929), ERF2 (M01057)
	3.8e-005	Lhx3a (M00510), IPF1 (M01233), IPF1 (M01235)
	6.9e-004	RBP-Jkappa (M01111), Ik-3 (M00088)
	1.8e-002	MADS-A (M00408), MADS-B (M00404), YY1 (M01894)
	2.1e-002	Sp1 (M00931, M00933, M00196)
	2.1e-002	SIG1 (M01631), Croc (M00266)

5.2.8 *Dali* binding loci are enriched for consensus motifs of multiple transcription factors

In addition to CTCF, the *de novo* motif discovery analysis performed on the *Dali* CHART-Seq data revealed significant enrichment of motifs matching consensus binding sequences for numerous other transcription factors (Table 5.2, above). This observation is in agreement with the hypothesis that *Dali*'s interaction with chromatin might be mediated by the interaction of its protein partners, such as DNMT1 and BRG1 (or BAF complex), with

sequence-specific transcription factors. To further evaluate this information, I decided to compare the list of transcription factors whose consensus motifs were enriched under the *Dali* peaks to the two lists of transcription factors that have been reported as either interacting specifically with DNMT1 (but not with other DNMTs), or with other DNMTs (but not with DNMT1) in a protein array experiment (Hervouet et al., 2010). 9 of 52 transcription factors that were found to interact specifically with DNMT1 were also on the list of *de novo* motifs enriched in the *Dali* CHART-Seq data. At the same time, none of the 42 proteins reported as interacting with other DNMTs but not with DNMT1 were present on that list. Therefore, motifs recognised by transcription factors that interact with one of *Dali*'s partners, DNMT1, are over-represented in *Dali*-bound regions, which supports the hypothesis that transcription factors provide sequence-specificity for *Dali* recruitment to chromatin.

5.2.9 *Dali* and POU3F3 directly interact and co-occupy a number of genomic binding sites within the regulatory regions of shared target genes

Previously, I showed that *Dali* and *Pou3f3* regulate a set of common genes (Figure 4.14 D). Moreover, *Dali* occupies regulatory regions of 135 genes that are differentially expressed upon *Pou3f3* knockdown, including 32 genes that are also bound by *Dali* (Figure 5.6). Based on these results, I hypothesized that *Dali* and POU3F3 protein may both bind and regulate regulatory regions of a set of common target genes. In support of this hypothesis, *de novo* motif discovery analysis of the CHART-Seq data revealed a motif for the POU III family of transcription factors in *Dali*-bound regions. This family includes

POU3F3. All members of the POU III family have a nearly identical DNA-binding domain and, therefore, can recognise very similar DNA sequences (Figure 5.11 A; (Zhao, 2013)). The POU III motif was discovered in 115 out of 1427 peaks (8.1% peaks, E-value = $3.8x10^{-5}$; Figure 5.11 A). Therefore, *Dali* may not only regulate some genes indirectly via its effect on *Pou3f3* transcription, but may also regulate some of their shared targets through its interaction with the POU3F3 protein.

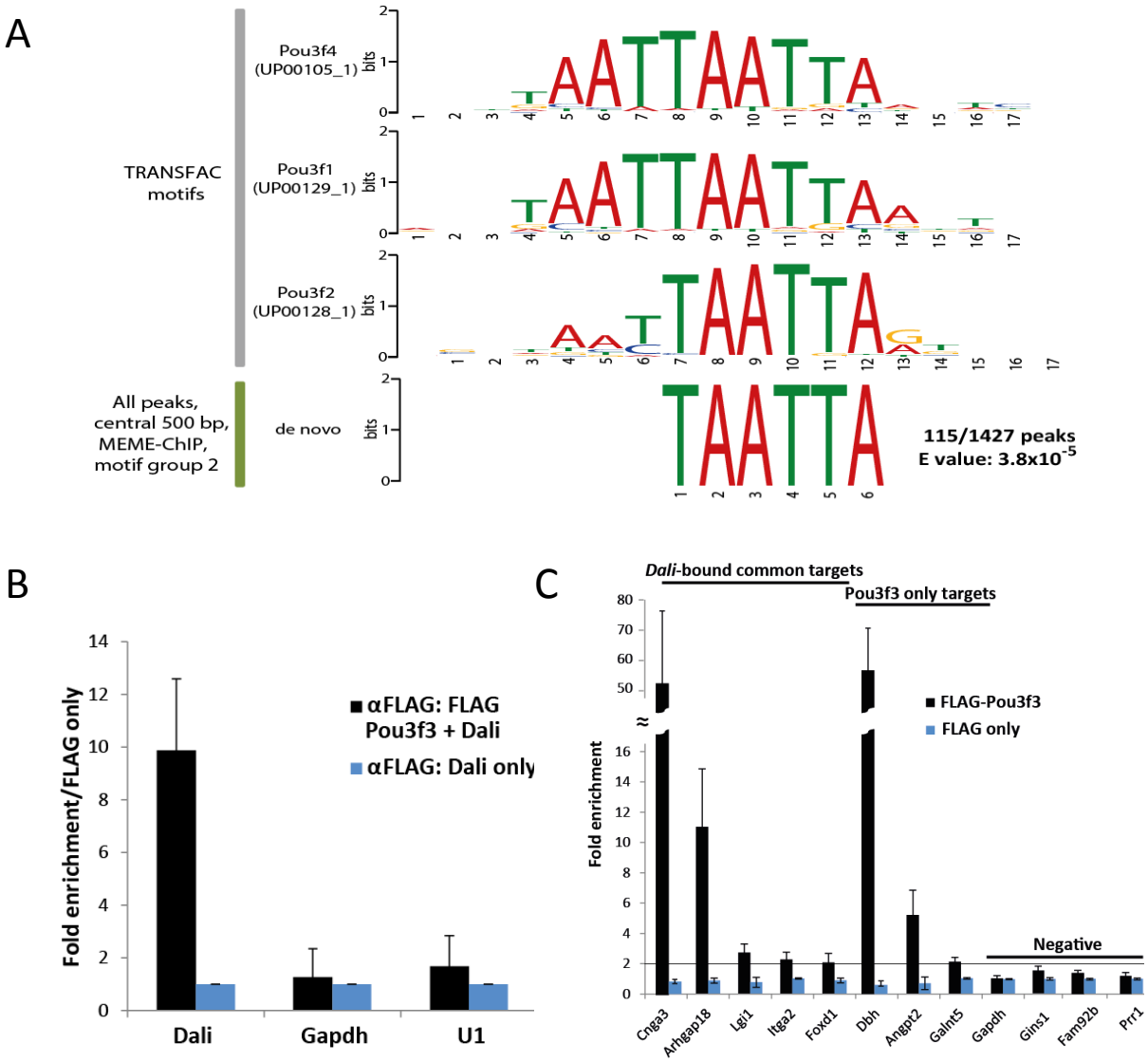


Figure 5.11 *Dali* interacts with the POU3F3 protein and occupies a set of shared loci with it. (A) *De novo* motif discovery identifies a motif for POU III family transcription factors (which includes POU3F3) in 115 out of 1427 (8.1%) *Dali* CHART-seq peaks. (B) UV-RIP in N2A cells shows that FLAG-tagged POU3F3 protein directly interacts with *Dali*. (C) ChIP-qPCR in N2A cells shows that POU3F3 occupies a subset of loci associated with *Dali* CHART-seq peaks and that are regulated by both *Pou3f3* and *Dali*. Loci associated with known (*Dali*-independent) *Pou3f3* targets were used as positive control, while loci not regulated by either *Pou3f3* or *Dali* and not bound by *Dali* were used as negative control.

5.2.10 *Dali* directly interacts with the POU3F3 protein which binds to a set of loci associated with genes bound by *Dali* and regulated by both factors

Previously I showed that *Dali* and *Pou3f3* together regulate a set of common genes, some of which are also directly bound by *Dali*. Also, 8.1% of *Dali*-bound regions contain a motif which can be bound by the POU3F3 protein. Together, these results raised the possibility that *Dali* and POU3F3 protein may physically interact. To investigate this hypothesis, I performed UV-RIP in N2A cells (Chapters 2.2.2.10 and 2.2.3.11). Due to the poor performance of commercially available anti-POU3F3 antibodies in this application and low abundance of the wild type POU3F3 in N2A cells, I therefore performed UV-RIP in N2A cells over-expressing both *Dali* and FLAG-tagged POU3F3, as well as N2A cells over-expressing *Dali* and carrying empty FLAG vector as negative control (Chapters 2.2.2.10 and 2.2.3.11). UV-RIP results revealed robust physical interaction between *Dali* and POU3F3 (Figure 5.11 B) supporting the hypothesis that *Dali* and POU3F3 may jointly bind and regulate a set of common target genes.

To further test this hypothesis, I studied POU3F3 occupancy at *Dali*-bound regions associated with commonly regulated target genes, as well as regulatory regions of known *Pou3f3* targets (positive controls) and regions not predicted to be bound by *Dali* and not regulated by *Pou3f3* (negative controls) using ChIP-qPCR (Chapter 2.2.2.19). I observed that some of the *Dali*-bound regions associated with common target genes were also occupied by POU3F3 protein. Similar enrichments were observed at known POU3F3 targets, while negative control regions showed no enrichment over empty vector control (Figure 5.11 C).

In summary, these results suggest that *Dali* and POU3F3 physically interact and jointly bind to and regulate a subset of promoters in N2A cells.

5.3 Discussion

As previously discussed, the paucity of studies reporting direct transcriptional targets of nuclear localised lncRNAs acting *in trans* means that the mechanisms by which these transcripts target and regulate expression at distal genomic locations are poorly understood. The work presented here adds to the small number of studies that have mapped the genome-wide binding profiles of chromatin-associated lncRNAs and identified their direct regulatory targets (Chu et al., 2011; Ng et al., 2013; Simon et al., 2011; Vance et al., 2014). Fortunately, expanding access to next-generation sequencing facilities together with the emergence and development of experimental and computational tools for studying RNA-DNA and RNA-protein interactions means that the number of such studies is likely to grow

dramatically over the coming years. Drawing on the data from these will help understand not only the scope and scale, but also the mechanistic details of lncRNA-mediated gene-expression regulation.

Theoretically, lncRNAs can associate with chromatin via the following mechanisms: 1) direct complementary base-pairing between RNA and DNA (or nascent RNA) sequences, 2) non-canonical RNA-DNA structures, such as RNA-DNA:DNA triplexes, whose formation may be governed by Hoogsteen base-pairing rules, 3) non-base pairing mediated RNA-DNA interactions, and 4) indirect interactions mediated through proteins or protein complexes. It is presently unknown how important direct complementary RNA-DNA (or RNA-RNA) interactions may be for targeting lncRNAs in the nucleus. Likewise, only individual examples are known where a lncRNA is targeted via RNA-DNA:DNA triplex formation. For example, pRNA (promoter associated RNA), a low abundance RNA upstream of the pre-rRNA TSS, appears to be targeted to a functional element within the rDNA promoter via such a mechanism (Schmitz et al., 2010). Hypothetically, direct RNA-DNA interactions could efficiently and sequence-specifically recruit lncRNAs to the genome. However, such interactions may also expose the genome to deamination and damage (Aguilera and Garcia-Muse, 2012; Buske et al., 2012). Also, RNA three-dimensional structure may create DNA-binding pocket analogous to the DNA-binding domains of DNA-binding proteins. For example, reminiscent of protein TFs, a DNA sequence motif was found to be enriched in the binding sites of *HOTAIR* lncRNA (Chu et al., 2011). However, a lack of large numbers of observations means that it is unclear how prevalent this hypothetical mode of lncRNA recruitment may be.

On the other hand, examples abound of lncRNAs that associate with and may be targeted to chromatin through proteins. DNA sequence-specific transcription factors constitute one group of such proteins. For example, YY1, a zinc finger-containing transcription factor, is thought to provide a bridge between *Xist* and chromatin via its distinct RNA- and DNA-binding domains (Jeon and Lee, 2011). Likewise, the androgen receptor protein (AR) recruits the *Prncr1* lncRNA to Androgen Response Elements (AREs) genome-wide, while PAX6 recruits *Paupar* to its targets in neuronal cells (Vance et al., 2014; Yang et al., 2013). For both *Prncr1* and *Paupar*, depletion of the lncRNA did not disrupt protein occupancy, which speaks in favour of the lncRNA being recruited by the protein, and not the other way round. Among other examples of RNA-TF interactions are *Lethe* and NF-κB (Rapicavoli et al., 2013), *Gas5* and glucocorticoid receptor (Kino et al., 2010), *Jpx* and CTCF (Sun et al., 2013), *Panda* and NF-YA (Hung et al., 2011), and *Rmst* and SOX2 (Ng et al., 2013). However, unlike *Paupar* and *Prncr1*, *Lethe*, *Gas5*, *Jpx* and *Panda* each inhibit DNA association of their TF partner at a number of loci tested, while *Rmst*, on the contrary, facilitates the functional association of SOX2 with promoters of its neurogenic target genes. Together, these studies show that lncRNAs can modulate the DNA binding capacity of their TF partners. It is expected that lncRNAs may also act as non-DNA binding cofactors of the associated TFs, as has been reported for *Prncr1* which strongly enhances both ligand-dependent and ligand-independent AR-mediated gene activation (Yang et al., 2013).

Histone modifications and chromatin structure may also be involved in some aspects of lncRNA-chromatin interactions, even if not in site-specific recruitment *per se*. For example, AR-associated lncRNAs *Pcgem1* and *Prncr1* have been shown to preferentially

associate with enhancer chromatin marks *in vitro* (Yang et al., 2013). Similarly, domains of initial *Xist* binding are highly enriched in active chromatin marks, suggesting that histone modifications may somehow facilitate *Xist* recruitment (Simon et al., 2013). *Dali* has been found to bind to regions associated with active chromatin marks.

In this Chapter, I described the identification of direct transcriptional targets of *Dali* by integrating its genome-wide occupancy data with previously described sets of putative target genes. I observed that 8.7% (162) of all genes whose expression altered following *Dali* depletion were associated with *Dali* binding sites within 1 Mb of surrounding sequence (although 30% of peaks resided within 5 kb, and 70% within 50 kb of a TSS). Therefore, these genes are likely directly regulated by *Dali*. This result is similar to the proportion of functional sites reported by others for transcription factors (Cusanovich et al., 2014). Computational analysis of the *Dali* genome-wide binding profile found no evidence for this lncRNA being recruited to chromatin via direct base-pairing interactions, both canonical complementary and non-canonical triplex-mediated. Direct non base-pairing mediated interactions similar to what has been proposed for *HOTAIR* (Chu et al., 2011) were not experimentally assessed, but no single sequence motif not corresponding to known TF recognition sites was discovered under the peaks.

Supporting the conclusion that *Dali* is likely recruited to the genome by chromatin-associated proteins, I identified some of the protein interactors of *Dali*. Because the experimental techniques for studying RNA-protein interactions available at the time are known to be prone to false positives (Chu et al., 2011; Hogg and Collins, 2007; Lee et al., 2013), I took a two-stage approach to identifying proteins associated with *Dali*. Firstly, I used a naïve proteome-wide approach to obtain a list of putative *Dali*-interacting partners,

which was then refined using the UV-RIP technique that is used for identifying direct molecular interactions with the endogenous *Dali* transcript. I showed that both mouse and human *Dali* interacts with the DNMT1 DNA methyltransferase, suggesting that the function of *Dali* may be conserved between the two species. In future, it would be extremely interesting to study if DNMT1 may interact with other intergenic lncRNAs genome-wide.

Also, among the direct regulatory targets of *Dali* I observed genes also regulated by *Pou3f*3. Computational analysis of *Dali* binding motifs suggested that some of these loci may also be bound by POU3F3. I confirmed these observations by showing that *Dali* directly interacts with the POU3F3 protein encoded by its neighbouring gene and together they bind and regulate a set of common targets. Therefore, *Dali* provides yet another example of a functional lncRNA-TF interaction. As such, studies of genomic occupancy of *Dali* and POU3F3 in the absence of the other factor would be an interesting extension of the present work.

Importantly, *Dali* provides the second example of an intergenic lncRNA in the vicinity of a neuronal transcription factor that both regulates the transcription and directly interacts with the protein product of this gene. The first example of this mode of action was provided by *Paupar*, which is transcribed in the vicinity and regulates transcription of *Pax6* (Vance et al., 2014). Similar to *Dali*, *Paupar* also binds to the product of the adjacent gene (PAX6) and together they bind and regulate a set of genes *in trans*. Of note, both *Dali* and *Paupar* were derived from a set of loci identified by Ponjavic et al. (Ponjavic et al., 2009) as CNS-expressed evolutionarily constrained transcripts. Both *Dali* and *Paupar* are spatially and temporally co-expressed with their neighbouring transcription factor genes. These

observations, together with the preferential genomic location of intergenic lncRNA loci adjacent to transcription factor genes (Ponjavic et al., 2009) imply that lncRNAs may commonly regulate the transcription and interact with the product of genomically adjacent transcription factor genes to act *in trans* on distal genes. Investigating this link further could prove fruitful for future research.

In Chapter 4, I showed that *Dali* regulates genes involved in neural development and function and its depletion disrupts terminal stages of neuronal differentiation, more particularly neurite outgrowth development. Here, this observation is supported by finding that *Dali* RNA binds to and regulates the promoters or promoter-proximal regions of important regulators of neuronal differentiation. Among these are *E2f2*, *Fam5b*, *Sparc* and *Dkk1*, all of which promote neuronal differentiation and are positively regulated by *Dali*. *E2f2* has a role in maintaining the postmitotic state of terminally differentiated neurons and is inducing apoptosis of those attempting to re-enter the cell cycle (Persengiev et al., 2001). *Fam5b* negatively regulates neuronal proliferation (Terashima et al., 2010). *Sparc* inhibits cell cycle progression and induces neuronal marker expression through Notch-mediated STAT3 signalling (Bhoopathi et al., 2011). *Dkk1* is an inhibitor of Wnt signaling and both enhances neuronal differentiation of mouse ES cells (Cajanek et al., 2009) and increases generation of early neuroectoderm (Watanabe et al., 2005). *Dali* also binds to and negatively regulated *Kif2c* and *Kif11* which are each known to block neurite outgrowth during differentiation (Laketa et al., 2007; Myers and Baas, 2007; Nadar et al., 2012). Therefore, *Dali* may work as a pro-differentiation factor in neural development by regulating the balance between proliferation and differentiation, as well as processes associated with terminal neuronal differentiation, such as neurite outgrowth.

In conclusion, my findings are consistent with a model in which mouse or human *Dali* is recruited to chromatin indirectly via RNA-protein interactions with both sequence-specific transcription factor proteins, such as POU3F3, or non-sequence specific DNA binding cofactors including DNMT1, which in turn may interact with sequence-specific DNA-binding proteins, to regulate genes involved in neuronal differentiation and function. In this model, *Pou3f3*-dependent target genes are regulated by *Dali* both indirectly, via its transcriptional regulatory effect on the *Pou3f3* gene, and directly via its physical interaction with the POU3F3 protein and their co-occupancy at regulatory regions of target genes.

Chapter 6 *Dali* regulates gene expression by binding to chromatin *in trans* and modulating DNA methylation status of its target CGI-associated promoters

6.1 Introduction

DNA methylation has long been recognised as one of the key epigenetic marks for mammalian development. It has been shown to play important regulatory roles in transcription, imprinting, and genomic defence against mobile elements (Karimi et al., 2011; Mohn et al., 2008; Smallwood and Kelsey, 2012). In mammals, methylation of cytosine on the fifth position occurs predominantly in the CpG context (Ziller et al., 2011). DNA methylation is catalysed by three conserved members of the DNA methyltransferase (DNMT) family of enzymes, all of which are essential for normal development (Li et al., 1992; Okano et al., 1999). DNMTs are traditionally classified as maintenance (DNMT1) and *de novo* (DNMT3A and DNMT3B).

Globally, mammalian genomes are depleted of CpG dinucleotides (approximately 1% as opposed to 4.41% expected by chance in the genome with GC content of 42%) which instead form CG-rich clusters referred to as CpG islands (CGIs); these contain approximately 10% of genomic CpGs (Smith and Meissner, 2013). CGIs are usually located in proximity of transcription start sites (TSSs). Approximately 70% of annotated gene promoters are associated with a CpG island (Long et al., 2013; Saxonov et al., 2006), including nearly all housekeeping genes, as well as some tissue-specific and developmentally regulated genes (Zhu et al., 2008). In fact, it has been suggested that all CGIs may be associated with transcripts, as those CGIs currently not associated with

annotated TSSs have been also shown to function as promoters (Illingworth et al., 2010; Maunakea et al., 2010). Between 60% and 80% of CpG dinucleotides elsewhere in the genome are methylated, while CGIs remain predominantly unmethylated in somatic cells (Smith and Meissner, 2013).

Vertebrate CGIs are short (1000 bp on average) stretches of genomic DNA characterized by a higher than genomic average G+C base composition, high CpG dinucleotide content, and lack of cytosine methylation (Gardiner-Garden and Frommer, 1987). CGIs are enriched in binding sites for ubiquitous transcription factors and predominantly exist in an open chromatin conformation (Gardiner-Garden and Frommer, 1987). It is thought that most, if not all, CGIs mark transcription initiation sites, including those of yet unannotated transcripts (Deaton and Bird, 2011). CGIs are frequently associated with promoters or first exons of ubiquitously expressed genes, but can also be found at tightly regulated developmental genes. In all cases, they are usually hypomethylated, whether the gene is active or not (Bird, 1987; Fouse et al., 2008; Gal-Yam et al., 2008; Rollins et al., 2006; Saxonov et al., 2006; Weber and Schubeler, 2007). Methylation of CGIs was shown to cause stable gene silencing (reviewed in (Blackledge and Klose, 2011)).

CGI-associated genes can undergo abnormal silencing, which remains poorly understood mechanistically, but involves both the closing of chromatin structure of the CGI, as well as hypermethylation (Gal-Yam et al., 2008). It has also been proposed that CGI methylation may play a role in regulating genes during normal development (Illingworth et al., 2008; Mohn et al., 2008; Payer and Lee, 2008; Stein et al., 1982). For example, some CGI-associated promoters acquire methylation during neuronal differentiation of ES cells, especially during the early stages of differentiation (Mohn et al., 2008). However, the

majority of genes associated with such CGIs are already repressed in ES cells (Mohn et al., 2008) supporting the idea that methylation may rather lock in the silent state than initiate it (Deaton and Bird, 2011).

DNA methylation may influence transcription via one of the following mechanisms: 1) disruption of methylation-sensitive transcription factor binding; 2) direct chromatin structure change; or 3) binding of methyl-binding domain (MBD) proteins that then recruit chromatin-modifying factors to the region (Bogdanovic and Veenstra, 2009). Many transcriptionally active promoters contain low levels of DNA methylation (4-7%) suggesting that DNA methylation-mediated gene repression may be density-dependent (Weber and Schubeler, 2007). However, the opposite was shown for the cAMP-responsive element (CRE)-binding sites found in the promoters of numerous tissue-specific genes (Roesler et al., 1988). Here, methylation of the CpG dinucleotide in the CRE recognition sequence was demonstrated to inhibit transcription factor binding and, consequently, transcription. Therefore, methylation of specific individual CpGs can be important for regulating at least some promoters (Iguchi-Ariga and Schaffner, 1989).

Although generally associated with gene silencing, it has been reported that CGI methylation may also contribute to gene activation. It appears that the position of a CGI within a gene affects the transcriptional response of that gene to the CGI methylation. Promoter CGI methylation was found to be associated with transcriptional silencing, while 3′ CGI methylation, on the contrary, resulted in transcriptional activation (Yu et al., 2013). It was also observed that intragenic CGIs are more prone to acquiring methylation than the promoter-associated ones (approximately 20-34%) (Deaton and Bird, 2011), which may also indicate their regulatory importance.

At present, most of the global genomic methylation is considered relatively stable in different tissues and throughout life, being altered only occasionally in specialized contexts. Two important exceptions are during germ cell and pre-implantation development, where fertilization triggers rapid de-methylation of the paternal genome followed by both parental genomes being hypomethylated during early embryogenesis (Seisenberger et al., 2013). Despite DNA methylation being one of the most extensively studied epigenetic signatures, how it is altered globally and locally, and what role these alterations play, remain poorly understood.

DNMT1 in particular has the unique ability to recognise the hemimethylated strand of newly replicated DNA. This property allows for the epigenetic inheritance of DNA methylation patterns during cell division (Hermann et al., 2004). However, it does not explain how DNA methylation patterns are changed, particularly in development and disease.

Traditionally, DNMT1 was thought to propagate pre-existing DNA methylation patterns in a replication-dependent manner (Hervouet et al., 2010; Liu et al., 2013). However, accumulating data suggest that DNA methylation inheritance may also occur via targeted mechanisms. Targeting of DNMT1 to specific loci is believed to be facilitated by DNMT1-interacting transcription factors (Esteve et al., 2005, 2007; Hervouet et al., 2012; Hervouet et al., 2010; Liu et al., 2013; Robertson et al., 2000; Zhang et al., 2005). The fact that 58 transcriptional factors have been reported as potentially interacting with DNMT1 implies that transcription-factor targeted DNA methylation may play a substantial role in genome-wide DNA methylation (Hervouet et al., 2010).

Recently, lncRNAs have been implicated in regulating the DNA methylation at their own site of synthesis. Specifically, *ecCEBPA* lncRNA transcribed from the *CEBPA* locus was found to interact with DNMT1 and prevent it from methylating DNA locally (Di Ruscio et al., 2013). Interestingly, deposition of DNA methylation upon *ecCEBPA* depletion immediately resulted in the down-regulation of *CEBPA* mRNA. Using high-throughput approaches, the same authors identified other loci genome-wide that may be subject to analogous regulation of local DNA methylation by nascent transcription (Di Ruscio et al., 2013).

In this Chapter, I explored the link between the genome-wide binding of *Dali* at or in proximity of CpG islands and the interaction of both mouse and human *Dali*, with the DNMT1 DNA methyltransferase. Using targeted DNA methylation studies, I showed that at three of the *Dali*-bound CGI-associated promoters depletion of *Dali* is accompanied by increased DNA methylation and by altered transcriptional response to differentiation-inducing retinoic acid treatment. Therefore, *Dali* appears to inhibit DNA methylation at a subset of bound and regulated regions, presumably deposited by the DNMT1 DNA methyltransferase, to which it binds. Importantly, unlike all previously reported lncRNAs involved in DNA methylation regulation, *Dali* appears to work *in trans* to modulate DNA methylation patterns away from its own site of synthesis. Also, enrichments of binding motifs for transcriptional factors that were reported by others as interacting with DNMT1 (Hervouet et al., 2010) were observed from the *Dali* CHART-Seq data. This observation may indicate that *Dali* regulates targeted but not the non-specific replication-dependent arm of DNA methylation maintenance.

6.2 Results

6.2.1 *Dali* binds in proximity to CGIs, including those associated with *Dali*-regulated promoters

Previously, analysis of the *Dali* CHART-Seq data revealed that CGI annotations were a prominent feature of the *Dali* genome-wide binding profile (Figure 5.5 C). *Dali* peaks overlapped CGI annotations approximately 10 times more often than expected by chance. Overall, 1727 out of 1989 genes associated with 1427 CHART-Seq peaks were also associated with a CGI (Figure 6.1). 130 of these 1727 genes were also previously found to be regulated by *Dali* (Chapter 5.2.3).

Considering the role of CGIs in gene regulation, as well as the fact that *Dali* was found to physically interact with DNMT1 in both mouse and human (Chapter 5.2.6; Figure 5.9 C), the enrichment of CGIs in *Dali* CHART-Seq peaks suggests that DNA methylation may play a role in transcriptional regulation of some of the *Dali* targets.

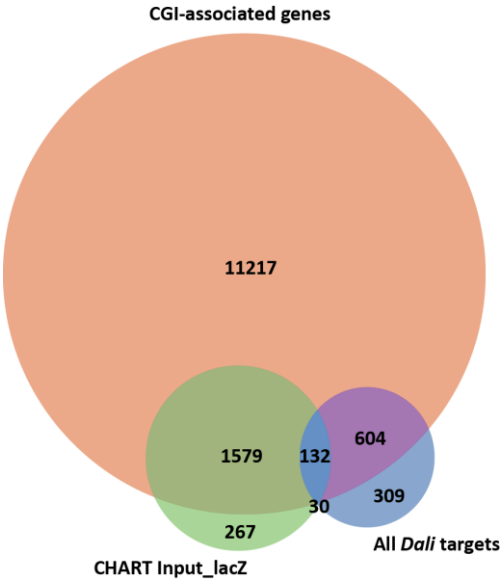


Figure 6.1 *Dali* binds and regulates CGI-associaed genes. The intersection of genes associated with CGIs (- 5kb/+1 kb), genes proximal (<1Mb) to *Dali* peaks and genes regulated by *Dali* identifies a set of direct targets of *Dali* (bound and regulated by it) also associated with a CGI.

6.2.2 Depletion of *Dali* leads to DNA methylation changes at bound and regulated promoters

The direct interaction of *Dali* with DNMT1 suggested that it may regulate DNMT1-mediated methylation at CGI-associated promoters of *Dali*-bound and -regulated genes. To test this hypothesis, I chose to use the Combined Bisulfite Restriction Analysis (COBRA; Chapter 2.2.2.21) assay (Xiong and Laird, 1997). This assay relies on bisulfite conversion of DNA template followed by PCR amplification and restriction digestion of resulting products to semi-quantitatively identify DNA methylation differences in a region of interest. Restriction endonucleases used in the COBRA assay have a recognition sequence that contains a CG dinucleotide. Therefore, the DNA methylation status of a particular cytosine residue will determine if the recognition sequence is preserved after bisulfite treatment (methylated residue) or disappears when an unmethylated cytosine is converted to thymine. After restriction digestion, fragments are separated by agarose electrophoresis and differences between experimental and control DNA samples are assessed visually (or quantified, if required). A control restriction digestion is performed using restriction endonucleases with recognition sequences containing a cytosine in a non-CG context to ensure complete bisulfite conversion of the template. In this case, all recognition sequences should be ablated by bisulfite treatment, and no significant restriction digestion should be observed (Chapter 2.2.2.21).

I chose to perform COBRA of *Dali*-bound and regulated CGI-associated promoters in stable *Dali* knockdown and stable non-targeting control N2A cells of an early passage. I reasoned that the usual 72 h transient transfection-mediated *Dali* knockdown may not be sufficient to produce DNA methylation changes detectable by COBRA. This is because

there is a delay between transfection and when efficient *Dali* knockdown is achieved caused by cellular uptake and expression of the transfected shRNA construct. Therefore, cells transfected in this way will only undergo 1-2 replicative cycles after *Dali* is depleted. This might not be sufficient to cause a DNA methylation maintenance defect, as might be the case if *Dali* regulates DNA methylation of the chosen regions via DNMT1.

I performed COBRA in parallel at 10 representative CGIs bound and regulated by *Dali* (all primers are listed in Table 2.8). I amplified regions lying fully within the boundaries of computationally annotated CGIs to avoid the potential detection of changes that are believed to occur at CGI "shores" and whose regulatory significance remain unclear (Sproul and Meehan, 2013). For five of the chosen regions (corresponding to four genes), restriction profiles of control and stable *Dali* knockdown cells were different, indicating an altered DNA methylation status upon *Dali* depletion (Figure 6.2). I also included samples from RA-differentiated cells of the same stable *Dali* knockdown and control lines as well as transient *Dali* knockdown samples. The profiles of the RA-treated samples were identical to those of their matched undifferentiated samples, i.e. they reflected the *Dali* status of cells, as opposed to their differentiation status.

Of note, while a positive COBRA result is likely to reflect the underlying DNA methylation differences, a negative result in most cases should be interpreted as inconclusive. This is because the assay relies on the availability of both efficient PCR primers to specifically amplify regions of interest from low-complexity bisulfite-converted samples, as well as recognition sequences for the COBRA-compatible and control restriction endonucleases within regions of interest. Therefore, the sensitivity of the assay is limited by the density of endonuclease recognition sequences within these regions.

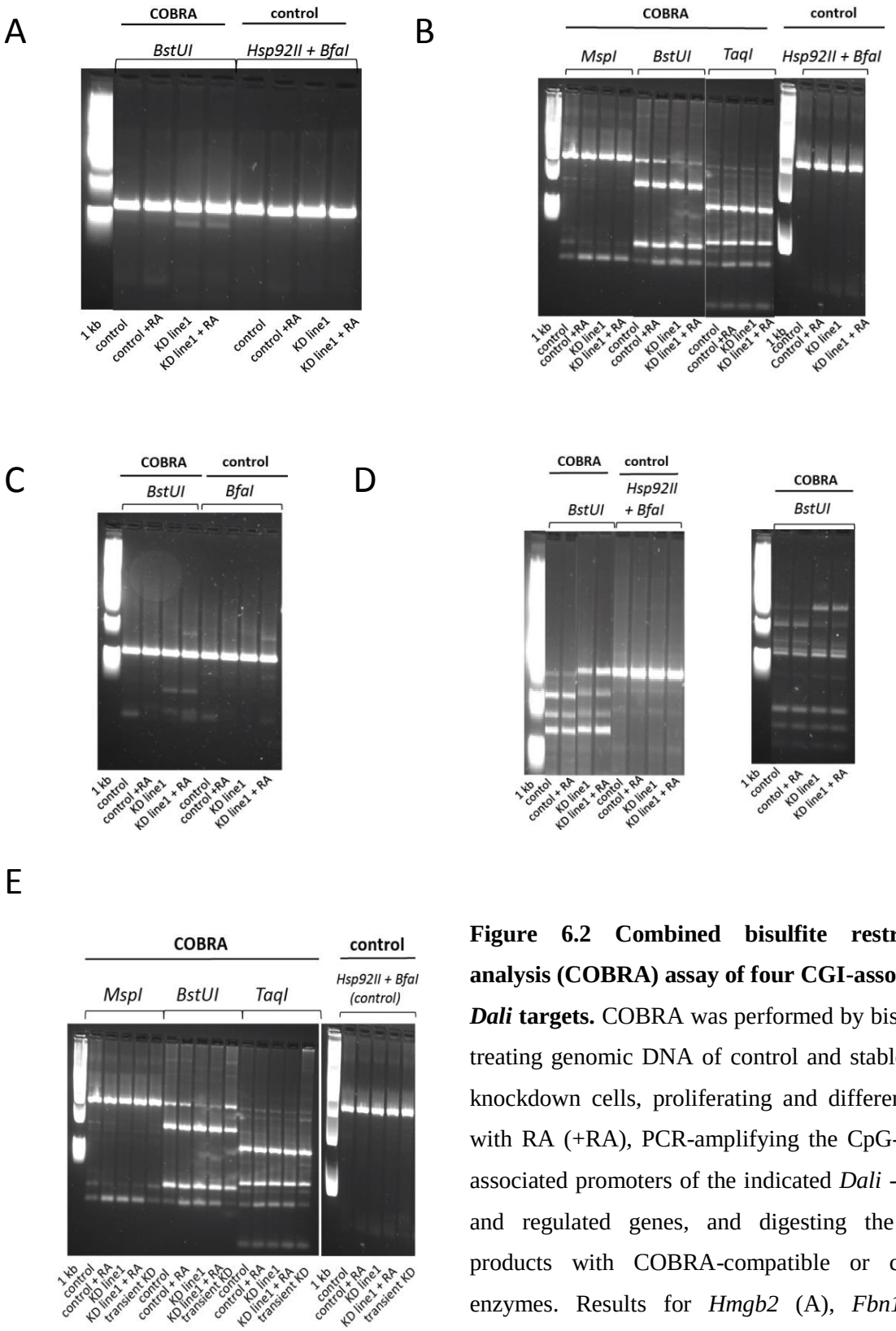


Figure 6.2 Combined bisulfite restriction analysis (COBRA) assay of four CGI-associated *Dali* targets. COBRA was performed by bisulfite-treating genomic DNA of control and stable *Dali* knockdown cells, proliferating and differentiated with RA (+RA), PCR-amplifying the CpG-island associated promoters of the indicated *Dali* -bound and regulated genes, and digesting the PCR products with COBRA-compatible or control enzymes. Results for *Hmgb2* (A), *Fbn1* (B), *Dlgap5* (C), *Nos1* (D) are shown.

To confirm the COBRA results, I performed bisulfite sequencing of the four COBRA-positive regions (associated with four *Dali*-bound and regulated genes) in two independent stable *Dali* knockdown cell lines bearing two different *Dali*-targeting shRNA constructs (including one cell line previously used for COBRA) and a stable control line (Chapter 2.2.2.15 and 2.2.2.20). The results of sequencing demonstrated that the *Dlgap5*, *Hmgb2*, and *Nos1* promoters each display increased CpG methylation in two independent stable *Dali* knockdown lines compared to control (Figure 6.3). The *Fbn1* promoter-associated CGI, despite exhibiting an altered restriction profile in the COBRA assay (Figure 6.2 E), was found to be almost entirely methylated (93.7%) in control cells, becoming fully methylated (98.1%) in one of the knockdown lines (sequencing in the other line was judged unnecessary and was not performed; Figure 6.3 G, H). Therefore, upon *Dali* depletion, *Dali*-bound and regulated CGI-associated promoters exhibit increased DNA methylation. Importantly, these changes occur within the CGIs themselves, and not on their "shores", which may indicate methylation of regulatory importance. Together, these results are consistent with the hypothesis that *Dali* acts *in trans* to inhibit CpG methylation at a subset of bound and regulated promoters.

Hypothetically, these changes may be mediated by DNMT1, which was previously shown to interact with lncRNAs *in cis* resulting in impaired methylation activity at the lncRNA-associated loci (Di Ruscio et al., 2013). Unfortunately, experimentally manipulating DNMT1 levels to investigate if it is involved in *Dali*-mediated regulation poses a methodological challenge, as *in vivo* increased expression of DNMT1 is lethal in mouse (Biniszkiewicz et al., 2002), and its loss results in neuronal death (Hutnick et al., 2009). Cells in culture with experimentally altered DNMT1 levels also manifest severe defects,

including cell death, grossly impaired differentiation and self-renewal (Fan et al., 2005; Lei et al., 1996; Panning and Jaenisch, 1996; Trowbridge et al., 2009).

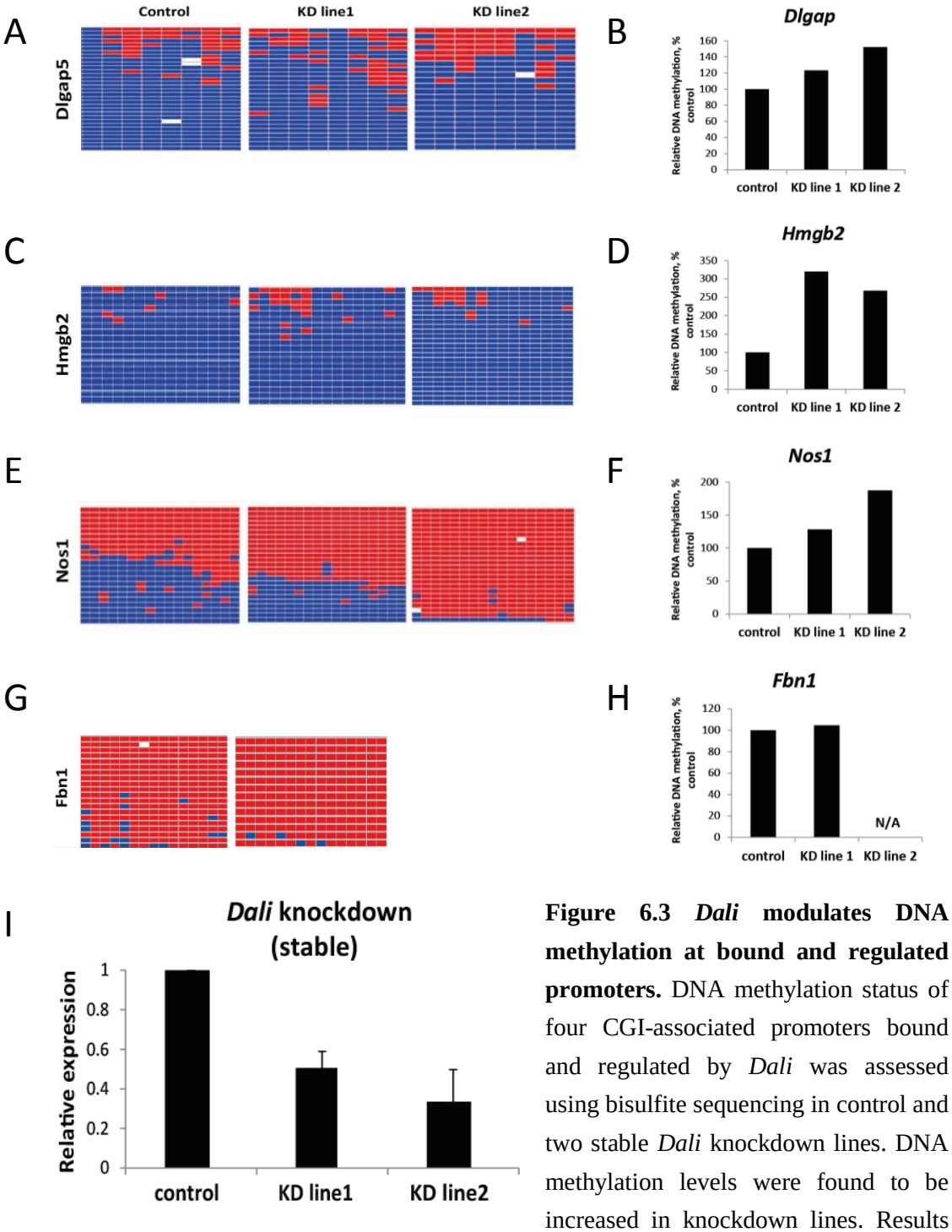


Figure 6.3 *Dali* modulates DNA methylation at bound and regulated promoters. DNA methylation status of four CGI-associated promoters bound and regulated by *Dali* was assessed using bisulfite sequencing in control and two stable *Dali* knockdown lines. DNA methylation levels were found to be increased in knockdown lines. Results for *Dlgap5* (A, B), *Hmgb2* (C, D), *Nos1* (E, F), and *Fbn1* (G, H) are shown. The degree of increase was correlated with the degree of *Dali* knockdown (panel I).

With regards to the expression of genes exhibiting DNA methylation changes upon *Dali* depletion, their steady-state levels, as well as response to RA, vary. *Nos1* is down-regulated following RA treatment in control, but not in stable knockdown cells. However, *Nos1* levels in undifferentiated knockdown cells are already lower than in even RA-treated control cells. Therefore, in this case, increased DNA methylation coincides with down-regulation of the gene in proliferating cells, while normally it only occurs upon RA stimulus. *Fbn1* is also down-regulated following RA treatment in control, but not in stable knockdown cells. In this case, however, *Fbn1* expression levels in knockdown cells are completely unaffected by RA, leading to its expression being higher in RA-differentiated knockdown cells compared to control.

The other two genes, *Dlgap5* and *Hmgb2*, are both up-regulated upon RA differentiation in both knockdown and control, although the degree of up-regulation is greater in control cells. Consequently, expression levels of both genes are higher in control RA-differentiated cells. Therefore, in these two cases, DNA methylation acquisition associated with *Dali* depletion may prevent the differentiation-induced up-regulation of transcription.

These three expression patterns described above correspond to three distinct methylation profiles. *Nos1* displays intermediate DNA methylation levels in control cells which appear to increase upon the depletion of *Dali* (Figure 6.3 E, F). This, together with the corresponding repression of the gene, is in agreement with the previously proposed idea of DNA methylation-mediated gene repression being density-dependent (Weber et al., 2007). *Fbn1* promoter, on the other hand, is already unusually heavily methylated in control cells, and is completely methylated in knockdown cells (Figure 6.3 G, H). Moreover, although it is difficult to explain the behaviour of this gene using a classical “methylation equates to

repression" framework, DNA methylation here may somehow contribute to a locking in of the expression levels, thereby preventing further regulation. Finally, *Dlgap5* and *Hmgb2* promoters display relatively low levels of DNA methylation, which appear increased upon *Dali* knockdown (Figure 6.3 A-D). As these two genes fail to be up-regulated to the same degree in knockdown as in control cells, *Dlgap5* and *Hmgb2* may belong to a category of genes that are regulated by methylation of specific CpG dinucleotides that lie within transcription factor recognition elements, as previously described for CRE-elements, for example (Iguchi-Ariga and Schaffner, 1989; Roesler et al., 1988). In fact, a growing number of transcriptional regulators have been found to be sensitive to DNA methylation of their binding motif. Among these are CTCF (Wang et al., 2012; Zampieri et al., 2012), Sp1 (Macleod et al., 1994; Mancini et al., 1999), and YY1 (Kim et al., 2003), whose motifs where found enriched in *Dali*-bound sequences, as well as CREB (Mancini et al., 1999), n-Myc, C/EBPα (Chatterjee and Vinson, 2012; Kim et al., 2003; Rishi et al., 2010), and likely many others (Medvedeva et al., 2014).

In summary, DNA methylation at the regions studied here may play different roles in regulation of the associated genes. Taken together, the results suggest that DNA methylation changes occur in undifferentiated N2A cells upon stable *Dali* depletion. DNA methylation profiles of the regions studied are subsequently not affected by the RA-induced differentiation. As the promoters tested exhibit differential transcriptional response to RA in knockdown cells compared to controls, I hypothesize that their altered methylation status may be one of the factors contributing to this alteration.

6.2.3 Depletion of *Dali*, and its associated DNA methylation changes, correlate with reduced usage of a *Dali*-bound alternative promoter of *Nos1*

As discussed above, one of the genes regulated by *Dali* and which appears to exhibit DNA methylation changes upon its depletion is *Nos1*. *Nos1* encodes for a neuronal isozyme of nitric oxide synthase, which is an enzyme crucial for producing nitric oxide (NO). NO is a messenger molecule with diverse functions, but in brain and peripheral nervous system it acts as a neurotransmitter (Ogura et al., 1993). *Nos1* has multiple alternative promoters falling into two separate regions (referred to here as Exon 1 and Exon 2 for simplicity) whose differentiated use is proposed to fine-tune *Nos1* expression in response to various physiological and developmental stimuli, while giving rise to protein of the same amino acid sequence (Bros et al., 2006). Only the 5′-most region contains a CpG island and is bound by *Dali* (Figure 6.4 A). Therefore, I decided to test if *Dali* binding specifically affects the expression only of the bound promoter cluster. By measuring expression levels of the three 5′-most exons of *Nos1* in stable *Dali* knockdown and control lines, I observed that the expression of the *Dali*-bound Exon 1 was reduced relative to that of Exon 3, when *Dali* was depleted, whereas the expression ratio between Exons 2 and 3 was unaffected (Figure 6.4 B). These results indicate that *Dali* can function by specifically promoting the use of a distantly located (and more rarely used) alternative promoter to which it binds. This effect may potentially be mediated through the modulation of DNA methylation of the promoter-associated CpG islands.

In fact, regulation of alternative promoter usage by *Dali* may be a common feature of several of its targets. For example, the *AldoC* gene also has several distinct clusters of transcriptional start sites, only one of which is bound by *Dali* (Figure 6.4 C). However,

AldoC gene does not have CGIs associated with it. Overall, 267 of the genes bound by *Dali* (among which 30 have evidence for being regulated) are not associated with CGIs. Nevertheless, some of the *Dali*-bound promoters can be *de facto* CGIs (or more precisely, non-methylated islands (NMIs)), but are not annotated as such by computational prediction algorithms because of the relatively shorter length and/or lower CG dinucleotide frequency, as shown by Long and colleagues (Long et al., 2013). These promoters, therefore, are likely to be subject to similar regulatory mechanisms as conventional CGIs. Also, reports appear to suggest that DNA methylation of intragenic CGIs (which are not considered here) is more dynamic and tissue-specific than promoter-associated CGIs and plays an important role in gene expression (Deaton et al., 2011; Kulis et al., 2013; Maunakea et al., 2010). Furthermore, methylation at non-CGI promoters may also be involved in gene regulation (Gal-Yam et al., 2008), but was not considered here due to technical reasons when choosing target regions for DNA methylation studies.

In summary, in addition to regulating DNA methylation at some CGI promoters, *Dali* may regulate some of its target genes either by modulating methylation of non-CGI loci, or an alternative DNA methylation-independent mechanism(s). In order to assess the full extent to which *Dali* may contribute to DNA methylation regulation *in trans*, high-throughput DNA methylation studies (such as for example, reduced-representation or, ideally, whole genome bisulfite sequencing) would be necessary.

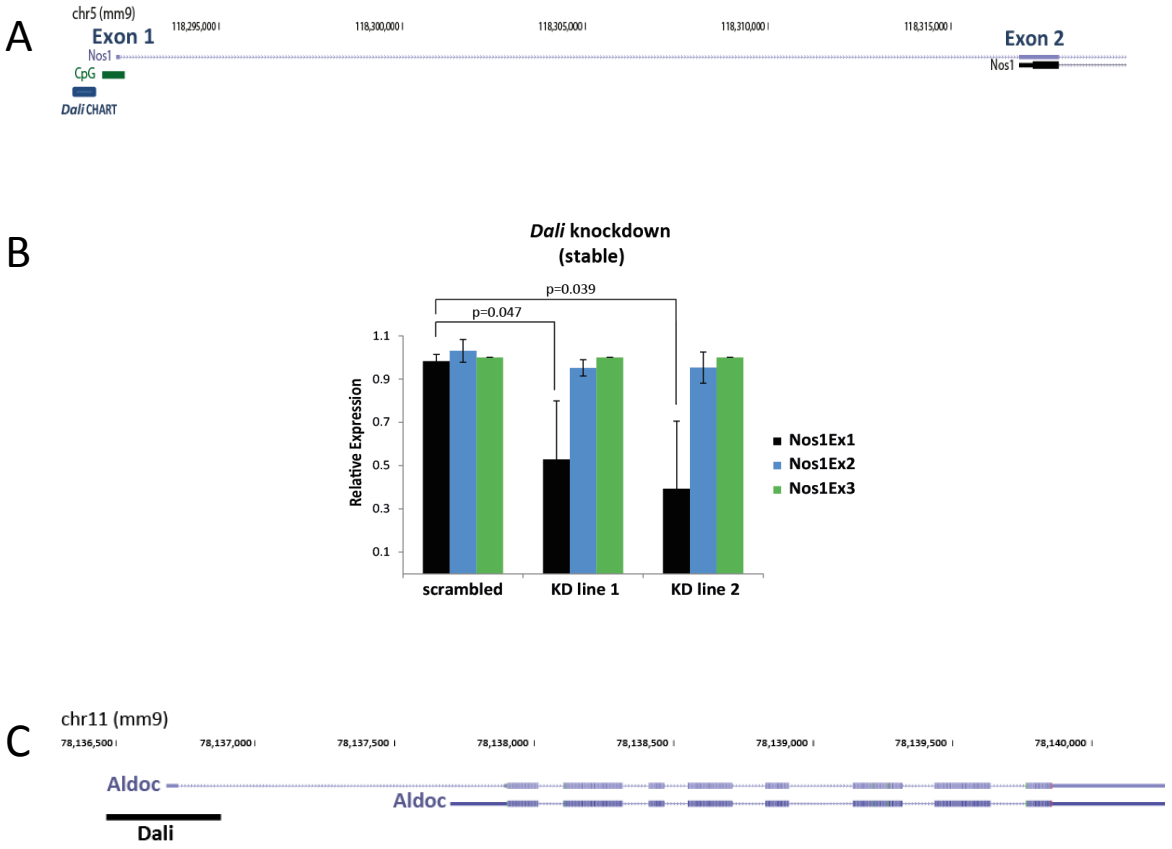


Figure 6.4 Depletion of *Dali* and associated DNA methylation changes correlate with reduced usage of the alternative *Nos1* promoter bound by it. (A) *Nos1* gene has two clusters of alternative TSSs (Exon 1 and Exon 2). The upstream neuronal tissue-specific cluster (Exon 1) is associated with a CpG island and is bound by *Dali*. (B) Down-regulation of *Nos1* observed in stable *Dali* knockdown lines can be explained by reduced initiation from the *Dali*-bound TSS (Exon 1), as the ratio between Exon1 and an internal Exon 3 is diminished, while the ration between Exon 2 and Exon 3 is not changed. Mean values +/- s.e, n = 3, one tailed t-Test, unequal variance. (C) *Dali* is bound at only one of the alternative promoters of *AldoC* regulated by it.

6.3 Discussion

Recent lncRNA functional studies have provided new insights into how epigenetic modifications are controlled. In particular, nuclear localised lncRNAs have been

demonstrated to interact with chromatin modifiers and transcription factors to regulate gene expression both locally and genome-wide.

Although non-coding RNAs have long been known to be extensively involved in regulating DNA methylation in plants (Matzke and Mosher, 2014), mammalian lncRNAs have been implicated in modulating DNA methylation only relatively recently. A growing body of literature suggest that all three catalytically active DNA methyltrasferases may be involved. As DNA methylation has been implicated in various aspects of gene expression regulation in both normal and pathologic states, revealing how lncRNAs contribute to this phenomenon is of great interest and practical importance.

Both *de novo* DNA methyltransferases, DNMT3A and DNMT3B, have been reported to associate with lncRNAs. For instance, the *CHD* lncRNA which is transcribed antisense to the *E-cadherin* promoter binds to the catalytic domain of DNMT3A to inhibit its activity (Holz-Schietinger and Reich, 2012). *ssEF1a*, a lncRNA antisense to the *EF1a* promoter, allosterically binds to DNMT3A outside of its catalytic domain and may play a scaffolding role for DNMT3A-associated proteins (Holz-Schietinger and Reich, 2012). The *Tsix* RNA binds and activates DNMT3A to methylate DNA at the promoter of *Xist*, the lncRNA crucial for X-chromosome inactivation (Sun et al., 2006; Sun et al., 2013). DNMT3B interacts with lncRNAs that are complementary to the rDNA promoter and that form a DNA:RNA triplex which is recognised by DNMT3B whose activity helps silence rRNA genes (Schmitz et al., 2010).

Conventionally referred to as a maintenance methyltransferase, DNMT1 which I found to be interacting with both mouse and human *Dali* orthologs was also previously reported to

be associated with lncRNAs. Among these are *Kcnq1ot1* and *ecCEBPA*, which both have been implicated in modifying CpG methylation in a DNMT1-dependent manner. *Kcnq1ot1* is expressed from the imprinted *Kcnq1* domain and maintains the silencing of ubiquitously imprinted genes in this locus in part through its interaction with DNMT1 (Mohammad et al., 2010). Interestingly, *Kcnq1ot1* was found to be highly unstable, similar to *Dali* (Clark et al., 2012). *ecCEBPA* originates from the *CEBPA* locus and binds to DNMT1 in a manner that prevents the subsequent deposition of CpG methylation marks at the *CEBPA* promoter. Similar to what was also observed for *Dali* and its targets, depletion of *ecCEBPA* results in increased DNA methylation of the *CEBPA* promoter (Di Ruscio et al., 2013). The same authors also reported a set of lncRNAs genome-wide that may act via analogous *cis*-acting mechanisms in human promyelocytic leukemia HL-60 cells (Di Ruscio et al., 2013).

In this and previous chapters, a combination of computational and experimental approaches were used to predict how *Dali* might interact with chromatin. Computational analysis of the genome-wide binding profile revealed enrichment of *Dali* occupancy sites at transcriptionally active promoters. Particularly prominent was the association of these sites with CpG islands. Together with the experimentally validated direct and evolutionarily conserved interaction of *Dali* with the DNA methyltransferase DNMT1, these observations suggested that *Dali* may be involved in modulating promoter activity via DNA methylation.

The data to date indicate that at three *Dali* bound and regulated promoters *in trans*, depletion of *Dali* levels may result in increased CpG methylation. This suggests that *Dali* may inhibit DNA methylation at a subset of bound and regulated regions, presumably

deposited by DNMT1 to which it binds in both mouse and human. DNMT1 was previously found to have higher affinity for structured RNA than to its DNA substrate (Di Ruscio et al., 2013). Hence, it is possible that *Dali* competes for binding to DNMT1 with either protein co-factors such as UHRF1, which loads and orients the enzyme on the DNA substrate (Inomata et al., 2008), or its DNA substrate.

Apart from the classic DNA replication-coupled maintenance of DNA methylation, targeted methylation by DNMT1 has been suggested to play an important role in genome-wide methylation. Targeting of DNMT1 to specific loci is believed to be mediated by transcription factors. 58 transcription factors were found to specifically interact with DNMT1 (but not to other DNMTs) *in vitro* (Hervouet et al., 2010). Of these, 9 have enriched sequence motifs in *Dali* CHART-Seq peaks. Hence, it is tempting to speculate that *Dali* may regulate TF-targeted, but not non-specific replication-coupled, DNA methylation. Conceptually, this could occur via *Dali* competing with the DNA substrate for DNMT1, inhibiting its enzymatic activity, or modulating its interaction with either transcription factors or other co-factors.

Studying genomic co-localisation of DNMT1 and transcription factors using ChIP remains challenging. This is primarily owing to the poor performance of the available anti-DNMT1 antibodies in this application. Also, an altogether non-focal distribution of most of the DNMT1 protein might be expected, as the enzyme is known to move along the replicated DNA together with the replication complex. However, the findings of this thesis may be substantiated and generalised by determining which other transcripts genome-wide also associate with DNMT1 and, therefore, may be involved in regulating its targeting or

activity. With advances in RNA-protein interaction techniques, such studies will soon be possible, and are of great interest.

The work presented in this thesis may also be extended by studying the *Dali*-DNMT1 interaction functionally *in vitro*. For example, of interest would be if *Dali* alone can inhibit DNA methylation of a suitable synthetic template by recombinant DNMT1. Or else, if *Dali* can modulate the interaction of DNMT1 and some of the transcription factors that were previously shown to interact with it, *in vitro* or *in vivo*, and whose recognition motifs were enriched in *Dali*-bound loci

In summary, *Dali* may be one of a growing subset of lncRNAs that bind to DNA methyltransferases and modify DNA methylation levels. Importantly, previous examples of lncRNAs that regulate DNA methylation act at their sites of synthesis (Mohammad et al., 2010) (Di Ruscio et al., 2013). Consequently *Dali* is the first reported intergenic lncRNA that may regulate DNA methylation *in trans*.

Chapter 7 General discussion

7.1 *Dali* is a functional lncRNA

Exonic regions responsible for all known protein-coding genes account for between 1.5 and 2% of mammalian genomes. The remaining 98% have until recently been largely unexplored and referred to as the "dark matter" of the genome. Previously assumed transcriptionally silent and largely non-functional ("junk"), the intergenic space has now been revealed to be pervasively transcribed. The latest estimates put the transcribed portion of the genome to be > 85% (ENCODE, 2012; Hangauer et al., 2013). Furthermore, genome- wide association studies uncovered that no less than half of all disease and trait-associated genomic loci fall into the intergenic space (Hindorff et al., 2009), defying the long-standing notion that proteins (and a relatively small number short non-coding RNAs) perform almost all of the essential biological functions. While some of these regions may function exclusively as DNA elements, we now know that thousands of putative intergenic long non-coding RNAs are produced from the intergenic regions, including an increasing number of transcripts with experimentally validated molecular and cellular functions.

The recent explosion of transcriptomic and other types of genome-wide data has fuelled the debate about the extent to which the intergenically-produced and often low-abundance transcripts are functional. Despite the uncertainty, it has been argued that many intergenic lncRNAs display features consistent with functionality, such as evolutionary conservation and active expression regulation. Moreover, the loci producing these transcripts are highly enriched for trait-associated variants, suggesting that many intergenic lncRNAs may

indeed confer biological function as opposed to being the product of transcriptional noise (Hangauer et al., 2013).

Present work expands our current understanding of biological functionality of intergenic lncRNAs by providing an in-depth experimental analysis of a single lncRNA termed *Dali*. *Dali* is a nuclear localised transcript originating from a locus 50 kb away from its nearest protein-coding neighbour. It plays gene regulatory roles both locally and distally in the genome, which expands the functional versatility previously attributed to individual lncRNAs.

7.1.1 *Dali* is a functional single-exon lncRNA involved in transcriptional regulation

Tens of thousands of putative intergenic lncRNAs have now been identified from mammalian genomes. However, it is almost certain that not all of them will prove functional. It is estimated that only 19% of human organ-specific lncRNAs have been conserved over > 90 My of evolution (Necsulea et al., 2014) and, if conservation is a good proxy for functionality, are likely to be biologically important.

Single exon lncRNA models, in particular, can be artefacts arising from genomic DNA contaminating sequencing libraries. Also, transcripts that are expressed at very low levels are generally considered less likely to confer function. As a group, single-exon intergenic lncRNAs are under-represented among transcripts that have been studied functionally. This is partly because they have been avoided due to concerns associated with the reliability of single-exon transcript models built from high-throughput transcriptomic data. Also, noisy

intergenic transcription is expected to produce largely single-exon products. At the same time, viable, fertile, and otherwise normal knockout mouse models of some of the highly and broadly expressed, and *bona fide* monoexonic intergenic lncRNAs, such as *Neat1* and *Malat1*/*Neat2*, also suggest that such transcripts may not be essential (Eissmann et al., 2012; Zhang et al., 2012). Hence, transcript sequences and levels are not reliable predictors of functionality.

Additionally, single-exon transcripts are potentially more challenging to study experimentally, due to issues associated with reliable detection and quantification. Furthermore, few of the single-exon lncRNAs studied in depth to date, except for *Malat1* and *Neat1*, are among the highest expressed lncRNA transcripts known (Hutchinson et al., 2007). Overall, mammalian single-exon lncRNAs have been found to be expressed at lower levels and are also less conserved evolutionarily compared to multi-exonic lncRNAs (W. Haerty, personal communication).

Dali is a 3.5-kb long evolutionarily constrained mono-exonic lncRNA. As such, it represents one of a small number of mono-exonic intergenic lncRNAs characterized in depth and assigned a cellular function. Therefore, even if only a small proportion of tens of thousands of mono-exonic intergenic lncRNAs that have been catalogued to date are functional, these transcripts represent a rich layer of presently unknown gene regulatory mechanisms.

7.1.2 *Dali* functions at a low copy number in N2A cells

Although individual examples of highly abundant lncRNAs, such as *Xist*, *Malat1* or *Neat1* (Noncoding nuclear-enriched *abundant* transcript 1) are known (Clemson et al., 2009; Clemson et al., 1996; Eissmann et al., 2012; Ji et al., 2003), taken in bulk, lncRNAs are on average expressed at lower levels than protein-coding genes. Generally, very low abundance transcripts are considered less likely to be functional. Hence, some cataloguing studies employed an arbitrary copy number cut-off when reporting lncRNAs (for example 1 copy per cell was used by (Hangauer et al., 2013)). There can be several reasons for the apparent low abundance of many lncRNAs. Firstly, many lncRNAs may indeed be non-functional products of noisy transcription initiating at low-frequency from random regions in the genome. Also, some bi-directionally expressed enhancers are known to produce non-polyadenylated short-lived and very low abundance lncRNAs referred to as enhancer RNAs, or eRNAs (De Santa et al., 2010; Hah et al., 2013; Kim et al., 2010). Additionally, the apparent low expression of lncRNAs may be the result of more narrow expression profiles which are known to be characteristic of lncRNAs (Derrien et al., 2012; Djebali et al., 2012; Guttman et al., 2009; Necsulea et al., 2014). In this case, if lncRNAs are expressed only by a certain cell type within a tissue, when studied in bulk tissue samples, expression levels of lncRNAs will appear artificially reduced. However, the results of a recent study which employed single cell RNA sequencing in human lymphoblastoid GM12878 cells corroborated the previous finding that lncRNAs as a group are expressed at lower levels than protein-coding genes (Marinov et al., 2014). Some investigators in the field also proposed that even in a seemingly homogeneous population of cells (such as cell lines in culture), there may exist individual "jackpot" cells that express certain lncRNAs at

much higher levels that the rest of the population, which to them serve as an indicator of highly specialised expression and function of these transcripts. Others, however, suggested that "jackpot" cells may instead indicate aberrant expression in just a few cells (Pennisi, 2014).

Despite the aspects discussed above, some of the lncRNAs shown to be functional in the literature are also expressed at very low levels in the systems in which their function was uncovered. For example, the *NeST* lncRNA that binds WDR5, a subunit of the MLL/SET1 histone H3 lysine 4 methyltransferase complex, and which up-regulates the *IFN-γ* locus in $CD3^+$ T cells by increasing the levels of H3K4 tri-methylation thereby modulating pathogen susceptibility in mice, was found to do so while being expressed at a copy number of only 0.15 molecules per cell (Gomez et al., 2013). *Hottip*, yet another WDR5-interacting lncRNA, which promotes the expression of several genes in the *HOXA* cluster through increased deposition of H3K4me3 marks, is expressed at a mere 0.3 copies per cell (Gomez et al., 2013; Wang et al., 2011b). *THRIL*, a lncRNA reported to bind to and regulate *TNFα* and potentially multiple other innate immunity-related genes *in trans*, was expressed at approximately 8 copies per cell in human macrophages (Li et al., 2014). Likewise, *Paupar* which binds to and regulates hundreds of genes in the genome of mouse neuronal cells has a copy number of 15 per cell in N2A cells (Vance et al., 2014). Divided by the number of potential targets *in trans*, the latter two lncRNAs may be acting at a very low copy number per target.

In the present work, I identified 162 genes as being directly bound and regulated by *Dali* in N2A cells where it is expressed at a population average of two copies per cell. Therefore, *Dali* provides an additional example of a lowly expressed transcript with important roles in

gene expression. Unfortunately, most of the functional studies of lncRNAs published to date either do not measure or do not report the cellular copy number. Consequently, it is currently unclear how abundant are these lncRNAs with assigned function. Also, only very few studies of lncRNAs that regulate target genes *in trans* have thus far mapped their genome-wide occupancy, which further hinders our understanding of the stoichiometry of the lncRNA function.

7.1.3 *Dali* acts in a RNA-dependent manner both locally and distally

Intergenic lncRNAs with functions in transcription may act either on genes situated relatively close to them in the genome or away from their loci elsewhere in the genome. The function itself may be conferred either by the mere act of transcription across the lncRNA locus (transcription-dependent mode of action) or by the resulting RNA molecule itself (RNA-dependent mode of action). In the former case, the transcript produced represents a non-functional by-product and does not function on its own. Attempts have been made to predict the likely mode of action of intergenic lncRNAs loci based on chromatin marks at their transcriptional start sites. Two categories of lncRNA loci were identified: those produced from loci with enhancer-like marks and those that look more like protein-coding promoters (Marques et al., 2013). Many, if not all, known enhancers are now known to be transcribed giving rise to typically exosome sensitive, mono-exonic RNAs referred to as enhancer RNAs (eRNAs). The levels of eRNAs were found to be positively correlated with those of adjacent protein-coding genes, and therefore, are likely to reflect enhancer activity (Andersson et al., 2014; Kim et al., 2010). However, some

enhancers produce more stable, polyadenylated and often multi-exonic transcripts, referred to as elncRNAs (also 1d-eRNAs, or ncRNA-activating lncRNAs (ncRNA-a)) (De Santa et al., 2010; Natoli and Andrau, 2012; Orom et al., 2010). All of the enhancer-associated lncRNAs may be transcribed in both directions with transcripts from either one or both directions being rapidly turned over (Andersson et al., 2014). It is currently unclear if mechanisms of enhancers that produce structurally different classes of transcripts are different or even how prevalent different categories are.

Furthermore, it remains an open question of whether enhancer-associated RNAs (of both eRNA or elncRNA types) are involved in enhancer function or are non-functional by-products of transcription-dependent mode of action. For a small but growing number of cases, enhancer function was shown to be RNA-dependent (Hah et al., 2013; Lam et al., 2013; Li et al., 2013b; Melo et al., 2013; Mousavi et al., 2013). In some of these cases, eRNAs were found to mediate enhancer-promoter looping interactions to regulate genes locally (Lai et al., 2013; Li et al., 2013b). Also, an eRNA produced by the *MyoD1* core enhancer was shown to modulate chromatin accessibility and PolII occupancy of the *MyoD1* promoter via an unknown mechanism (Mousavi et al., 2013). Together, these studies demonstrate that enhancer-derived RNAs can participate in enhancer function by modulating local chromatin structure or spatial conformation. On the other hand, p53-bound enhancers were shown to produce RNAs that augmented enhancer function via an unknown mechanism but had no effect on local chromatin state (Melo et al., 2013), while blocking the production of eRNAs from some estrogen receptor (ER) bound enhancers disrupted gene activation but not enhancer-promoter looping (Hah et al., 2013). Therefore,

results to date indicate that enhancer-associated lncRNAs may be involved in enhancer function via multiple different RNA-dependent mechanisms.

Likewise, examples of transcript-dependent but RNA-independent roles of lncRNA loci have been identified. The activity of the human growth hormone (*hGH*) HS1 enhancer is stimulated by transcription across a lncRNA locus immediately downstream of it, but non-overlapping with the *hGH* promoter itself (Yoo et al., 2012). The transcription-dependent and RNA-independent nature of this process was elegantly demonstrated in a series of genomic engineering experiments. First, when transcriptional terminators were inserted into the genome to prevent transcription from proceeding across the locus, a reduction of enhancer activity was observed. When the original genomic sequence downstream of *HS1* was replaced with an unrelated one, transcription across the locus was sufficient to recapitulate the original effect on the *HS1* enhancer activity. In summary, both transcription-dependent and RNA-dependent modes of action have been ascribed to transcribed enhancer elements.

Apart from acting at their site of synthesis to regulate chromatin and gene expression locally in an allele-specific manner, lncRNAs may also act on their neighbouring genes via *de facto in trans* mechanisms. For example, in female mammals *Jpx* transcribed from the active X chromosome can induce the expression of *Xist* which is genomically neighbouring but is located on the other copy of the X chromosome which is to be inactivated (Tian et al., 2010). *Evf-2* regulates expression of its neighbouring *Dlx6* gene by binding to DLX2 and MECP2 proteins and inhibiting DNA methylation of a local enhancer (Berghoff et al., 2013; Feng et al., 2006). The *trans*-acting mechanism of local regulation is supported by the findings that *Evf-2* and DLX2 can activate the same enhancer in a reporter assay when

provided on a transiently transfected plasmid, while ectopically expressed *in vivo Evf-2* can still inhibit DNA methylation. In summary, local gene expression and chromatin regulation by lncRNAs may be mediated via both *cis*- and *trans*-acting mechanisms.

Despite initial reports largely describing local regulatory effects of lncRNAs, a growing number of studies have demonstrated that lncRNAs may also regulate genes situated elsewhere in the genome. For example, the *NRIP1* eRNA appears to mediate long-range looping interactions between its own locus and *TFF1* 27 Mb away induced by E2 (Li et al., 2013b; Melo et al., 2013). Hence, some enhancer lncRNAs may possess previously unsuspected distal regulatory roles.

Trans-acting modes of lncRNA action have hitherto been less studied and remain very poorly understood. Recently developed techniques for mapping genome-wide occupancy of lncRNAs have revealed that a single transcript can associate with multiple genomic loci situated on different chromosomes. Such mapping studies have also shone new light on lncRNAs with previously identified roles. For example, *Hotair* transcribed from the *HOXC* cluster and acting on the genes in the *HOXD* locus situated *in trans* was shown to bind to approximately 800 other sites of up to a 1000 bp in length across multiple chromosomes (Chu et al., 2011; Rinn et al., 2007). These *Hotair*-bound loci are found within larger Polycomb-targeted regions and are enriched for genes regulated upon *Hotair* knockdown (Chu et al., 2011). The androgen receptor (AR) associated lncRNA *PCGEM1* binds to 2142 loci genome-wide, most of which are AR-bound enhancers, to promote AR-mediated gene activation (Yang et al., 2013). *Paupar* transcribed from an intergenic locus adjacent to *Pax6* binds to >2800 loci genome-wide including those associated with 242 genes differentially expressed upon its depletion in neuronal cells (Vance et al., 2014). Of note,

Paupar is transcribed from a conserved enhancer for *Pax6* and was shown to regulate the expression of *Pax6* itself, providing evidence that a single lncRNA may have both local and distal regulatory roles. Moreover, *Paupar* appears to bind the PAX6 protein to jointly regulate some of its distal targets (Vance et al., 2014). *Ctbp1-as* was also shown to possess both local and distal roles. Locally, it represses *Ctbp1* via a sense-antisense mechanism, while distally it promotes AR transcriptional activity in prostate cancer cells (Takayama et al., 2013).

In summary, a growing body of literature points to a conclusion that lncRNAs can be extensively involved in both local and distal regulatory mechanisms. As more of these studies appear, including those reporting directly bound and regulated genes, it will become possible to evaluate both the scope of such regulation and the detailed molecular mechanisms involved.

In this thesis, I showed that *Dali* acts in an RNA-dependent manner both locally to regulate transcription of the wider *Pou3f3* locus, including the protein-coding *Pou3f3* gene and other lncRNAs in its vicinity, and distally to regulate numerous genes genome-wide involved in neuronal differentiation and function. To my knowledge, *Dali* provides only a second example of an intergenic lncRNA possessing such mixed roles. Of note, both *Dali* and the previously reported *Paupar* do not only regulate their neighbouring transcription factors, but also interact with their protein products to regulate other genes *in trans*. Given this, as well as the fact that the loci of evolutionarily conserved lncRNAs in mammals preferentially located in the vicinity of neuronal transcription factors, it would be interesting to investigate if other co-expressed lncRNA-TF loci may recapitulate some or all of the features described for *Dali* and *Paupar*.

The current work also provides a list of genes that are directly bound and regulated by *Dali*, which has so far been done only for only a few lncRNAs. Therefore, together with the copy number calculations, this work is one of the first studies to provide insights into the stoichiometry of distal regulatory mechanisms of any lncRNA.

7.1.4 *Dali* is likely targeted to chromatin and acts as a part of a ribonucleoprotein complex

One of the most interesting and so far unresolved questions in the lncRNA field is how the chromatin-associated transcripts that work *in trans* are targeted to specific locations genome-wide. It is difficult to envisage how some transcripts, which act in this way and have been shown to be very low abundance and are unstable, find their targets in the nuclear space packed with non-target DNA. The recently proposed model of "proximity transfer" postulates that some lncRNAs may be guided to their distally located targets by the spatial folding of the genome itself. For example, *Xist* may spread from its own locus to genomically distal but spatially proximal loci along the X chromosome. However, the resolution of the currently available data does not allow us to conclusively accept or reject this model for *Xist* (Engreitz et al., 2013). On the other hand, support for this model is provided by observations that the *Hottip* locus is brought into proximity of the *HOXA* cluster regulated by the *Hottip* RNA by chromosome looping. Moreover, *Hottip* can only activate transcription from a reporter construct when tethered to its promoter, but not when expressed ectopically (Wang et al., 2011b). Also, the spatial organization of the genome in

nuclear space may facilitate simultaneous interactions of a lncRNA with multiple distal sites. For example, genomic loci bound by *Firre* were found to occur in spatial proximity within the nucleus (Hacisuleyman et al., 2014).

Those nuclear lncRNAs that can act upon their targets when expressed ectopically must be able to locate them in the genome without the aid of chromatin fold (Feng et al., 2006; Gomez et al., 2013; Tian et al., 2010; Vance et al., 2014). This can occur, for example, by the lncRNA being targeted to specific loci *in trans* as a part of a ribonucleoprotein complex. The previously discussed *Evf-2*, which binds to the DLX2 protein, can locate and activate the *Dlx-5/6* enhancer on a reporter construct when expressed from a transiently transfected plasmid (Feng et al., 2006). Likewise, depletion of *Paupar* was shown to modulate, in a dose-dependent manner, the transcriptional activity of DNA elements bound and regulated by it in the genome when these were inserted into a transiently transfected reporter construct (Vance et al., 2014). Another previously discussed lncRNA, *NeST*, is also capable of activating the *IFN-γ* locus by interacting with WDR5 and promoting H3K4me3 deposition when expressed from an exogenous construct (Gomez et al., 2013). In contradiction to the "proximity transfer" model, *Xist* was also shown to be able to act upon the endogenous X inactivation centre when expressed from a transgene in female mouse fibroblast cells, presumably being targeted there in a complex with the YY1 transcription factor which has both RNA and DNA-binding domains (Jeon and Lee, 2011).

In these cases where a lncRNA has been found to interact with a transcription factor, it is likely that the protein may provide sequence-specificity for the genomic binding of the complex. Apart from *Xist* and YY1, known lncRNA-TF complexes include: *Gas5* and GR (Kino et al., 2010), *Lethe* and NF-kB (Rapicavoli et al., 2013), *Jpx* and CTCF (Sun et al.,

2013), *Paupar* and PAX6 (Vance et al., 2014), *Prncr1* and AR (Yang et al., 2013), *Panda* and NF-YA (Hung et al., 2011), and *Rmst* and SOX2 (Ng et al., 2013).

Additionally, a lncRNA itself can provide sequence-specificity to the ribonucleoprotein complex of which it is part. For example, the pRNA regulating transcription of the rDNA cluster *in vivo* was shown to interact with its target promoter directly *in vitro* by forming a RNA-DNA:DNA triplex structure (Schmitz et al., 2010). Likewise, apart from the *HOXD* cluster, *Hotair* is targeted to approximately 800 other loci genome-wide that are enriched for a motif not matching any of the known transcription factor motifs, which therefore, may be recognised by the lncRNA itself (Chu et al., 2011). Furthermore, complementary RNA–RNA interactions may theoretically sequence-specifically guide lncRNAs to the transcribed loci, although the prevalence of this mechanism as well as direct lncRNA-DNA associations remains unknown.

In summary, genomic targeting of lncRNAs is the result of the interplay between the spatial organization of the genome, specific interactions of lncRNAs with transcription factors, chromatin remodellers and possibly other chromatin-associated factors, as well as potentially sequence recognition properties of lncRNAs themselves.

In this thesis, I showed that *Dali* is unlikely to directly recognise its DNA binding sites and instead may be reliant on its protein partners for genomic targeting. In support of this conclusion, no significant sequence complementarity was found between *Dali* and regions bound by it. Likewise, the currently available computational tools predicted that *Dali* does not form RNA-DNA:DNA triplexes with the vast majority of its binding sites. On the other hand, multiple transcription factor motifs were found enriched in the regions occupied by

Dali, including POU3F3 that was also shown to directly bind to *Dali* to regulate common targets. I also showed that local regulatory effects of *Dali* in the *Pou3f3* locus may be mediated by the spatial proximity of its locus to its targets. However, *Dali* RNA itself is not involved in setting these interactions, as the locus exists in the same folded conformation prior to the onset of *Dali* expression. Together with previous findings, such as those on *Hottip* and *Xist*, this work highlights that a single nuclear lncRNA may employ multiple different mechanisms to target to local and distal sites.

7.1.5 *Dali* may regulate gene expression *in trans* by modulating DNA methylation

A growing number of nuclear lncRNAs have been found to bind and function in conjunction with transcriptional regulators and chromatin-modifiers. Moreover, some of the chromatin-modifying complexes, such as PRC2 and MLL, have now thousands of lncRNAs listed as interactors (Yang et al., 2014; Zhao et al., 2010). Apart from potential issues with identifying specific lncRNA-protein interactions, these findings underscore the importance of lncRNAs in the grand scheme of chromatin regulation. That is not to say that a strong propensity of some protein complexes to non-specifically interact with RNA, which constitutes an important part of the nucleus by mass, may not be important for regulation in its own right.

LncRNAs that specifically bind regulatory proteins may perform several functions in such interactions. As was alluded to previously, lncRNAs may serve the purpose of targeting protein complexes to the genome. For example, the MLL complex is recruited to the

genomic region between *Hoxa6* and *Hoxa7* by recognizing the RNA-DNA hybrid formed by the *Mistral* lncRNA and its endogenous locus located there (Bertani et al., 2011). Likewise, PRC2 may be recruited to the *Pitx2* promoter by *Fendrr* which is predicted to interact with its short (< 40 bp) sequence potentially via RNA-DNA:DNA triplex formation (as shown *in vitro*). This is supported by reduced PRC2 occupancy and H3K27me3 at the *Pitx2* promoter in the *Fendrr* knockout embryos (Grote et al., 2013). Together, these findings demonstrate that lncRNAs can recruit chromatin-modifying complexes to both their own sites of synthesis as well as to distally located sites.

LncRNAs appear to not only recruit, but also negatively modulate binding of their protein partners to DNA. For example, the previously mentioned TF-associated lncRNAs *Gas5*, *Jpx*, *Lethe*, and *Panda* were shown to inhibit the binding of their partners to DNA at several loci tested (Hung et al., 2011; Kino et al., 2010; Rapicavoli et al., 2013; Sun et al., 2013).

LncRNAs may also modulate the structure and function of protein complexes. This is illustrated by *rox2* which is involved in sex chromosome dosage compensation in *Drosophila* males. This lncRNA appears to be critical for the assembly of the MSL dosage compensation complex where it sequentially binds to the MLE RNA helicase and MSL2 ubiquitin ligase subunits via its stem-loop domains (Ilik et al., 2013; Maenner et al., 2013). Notably, the initial association with MLE induces an ATP-dependent conformational switch in the *roX2* structure, which stimulates its subsequent binding to MSL2. In mammals, the lncRNAs *Prncr1* and *Pcgem1* were shown to sequentially bind to and stimulate the ability of AR to activate gene expression. First, *Prncr1* together with DOT1L methyltransferase binds to the acetylated enhancer-associated AR which is subsequently

methylated by DOT1L. Methylated AR then acquires increased capacity to bind *Pcgem1*-PYGO2 complex, which is thought to promote enhancer-promoter looping due to the H3K4me3-binding ability of PYGO2 (Li et al., 2013b).

All of these examples serve to highlight that the functions of many lncRNAs are mediated by their modular structures. Moreover, due to the unique chemistry of RNA, lncRNAs individually and as a class can serve as bridges between all and any combination of classes of cellular biopolymers, i.e. RNA, DNA and proteins (Guttman and Rinn, 2012; Mercer and Mattick, 2013; Wang and Chang, 2011). For example, *Hotair* can bring together the PRC2 complex and CoREST transcription factor which it binds via its 5′ and 3′ domains, respectively (Tsai et al., 2010). If acting on the same locus, different "scaffolding" lncRNAs with affinities for distinct types of regulatory factors can provide dynamic regulatory responses of such loci to the incoming developmental or environmental cues. For example, growth control genes were shown to be re-positioned from repressive Polycomb bodies to transcriptionally permissive interchromatin granules by, respectively, *Tug1* and *Malat1* lncRNAs which have differential affinity for methylated and unmethylated Pc2 protein which is localised to these nuclear domains (Yang et al., 2011).

Apart from recruitment or inhibition of binding and scaffolding functions, lncRNAs can also directly modulate the activity of the associated regulatory proteins or act as non-DNA binding co-factors. Examples of such interactions are provided, among others, by lncRNAs associated with DNA methyltransferases. DNMT1 was recently reported to bind to *ecCEBPA* and other similar transcripts genome-wide which inhibit its DNA methylation activity locally via an unknown mechanism (Di Ruscio et al., 2013). *CHD* transcribed antisense to the *E-cadherin* promoter binds to the catalytic domain of DNMT3A to inhibit

its activity (Holz-Schietinger and Reich, 2012). The *Tsix* RNA binds and activates DNMT3A to methylate DNA at the promoter of *Xist*, the lncRNA crucial for X-chromosome inactivation (Sun et al., 2006). *Dali* thus is one of a growing subset of lncRNAs that bind to DNA methyltransferases and modify DNA methylation levels.

In this thesis, I investigated how *Dali* may regulate target genes to which it binds directly. To that end I identified a number of chromatin-associated proteins with which *Dali* interacts both *in vitro* and *in vivo*. Among these proteins are the BAF SWI/SNF family chromatin remodelling complex component BRG1, the transcriptional co-factors P66 and SIN3A, the DNA methyltransferase DNMT1 and the transcription factor POU3F3. The interaction with DNMT1 was observed in both mouse and human, which indicates that the function of *Dali* may be evolutionarily conserved. Targeted DNA methylation studies revealed that depletion of *Dali* is associated with increased DNA methylation at three *Dali*-bound and regulated CpG island-associated promoters, presumable deposited by DNMT1. Therefore, *Dali* appears to regulate gene expression of these genes by inhibiting DNA methylation of their promoters, similar to what was recently reported for *ecCEBPA* and other lncRNAs that inhibit DNMT1-mediated methylation at their sites of synthesis in human. However, unlike all previously reported lncRNAs involved in DNA methylation regulation in mammals, *Dali* appears to work on multiple loci genome-wide away from its own site of synthesis. Conceptually, this can occur via *Dali* competing with the DNA substrate for DNMT1 (similar to what was proposed for *ecCEBPA*), inhibiting its enzymatic activity (as exemplified by *CHD* and DNMT3A), or modulating its interaction with either transcription factors or other co-factors. In fact, *Dali* was also shown to interact with the POU3F3 transcription factor to jointly regulate a set of common targets. In this

respect, *Dali* joins the set of previously identified TF-interacting lncRNAs, although it is not known what the mechanistic significance of this interaction may be. It would be interesting to investigate whether POU3F3 interacts with DNMT1 or, for example, BAF complex, and if so, whether *Dali* may modulate this interaction.

In summary, *Dali* is one of a growing subset of lncRNAs that bind to DNA methyltransferases to modify DNA methylation levels. Importantly, previous examples of lncRNAs that regulate DNA methylation act at their sites of synthesis. Consequently *Dali* is the first reported intergenic lncRNA that regulates DNA methylation *in trans*. *Dali*, like some other lncRNAs, also bind transcription factors. Together, this work and the available literature raise an intriguing possibility that lncRNAs may serve as bridges between sequence-specific transcription-factors and generally sequence non-specific chromatin modifiers.

7.2 Considerations for studying lncRNA function

7.2.1 Approaches to prioritising lncRNAs for functional studies

With tens of thousands of lncRNAs identified to date from mammalian genomes, the important question is how to predict those that are likely functional and to prioritise them for functional studies. It is widely accepted that the mere fact of being transcribed is insufficient to provide evidence for biological function of a lncRNA locus. Certain characteristics are thought to be inconsistent with a particular transcript being a product of

transcriptional noise, including active spatial and temporal regulation of expression (Hangauer et al., 2013), evidence of individual evolutionary constraint (Ponjavic et al., 2009), and regulated post-transcriptional processing (Clark et al., 2012). These characteristics can be considered primary indicators of likely functionality. On the other hand, some other features have been used either arbitrarily, or for technical reasons (such as unavailability of strand information for data generated with some earlier transcriptomics protocols or concerns over potential residual genomic DNA contamination of RNA samples), and are overall poor predictors of functionality. Among these are transcript structure (mono- vs multi-exonic) and expression levels. Multiple studies, including the current one, have shown that both mono-exonic and lowly expressed transcripts can play important biological roles. Therefore, one of the challenges in the field is to develop approaches to reveal the secrets of this potentially rich layer of regulatory information.

One reasonable approach would be to seek and integrate the data on all three primary indicators of functionality for a particular lncRNA set. We are now in a good position to do so, as both computational and experimental tools are evolving rapidly, and the amount of publicly available regulatory and other genome-wide data is growing at an ever-increasing rate. In this study, I tried to follow this framework. *Dali* and several other putative lncRNA candidates were selected form a set of loci evolving under individual evolutionary constraint. Active temporal and spatial regulation was inferred from experimentally determined expression profiles across multiple adult tissues, during a differentiation process modelled in culture, and (for *Dali* only) across an important window of embryonic brain development. I also assessed the chromatin status of both mouse and human *Dali* loci and concluded that it is consistent with the previously determined expression profiles,

indicating that their expression is likely regulated at a chromatin level. I also determined that the *Dali* transcript is capped and polyadenylated, indicating post-transcriptional processing and regulation.

In conclusion, the approach described above to prioritising candidate lncRNAs for in-depth studies is likely to result in the identification of functional lncRNAs. Although performed here on a small set of transcripts, an analogous approach could be applied to genome-wide datasets, yielding a large number of candidates enriched for those that are truly functional.

7.2.2 Experimental and computational approaches to studying nuclear lncRNA function

7.2.2.1 Identifying lncRNA binding loci genome-wide

To have a good chance of uncovering molecular mechanisms of lncRNA function, it is important to study only genes that are true regulatory targets of a particular transcript. For a nuclear localised transcript acting on chromatin, direct targets are expected to be both regulated by it (in loss- and/or gain-of-function studies) and also associated with its binding site either within the target gene itself or in its genomic vicinity. Because there are limitations to both transcriptomics studies of gain- or loss-of function models and approaches used for mapping genome-wide binding profiles of lncRNAs, it is important to integrate both types of data.

The identification of lncRNA target genes with transcriptomic approaches will always be limited by the inherent redundancy of gene regulatory mechanisms, which makes it difficult to identify the specific effects of knockdowns (Biggin, 2011; Farnham, 2009). Also, the technology used to measure transcript levels may impose additional restrictions on the study. In this study, microarrays were used that are known to suffer from sensitivity issues as well as from the measurement of only the portion of the transcriptome represented by probes (Wang et al., 2009). As a result, the list of putative *Dali* target genes was likely to omit some genes that are true targets or miss alternative events that may also be regulated by the lncRNA. Well-designed RNA-seq experiments would potentially be more sensitive and allow more detailed questions to be asked about processes regulated by lncRNAs.

The development of experimental techniques to map the genome-wide occupancy of lncRNAs has been a major advance for the field (Chu et al., 2011; Engreitz et al., 2013; Simon, 2013). Three such techniques conceptually similar to ChIP have been developed recently to map genome-wide occupancy of lncRNAs: chromatin isolation by RNA purification (ChIRP); capture hybridisation analysis of RNA targets (CHART); and RNA antisense purification (RAP) (Chu et al., 2011; Engreitz et al., 2013; Simon, 2013). All employ biotinylated antisense oligos to purify the RNA of interest from cross-linked chromatin extracts and next generation sequencing to identify the associated DNA loci. The differences between the protocols lie in the conditions used for chromatin cross-linking, RNA capture, and elution of the associated DNA, as well as the size and number of antisense oligos used. ChIRP uses glutaraldehyde cross-linking and a pool of approximately 24 20-nt oligos spanning the whole length of the RNA target to be captured

(Chu et al., 2011). RAP, on the other hand, uses a very large number (1054 in the original publication) of much longer 120-nt oligos also spanning the whole lengths of the target to purify it from cells cross-linked with both glutarate and formaldehyde (Engreitz et al., 2013). Arguably, the use of such a large number of oligos might have been justified by the length (approximately 17 kb) and abundance of the target being purified, *Xist*. It was suggested that the use of a large number of oligos spanning the whole length of the target helps improve the performance of these protocols for capturing very long lncRNAs efficiently and uniformly (Chu et al., 2011; Engreitz et al., 2013). Unlike RAP and ChIRP, CHART which was employed in this study, uses a relatively small number of 18-28-nt oligos designed to target accessible regions of the transcript mapped by RNAse H sensitivity assay (Simon, 2013; Simon et al., 2011). In this protocol, formaldehyde cross-linking is performed, followed by capture hybridization at room temperature to reduce the chance of the DNA duplex melting locally and becoming accessible for direct binding to capture oligos. Also, DNA is eluted by RNAse H digestion to elute only DNA that interacts with DNA capture oligos indirectly via RNA, which serves to further reduce those artefacts caused by direct oligo-DNA interactions.

Overall, the optimal choice of a technique to use depends on the size, abundance, and subcellular localisation of the lncRNA of interest. However, all of these techniques are expected to identify both direct and indirect genomic associations of a given lncRNA. They also all have other important limitations. One such limitation is that false positive associations can result from the direct binding of the antisense capture oligos to DNA as well as from the cross-linking of genomic loci existing in close spatial proximity in the nuclear space. Furthermore, studies of transcription factors suggest that a large proportion

of binding events are non-functional and do not confer regulation of the associated loci (Cusanovich et al., 2014; Zhou and O'Shea, 2011). The same is also expected to be true for lncRNAs. For example, only approximately 7% of loci bound by *Paupar* were associated with genes differentially expressed upon its depletion (Vance et al., 2014). In another study, none of the 15 binding loci identified for an eRNA associated with an E2-regulated *FoxC1* enhancer were found in proximity to E2-responsive genes, and therefore, are unlikely to be functional (Li et al., 2013b). Additionally, bound loci are usually attributed to genes based on genomic proximity, which is certain to produce some spurious associations, assigning irrelevant binding in some cases, while in others missing true regulatory interactions. Peak-gene association can potentially be improved by integrating the available chromatin data to assign intergenic regulatory regions to genes independently of genomic proximity (Thurman et al., 2012). It is also important to keep in mind that thus far all studies reporting lncRNA binding profiles investigated them only in one cell type or under one particular set of conditions. It can be argued that some of the bound loci may require the presence of specific co-factors that only become available upon the arrival of a particular stimulus, or may function only under some other conditions.

Therefore, to focus on studying binding events that are functional, it is necessary to identify genes that are both directly bound and regulated by the lncRNA of interest. In the current study, this was achieved by intersecting the list of bound genes with a list of all genes for which there was evidence of being regulated by *Dali* in knockdown studies (both transient and stable, as well as in response to a differentiation stimulus).

7.2.2.2 Studying lncRNA-protein interactions

Although the number of newly identified lncRNAs has been growing at a rapid pace, functional studies are slow and time-consuming. However, a major theme that emerged from these is the wide-spread interaction between lncRNAs and proteins. Therefore, to determine the biochemical activity and ultimately biological function of a particular lncRNA, it is crucial to identify proteins that bind to it specifically and participate in it biochemical functions. A major hurdle for understanding lncRNA-protein interaction has been, and still is, the presence of a large number of non-specific interactors identified in biochemical assays. This is caused in part by the sheer abundance of some basic proteins, such as histones, that can interact with nucleic acids non-specifically and may dominate the final RNA-bound samples. Also, high throughput RNA-IP studies have shown that some chromatin-modifying complexes, such as PRC2 and MLL, interact with thousands of transcripts, including lncRNAs (Yang et al., 2014; Zhao et al., 2010). Furthermore, purified PRC2 complex binds RNA non-specifically *in vitro* in a size-dependent manner (Davidovich et al., 2013). On the other hand, a large number of transcripts also have been found in these studies to associate non-specifically and reproducibly with a variety of RNA-binding proteins (Friedersdorf and Keene, 2014). Additionally, the conditions under which biochemical experiments are performed are inherently prone to artefacts due to the disruption of cellular structure, which may cause structural and other changes in both proteins and RNAs, as well as expose to each other factors that do not normally exist in the same cellular compartment (Hogg and Collins, 2007; Lee et al., 2013). Also, the low abundance of many lncRNAs makes experiments *in vivo* challenging, resulting in much of the currently available data being derived from *in vitro* studies. Thus, when undertaking a

functional lncRNA study, it is important to distinguish specific from nonspecific RNA–protein interactions.

Currently available strategies for identifying proteins associated with specific transcripts employ various affinity tags, including biotin, aptamers, and certain protein-binding sequences to efficiently capture and securely immobilise the RNA of interest on a resin to enable stringent washes aimed at removing non-specific interactors that are thought to be bound with lower affinity than the specific ones (Bachler et al., 1999; Hartmuth et al., 2002; Hogg and Collins, 2007; Srisawat and Engelke, 2001; Youngman and Green, 2005).

Despite the multitude of increasingly technically complicated protocols, most still suffer from the same problem of a high rate of false positive associations. For that reason, in this study, I adopted a two-step approach to identifying protein partners of *Dali*. Initially, I identified a list of putative *Dali*-interacting proteins by performing an *in vitro* RNA pull down assay. At the next stage, individual putative *Dali*-interacting proteins were selected and their interactions with *Dali in vivo* were studied using the UV-RIP approach. UV-RIP relies on the direct molecular link created between the RNA and protein, as well as stringent purification steps which, therefore, should in principle identify only direct molecular interactions. Four out of six selected candidates were found to interact with *Dali* in N2A cells. Out of the two interactions further tested in human cells, the association with DNMT1 was found to be conserved. Therefore, the two-step approach described here is capable of identifying evolutionarily conserved functional lncRNA-protein interactions.

Together with previous reports (Di Ruscio et al., 2013; Mohammad et al., 2010), this thesis confirms DNMT1 DNA methyltransferase to be a non-canonical RNA-binding protein

(RBP). In general, there have been numerous recent studies that report non-canonical RNA-binding proteins, i.e. proteins that do not possess canonical RNA-binding domains and, therefore, were not previously thought to function as RBPs (Davidovich et al., 2013; Kaneko et al., 2010).

Recent advances in both experimental and computational techniques have provided a hitherto unprecedented insight into the world of RNA-protein interactions, including lncRNAs. However, technical problems remain and, therefore, experimental tools will need to develop further if we are to gain mechanistic understanding of the biochemical and biological functions of lncRNAs. Also important is to develop new approaches to study the structural rules governing lncRNA-protein interactions. For example, a tool that would identify lncRNA structure genome-wide and determine which of the structural lncRNA elements interact with what proteins would provide a major advancement.

References

Abarrategui, I., and Krangel, M.S. (2007). Noncoding transcription controls downstream promoters to regulate T-cell receptor alpha recombination. The EMBO journal *26*, 4380-4390.

Abrajano, J.J., Qureshi, I.A., Gokhan, S., Molero, A.E., Zheng, D., Bergman, A., and Mehler, M.F. (2010). Corepressor for element-1-silencing transcription factor preferentially mediates gene networks underlying neural stem cell fate decisions. Proceedings of the National Academy of Sciences of the United States of America *107*, 16685-16690.

Agarwala, K.L., Ganesh, S., Amano, K., Suzuki, T., and Yamakawa, K. (2001). DSCAM, a highly conserved gene in mammals, expressed in differentiating mouse brain. Biochemical and biophysical research communications *281*, 697-705.

Aguilera, A., and Garcia-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Molecular cell *46*, 115-124.

Ahlenstiel, C.L., Lim, H.G., Cooper, D.A., Ishida, T., Kelleher, A.D., and Suzuki, K. (2012). Direct evidence of nuclear Argonaute distribution during transcriptional silencing links the actin cytoskeleton to nuclear RNAi machinery in human cells. Nucleic acids research *40*, 1579-1595.

Allen, D.M., van Praag, H., Ray, J., Weaver, Z., Winrow, C.J., Carter, T.A., Braquet, R., Harrington, E., Ried, T., Brown, K.D*., et al.* (2001). Ataxia telangiectasia mutated is essential during adult neurogenesis. Genes & development *15*, 554-566.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Amaral, P.P., Neyt, C., Wilkins, S.J., Askarian-Amiri, M.E., Sunkin, S.M., Perkins, A.C., and Mattick, J.S. (2009). Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. RNA *15*, 2013-2027.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T*., et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Andrews, G.L., Yun, K., Rubenstein, J.L., and Mastick, G.S. (2003). Dlx transcription factors regulate differentiation of dopaminergic neurons of the ventral thalamus. Molecular and cellular neurosciences *23*, 107-120.

Aoki, M., Yamashita, T., and Tohyama, M. (2004). EphA receptors direct the differentiation of mammalian neural precursor cells through a mitogen-activated protein kinase-dependent pathway. The Journal of biological chemistry *279*, 32643-32650.

Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P., and Proudfoot, N.J. (1997). Intergenic transcription and transinduction of the human beta-globin locus. Genes & development *11*, 2494-2509.

Aubert, B., Bona, M., Boutigny, D., Couderc, F., Karyotakis, Y., Lees, J.P., Poireau, V., Tisserand, V., Zghiche, A., Grauges, E*., et al.* (2007). Observation of CP violation in B --> eta'K0 decays. Physical review letters *98*, 031801.

Bachler, M., Schroeder, R., and von Ahsen, U. (1999). StreptoTag: a novel method for the isolation of RNA-binding proteins. RNA *5*, 1509-1516.

Barry, G., Briggs, J.A., Vanichkina, D.P., Poth, E.M., Beveridge, N.J., Ratnu, V.S., Nayler, S.P., Nones, K., Hu, J., Bredy, T.W*., et al.* (2014). The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. Molecular psychiatry *19*, 486-494.

Bassett, A., Akhtar, A., Barlow, D.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., Higgs, D.R*., et al.* (2014a). Considerations when investigating lncRNA function in vivo. eLife.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W.*, et al.* (2014b). Considerations when investigating lncRNA function in vivo. eLife *3*, e03058.
Batista, P.J., and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell *152*, 1298-1307.
Bennett, C.F., and Swayze, E.E. (2010). RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. Annual review of pharmacology and toxicology *50*, 259-293.
Berezhna, S.Y., Supekova, L., Supek, F., Schultz, P.G., and Deniz, A.A. (2006). siRNA in human cells selectively localizes to target RNA sites. Proceedings of the National Academy of Sciences of the United States of America *103*, 7682-7687.
Berghoff, E.G., Clark, M.F., Chen, S., Cajigas, I., Leib, D.E., and Kohtz, J.D. (2013). Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. Development *140*, 4407-4416.
Berretta, J., Pinskaya, M., and Morillon, A. (2008). A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in S. cerevisiae. Genes & development *22*, 615-626.
Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Molecular cell *43*, 1040-1046.
Bhoopathi, P., Chetty, C., Dontula, R., Gujrati, M., Dinh, D.H., Rao, J.S., and Lakka, S.S. (2011). SPARC stimulates neuronal differentiation of medulloblastoma cells via the Notch1/STAT3 pathway. Cancer research *71*, 4908-4919.
Biggin, M.D. (2011). Animal transcription networks as highly connected, quantitative continua. Developmental cell *21*, 611-626.
Biniszkiewicz, D., Gribnau, J., Ramsahoye, B., Gaudet, F., Eggan, K., Humpherys, D., Mastrangelo, M.A., Jun, Z., Walter, J., and Jaenisch, R. (2002). Dnmt1 overexpression causes genomic hypermethylation, loss of imprinting, and embryonic lethality. Molecular and cellular biology *22*, 2124-2135.
Bird, A.P. (1987). Cpg Islands as Gene Markers in the Vertebrate Nucleus. Trends in Genetics *3*, 342-347.
Blackledge, N.P., and Klose, R. (2011). CpG island chromatin: a platform for gene regulation. Epigenetics : official journal of the DNA Methylation Society *6*, 147-152.
Blume, S.W., Meng, Z., Shrestha, K., Snyder, R.C., and Emanuel, P.D. (2003). The 5'-untranslated RNA of the human dhfr minor transcript alters transcription pre-initiation complex assembly at the major (core) promoter. Journal of cellular biochemistry *88*, 165-180.
Bogdanovic, O., and Veenstra, G.J. (2009). DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma *118*, 549-565.
Bolland, D.J., Wood, A.L., Afshar, R., Featherstone, K., Oltz, E.M., and Corcoran, A.E. (2007). Antisense intergenic transcription precedes Igh D-to-J recombination and is controlled by the intronic enhancer Emu. Molecular and cellular biology *27*, 5523-5533.
Bonasio, R., and Shiekhattar, R. (2014). Regulation of Transcription by Long Noncoding RNAs. Annual review of genetics.
Bond, C.S., and Fox, A.H. (2009). Paraspeckles: nuclear bodies built on long noncoding RNA. The Journal of cell biology *186*, 637-644.
Brackertz, M., Boeke, J., Zhang, R., and Renkawitz, R. (2002). Two highly related p66 proteins comprise a new family of potent transcriptional repressors interacting with MBD2 and MBD3. The Journal of biological chemistry *277*, 40958-40966.

Brand, F., Schumacher, S., Kant, S., Menon, M.B., Simon, R., Turgeon, B., Britsch, S., Meloche, S., Gaestel, M., and Kotlyarov, A. (2012). The extracellular signal-regulated kinase 3 (mitogen-activated protein kinase 6 [MAPK6])-MAPK-activated protein kinase 5 signaling complex regulates septin function and dendrite morphology. Molecular and cellular biology *32*, 2467-2478.
Bros, M., Boissel, J.P., Godtel-Armbrust, U., and Forstermann, U. (2006). Transcription of human neuronal nitric oxide synthase mRNAs derived from different first exons is partly controlled by exon 1-specific promoter sequences. Genomics *87*, 463-473.
Bultman, S., Gebuhr, T., Yee, D., La Mantia, C., Nicholson, J., Gilliam, A., Randazzo, F., Metzger, D., Chambon, P., Crabtree, G.*, et al.* (2000). A Brg1 null mutation in the mouse reveals functional differences among mammalian SWI/SNF complexes. Molecular cell *6*, 1287-1295.
Buske, F.A., Bauer, D.C., Mattick, J.S., and Bailey, T.L. (2012). Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome research *22*, 1372-1381.
Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development *25*, 1915-1927.
Caiafa, P., Guastafierro, T., and Zampieri, M. (2009). Epigenetics: poly(ADP-ribosyl)ation of PARP-1 regulates genomic methylation patterns. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *23*, 672-678.
Cajanek, L., Ribeiro, D., Liste, I., Parish, C.L., Bryja, V., and Arenas, E. (2009). Wnt/beta-catenin signaling blockade promotes neuronal induction and dopaminergic differentiation in embryonic stem cells. Stem Cells *27*, 2917-2927.
Carlessi, L., Fusar Poli, E., De Filippis, L., and Delia, D. (2013). ATM-deficient human neural stem cells as an in vitro model system to study neurodegeneration. DNA repair *12*, 605-611.
Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C.*, et al.* (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559-1563.
Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C.*, et al.* (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454-457.
Carvalho, B.S., and Irizarry, R.A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics *26*, 2363-2367.
Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358-369.
Chang, K.H., Vincent, F., and Shah, K. (2012). Deregulated Cdk5 triggers aberrant activation of cell cycle kinases and phosphatases inducing neuronal death. Journal of cell science *125*, 5124-5137.
Chang, M.C., Tsai, Y.L., Chen, Y.W., Chan, C.P., Huang, C.F., Lan, W.C., Lin, C.C., Lan, W.H., and Jeng, J.H. (2013). Butyrate induces reactive oxygen species production and affects cell cycle progression in human gingival fibroblasts. Journal of periodontal research *48*, 66-73.
Chatterjee, R., and Vinson, C. (2012). CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. Biochimica et biophysica acta *1819*, 763-770.
Chen, L.L., and Carmichael, G.G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Molecular cell *35*, 467-478.
Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnar, Z., and Ponting, C.P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome biology *11*, R72.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Molecular cell *44*, 667-678.
Clark, B.S., and Blackshaw, S. (2014). Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. Frontiers in genetics *5*, 164.
Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome research *22*, 885-898.
Clements, E.G., Mohammad, H.P., Leadem, B.R., Easwaran, H., Cai, Y., Van Neste, L., and Baylin, S.B. (2012). DNMT1 modulates gene expression without its catalytic activity partially through its interactions with histone-modifying enzymes. Nucleic acids research *40*, 4334-4346.
Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Molecular cell *33*, 717-726.
Clemson, C.M., McNeil, J.A., Willard, H.F., and Lawrence, J.B. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. The Journal of cell biology *132*, 259-275.
Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.
Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. PLoS genetics *10*, e1004226.
Davidovich, C., Zheng, L., Goodrich, K.J., and Cech, T.R. (2013). Promiscuous RNA binding by Polycomb repressive complex 2. Nature structural & molecular biology *20*, 1250-1257.
De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS biology *8*, e1000384.
Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes & development *25*, 1010-1022.
Deaton, A.M., Webb, S., Kerr, A.R., Illingworth, R.S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome research *21*, 1074-1086.
Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Molecular cell *54*, 766-776.
Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research *22*, 1775-1789.
Di Ruscio, A., Ebralidze, A.K., Benoukraf, T., Amabile, G., Goff, L.A., Terragni, J., Figueroa, M.E., De Figueiredo Pontes, L.L., Alberich-Jorda, M., Zhang, P.*, et al.* (2013). DNMT1-interacting RNAs block gene-specific DNA methylation. Nature *503*, 371-376.
Dinger, M.E., Amaral, P.P., Mercer, T.R., and Mattick, J.S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Briefings in functional genomics & proteomics *8*, 407-423.
Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F.*, et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.
Dominguez, M.H., Ayoub, A.E., and Rakic, P. (2013). POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. Cereb Cortex *23*, 2632-2643.

Dorus, S., Vallender, E.J., Evans, P.D., Anderson, J.R., Gilbert, S.L., Mahowald, M., Wyckoff, G.J., Malcom, C.M., and Lahn, B.T. (2004). Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell *119*, 1027-1040.
Easwaran, H.P., Schermelleh, L., Leonhardt, H., and Cardoso, M.C. (2004). Replication-independent chromatin loading of Dnmt1 during G2 and M phases. EMBO reports *5*, 1181-1186.
Eissmann, M., Gutschner, T., Hammerle, M., Gunther, S., Caudron-Herger, M., Gross, M., Schirmacher, P., Rippe, K., Braun, T., Zornig, M.*, et al.* (2012). Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. RNA biology *9*, 1076-1087.
Elvers, M., Pfeiffer, J., Kaltschmidt, C., and Kaltschmidt, B. (2005). TGF-beta2 neutralization inhibits proliferation and activates apoptosis of cerebellar granule cell precurors in the developing cerebellum. Mechanisms of development *122*, 587-602.
ENCODE (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology *9*, e1001046.
ENCODE (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.
Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S.*, et al.* (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.
Enokido, Y., Suzuki, E., Iwasawa, K., Namekata, K., Okazawa, H., and Kimura, H. (2005). Cystathionine beta-synthase, a key enzyme for homocysteine metabolism, is preferentially expressed in the radial glia/astrocyte lineage of developing mouse CNS. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *19*, 1854-1856.
Espada, J., Peinado, H., Lopez-Serra, L., Setien, F., Lopez-Serra, P., Portela, A., Renart, J., Carrasco, E., Calvo, M., Juarranz, A.*, et al.* (2011). Regulation of SNAIL1 and E-cadherin function by DNMT1 in a DNA methylation-independent context. Nucleic acids research *39*, 9194-9205.
Espinoza, C.A., Allen, T.A., Hieb, A.R., Kugel, J.F., and Goodrich, J.A. (2004). B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nature structural & molecular biology *11*, 822-829.
Esteve, P.O., Chin, H.G., and Pradhan, S. (2005). Human maintenance DNA (cytosine-5)-methyltransferase and p53 modulate expression of p53-repressed promoters. Proceedings of the National Academy of Sciences of the United States of America *102*, 1000-1005.
Esteve, P.O., Chin, H.G., and Pradhan, S. (2007). Molecular mechanisms of transactivation and doxorubicin-mediated repression of survivin gene in cancer cells. The Journal of biological chemistry *282*, 2615-2625.
Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St Laurent, G., 3rd, Kenny, P.J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nature medicine *14*, 723-730.
Fan, G., Martinowich, K., Chin, M.H., He, F., Fouse, S.D., Hutnick, L., Hattori, D., Ge, W., Shen, Y., Wu, H.*, et al.* (2005). DNA methylation controls the timing of astrogliogenesis through regulation of JAK-STAT signaling. Development *132*, 3345-3356.
Fan, M., Li, X., Jiang, W., Huang, Y., Li, J., and Wang, Z. (2013). A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. Experimental and therapeutic medicine *5*, 1143-1146.
Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nature reviews Genetics *10*, 605-616.

Fathi, A., Hatami, M., Vakilian, H., Han, C.L., Chen, Y.J., Baharvand, H., and Salekdeh, G.H. (2014). Quantitative proteomics analysis highlights the role of redox hemostasis and energy metabolism in human embryonic stem cell differentiation to neural cells. Journal of proteomics *101*, 1-16.
Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nature reviews Genetics *15*, 7-21.
Feng, J., Bi, C., Clark, B.S., Mady, R., Shah, P., and Kohtz, J.D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes & development *20*, 1470-1484.
Filipovska, A., and Rackham, O. (2012). Modular recognition of nucleic acids by PUF, TALE and PPR proteins. Molecular bioSystems *8*, 699-708.
Folch, J., Junyent, F., Verdaguer, E., Auladell, C., Pizarro, J.G., Beas-Zarate, C., Pallas, M., and Camins, A. (2012). Role of cell cycle re-entry in neurons: a common apoptotic mechanism of neuronal cell death. Neurotoxicity research *22*, 195-207.
Fornazari, M., Nascimento, I.C., Nery, A.A., da Silva, C.C., Kowaltowski, A.J., and Ulrich, H. (2011). Neuronal differentiation involves a shift from glucose oxidation to fermentation. Journal of bioenergetics and biomembranes *43*, 531-539.
Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell stem cell *2*, 160-169.
Friedel, C.C., Dolken, L., Ruzsics, Z., Koszinowski, U.H., and Zimmer, R. (2009). Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic acids research *37*, e115.
Friedersdorf, M.B., and Keene, J.D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome biology *15*, R2.
Gal-Yam, E.N., Egger, G., Iniguez, L., Holster, H., Einarsson, S., Zhang, X., Lin, J.C., Liang, G., Jones, P.A., and Tanay, A. (2008). Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proceedings of the National Academy of Sciences of the United States of America *105*, 12979-12984.
Gao, X., Yang, J., Tsang, J.C., Ooi, J., Wu, D., and Liu, P. (2013). Reprogramming to Pluripotency Using Designer TALE Transcription Factors Targeting Enhancers. Stem cell reports *1*, 183-197.
Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. Journal of molecular biology *196*, 261-282.
Georgia, S., Kanji, M., and Bhushan, A. (2013). DNMT1 represses p53 to maintain progenitor cell survival during pancreatic organogenesis. Genes & development *27*, 372-377.
Georgopoulou, N., Hurel, C., Politis, P.K., Gaitanou, M., Matsas, R., and Thomaidou, D. (2006). BM88 is a dual function molecule inducing cell cycle exit and neuronal differentiation of neuroblastoma cells via cyclin D1 down-regulation and retinoblastoma protein hypophosphorylation. The Journal of biological chemistry *281*, 33606-33620.
Giallourakis, C.C., Franklin, A., Guo, C., Cheng, H.L., Yoon, H.S., Gallagher, M., Perlot, T., Andzelm, M., Murphy, A.J., Macdonald, L.E*., et al.* (2010). Elements between the IgH variable (V) and diversity (D) clusters influence antisense transcription and lineage-specific V(D)J recombination. Proceedings of the National Academy of Sciences of the United States of America *107*, 22207-22212.
Gill, R.M., Gabor, T.V., Couzens, A.L., and Scheid, M.P. (2013). The MYC-associated protein CDCA7 is phosphorylated by AKT to regulate MYC-dependent apoptosis and transformation. Molecular and cellular biology *33*, 498-513.
Gitai, D.L., Fachin, A.L., Mello, S.S., Elias, C.F., Bittencourt, J.C., Leite, J.P., Passos, G.A., Garcia-Cairasco, N., and Paco-Larson, M.L. (2011). The non-coding RNA BC1 is down-regulated in the

hippocampus of Wistar Audiogenic Rat (WAR) strain after audiogenic kindling. Brain research *1367*, 114-121.
Gomez, J.A., Wapinski, O.L., Yang, Y.W., Bureau, J.F., Gopinath, S., Monack, D.M., Chang, H.Y., Brahic, M., and Kirkegaard, K. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. Cell *152*, 743-754.
Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature *470*, 284-288.
Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R., and Fraser, P. (2000). Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. Molecular cell *5*, 377-386.
Grimes, J.A., Nielsen, S.J., Battaglioli, E., Miska, E.A., Speh, J.C., Berry, D.L., Atouf, F., Holdener, B.C., Mandel, G., and Kouzarides, T. (2000). The co-repressor mSin3A is a functional component of the REST-CoREST repressor complex. The Journal of biological chemistry *275*, 9461-9467.
Grote, P., Wittler, L., Hendrix, D., Koch, F., Wahrisch, S., Beisaw, A., Macura, K., Blass, G., Kellis, M., Werber, M.*, et al.* (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Developmental cell *24*, 206-214.
Guastafierro, T., Cecchinelli, B., Zampieri, M., Reale, A., Riggio, G., Sthandier, O., Zupi, G., Calabrese, L., and Caiafa, P. (2008). CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery. The Journal of biological chemistry *283*, 21873-21880.
Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L.*, et al.* (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071-1076.
Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.*, et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.
Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L.*, et al.* (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295-300.
Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339-346.
Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240-251.
Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R.*, et al.* (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nature structural & molecular biology *21*, 198-206.
Haerty, W., and Ponting, C.P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome biology *14*, R49.
Haerty, W., and Ponting, C.P. (2014). No Gene in the Genome Makes Sense Except in the Light of Evolution. Annual review of genomics and human genetics.
Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forne, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nature protocols *2*, 1722-1733.
Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome research *23*, 1210-1223.
Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS genetics *9*, e1003569.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384-388.
Hansen, T.B., Wiklund, E.D., Bramsen, J.B., Villadsen, S.B., Statham, A.L., Clark, S.J., and Kjems, J. (2011). miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. The EMBO journal *30*, 4414-4422.
Hartmuth, K., Urlaub, H., Vornlocher, H.P., Will, C.L., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. Proceedings of the National Academy of Sciences of the United States of America *99*, 16719-16724.
Hartwig, C., Veske, A., Krejcova, S., Rosenberger, G., and Finckh, U. (2005). Plexin B3 promotes neurite outgrowth, interacts homophilically, and interacts with Rin. BMC neuroscience *6*, 53.
Heger, A., Webber, C., Goodson, M., Ponting, C.P., and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics *29*, 2046-2048.
Hekimoglu-Balkan, B., Aszodi, A., Heinen, R., Jaritz, M., and Ringrose, L. (2012). Intergenic Polycomb target sites are dynamically marked by non-coding transcription during lineage commitment. RNA biology *9*, 314-325.
Hermann, A., Goyal, R., and Jeltsch, A. (2004). The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. The Journal of biological chemistry *279*, 48350-48359.
Hervouet, E., Nadaradjane, A., Gueguen, M., Vallette, F.M., and Cartron, P.F. (2012). Kinetics of DNA methylation inheritance by the Dnmt1-including complexes during the cell cycle. Cell division *7*, 5.
Hervouet, E., Vallette, F.M., and Cartron, P.F. (2010). Dnmt1/Transcription factor interactions: an alternative mechanism of DNA methylation inheritance. Genes & cancer *1*, 434-443.
Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America *106*, 9362-9367.
Hjelm, B.E., Salhia, B., Kurdoglu, A., Szelinger, S., Reiman, R.A., Sue, L.I., Beach, T.G., Huentelman, M.J., and Craig, D.W. (2013). In vitro-differentiated neural cell cultures progress towards donor-identical brain tissue. Human molecular genetics *22*, 3534-3546.
Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E.*, et al.* (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic acids research *41*, 827-841.
Hogg, J.R., and Collins, K. (2007). RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. RNA *13*, 868-880.
Holz-Schietinger, C., and Reich, N.O. (2012). RNA modulation of the human DNA methyltransferase 3A. Nucleic acids research *40*, 8550-8557.
Hoshino, M., and Nakamura, S. (2003). Small GTPase Rin induces neurite outgrowth through Rac/Cdc42 and calmodulin in PC12 cells. The Journal of cell biology *163*, 1067-1076.
Hu, X., and Zuckerman, K.S. (2001). Transforming growth factor: signal transduction pathways, cell cycle mediation, and effects on hematopoiesis. Journal of hematotherapy & stem cell research *10*, 67-74.
Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P.*, et al.* (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nature genetics *43*, 621-629.

Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., and Chess, A. (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC genomics *8*, 39.
Hutnick, L.K., Golshani, P., Namihira, M., Xue, Z., Matynia, A., Yang, X.W., Silva, A.J., Schweizer, F.E., and Fan, G. (2009). DNA hypomethylation restricted to the murine forebrain induces cortical degeneration and impairs postnatal neuronal maturation. Human molecular genetics *18*, 2875-2888.
Ideue, T., Hino, K., Kitao, S., Yokoi, T., and Hirose, T. (2009). Efficient oligonucleotide-mediated degradation of nuclear noncoding RNAs in mammalian cultured cells. RNA *15*, 1578-1587.
Iguchi-Ariga, S.M., and Schaffner, W. (1989). CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes & development *3*, 612-619.
Ilik, I.A., Quinn, J.J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R.C., Luscombe, N., Backofen, R*., et al.* (2013). Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. Molecular cell *51*, 156-173.
Illingworth, R., Kerr, A., Desousa, D., Jorgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J*., et al.* (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS biology *6*, e22.
Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS genetics *6*, e1001134.
Inomata, K., Ohki, I., Tochio, H., Fujiwara, K., Hiroaki, H., and Shirakawa, M. (2008). Kinetic and thermodynamic evidence for flipping of a methyl-CpG binding domain on methylated DNA. Biochemistry *47*, 3266-3271.
Jeon, Y., and Lee, J.T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. Cell *146*, 119-133.
Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E*., et al.* (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031-8041.
Jiang, J., Jing, Y., Cost, G.J., Chiang, J.C., Kolpa, H.J., Cotton, A.M., Carone, D.M., Carone, B.R., Shivak, D.A., Guschin, D.Y*., et al.* (2013). Translating dosage compensation to trisomy 21. Nature *500*, 296-300.
Johnsson, P., and Morris, K.V. (2014). Expanding the functional role of long noncoding RNAs. Cell research.
Kaneko, S., Li, G., Son, J., Xu, C.F., Margueron, R., Neubert, T.A., and Reinberg, D. (2010). Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. Genes & development *24*, 2615-2620.
Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M*., et al.* (2011). DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell stem cell *8*, 676-687.
Karreth, F.A., Tay, Y., Perna, D., Ala, U., Tan, S.M., Rust, A.G., DeNicola, G., Webster, K.A., Weiss, D., Perez-Mancera, P.A*., et al.* (2011). In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. Cell *147*, 382-395.
Kim, H., and Kim, J.S. (2014). A guide to genome engineering with programmable nucleases. Nature reviews Genetics *15*, 321-334.
Kim, J., Kollhoff, A., Bergmann, A., and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Human molecular genetics *12*, 233-245.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S.*, et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.
Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Science signaling *3*, ra8.
Koerner, M.V., Pauler, F.M., Huang, R., and Barlow, D.P. (2009). The function of non-coding RNAs in genomic imprinting. Development *136*, 1771-1783.
Kornienko, A.E., Guenzl, P.M., Barlow, D.P., and Pauler, F.M. (2013). Gene regulation by the act of long non-coding RNA transcription. BMC biology *11*, 59.
Korostowski, L., Sedlak, N., and Engel, N. (2012). The Kcnq1ot1 long non-coding RNA affects chromatin conformation and expression of Kcnq1, but does not regulate its imprinting in the developing heart. PLoS genetics *8*, e1002956.
Kowalczyk, M.S., Higgs, D.R., and Gingeras, T.R. (2012). Molecular biology: RNA discrimination. Nature *482*, 310-311.
Kraus, P., Sivakamasundari, V., Lim, S.L., Xing, X., Lipovich, L., and Lufkin, T. (2013). Making sense of Dlx1 antisense RNA. Developmental biology *376*, 224-235.
Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J.*, et al.* (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231-235.
Kulis, M., Queiros, A.C., Beekman, R., and Martin-Subero, J.I. (2013). Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. Biochimica et biophysica acta *1829*, 1161-1174.
Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497-501.
Laketa, V., Simpson, J.C., Bechtel, S., Wiemann, S., and Pepperkok, R. (2007). High-content microscopy identifies new neurite outgrowth regulators. Molecular biology of the cell *18*, 242-252.
Lam, M.T., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M.*, et al.* (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511-515.
Latos, P.A., Pauler, F.M., Koerner, M.V., Senergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E.*, et al.* (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469-1472.
Lee, H.Y., Haurwitz, R.E., Apffel, A., Zhou, K., Smart, B., Wenger, C.D., Laderman, S., Bruhn, L., and Doudna, J.A. (2013). RNA-protein analysis using a conditional CRISPR nuclease. Proceedings of the National Academy of Sciences of the United States of America *110*, 5416-5421.
Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435-1439.
Lei, H., Oh, S.P., Okano, M., Juttermann, R., Goss, K.A., Jaenisch, R., and Li, E. (1996). De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development *122*, 3195-3205.
Lessard, J., Wu, J.I., Ranish, J.A., Wan, M., Winslow, M.M., Staahl, B.T., Wu, H., Aebersold, R., Graef, I.A., and Crabtree, G.R. (2007). An essential switch in subunit composition of a chromatin remodeling complex during neural development. Neuron *55*, 201-215.
Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915-926.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A., *et al.* (2013a). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. Cell reports *5*, 3-12.
Li, L.C., and Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. Bioinformatics *18*, 1427-1431.
Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., *et al.* (2013b). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516-520.
Li, Z., Chao, T.C., Chang, K.Y., Lin, N., Patil, V.S., Shimizu, C., Head, S.R., Burns, J.C., and Rana, T.M. (2014). The long noncoding RNA THRIL regulates TNFalpha expression through its interaction with hnRNPL. Proceedings of the National Academy of Sciences of the United States of America *111*, 1002-1007.
Ligtenberg, M.J., Kuiper, R.P., Chan, T.L., Goossens, M., Hebeda, K.M., Voorendt, M., Lee, T.Y., Bodmer, D., Hoenselaar, E., Hendriks-Cornelissen, S.J., *et al.* (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nature genetics *41*, 112-117.
Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., *et al.* (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476-482.
Liu, X., Gao, Q., Li, P., Zhao, Q., Zhang, J., Li, J., Koseki, H., and Wong, J. (2013). UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. Nature communications *4*, 1563.
Long, H.K., Sims, D., Heger, A., Blackledge, N.P., Kutter, C., Wright, M.L., Grutzner, F., Odom, D.T., Patient, R., Ponting, C.P., *et al.* (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. eLife *2*, e00348.
Ma, L., Bajic, V.B., and Zhang, Z. (2013). On the classification of long non-coding RNAs. RNA biology *10*, 925-933.
Maamar, H., Cabili, M.N., Rinn, J., and Raj, A. (2013). linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis. Genes & development *27*, 1260-1271.
Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696-1697.
Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes & development *8*, 2282-2292.
Maenner, S., Muller, M., Frohlich, J., Langer, D., and Becker, P.B. (2013). ATP-dependent roX RNA remodeling by the helicase maleless enables specific association of MSL proteins. Molecular cell *51*, 174-184.
Mager, J., Montgomery, N.D., de Villena, F.P., and Magnuson, T. (2003). Genome imprinting regulated by the mouse Polycomb group protein Eed. Nature genetics *33*, 502-507.
Mamalaki, A., Boutou, E., Hurel, C., Patsavoudi, E., Tzartos, S., and Matsas, R. (1995). The BM88 antigen, a novel neuron-specific molecule, enhances the differentiation of mouse neuroblastoma cells. The Journal of biological chemistry *270*, 14201-14208.
Mancini, D.N., Singh, S.M., Archer, T.K., and Rodenhiser, D.I. (1999). Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. Oncogene *18*, 4108-4119.
Mao, Y.S., Sunwoo, H., Zhang, B., and Spector, D.L. (2011). Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. Nature cell biology *13*, 95-101.

Mariner, P.D., Walters, R.D., Espinoza, C.A., Drullinger, L.F., Wagner, S.D., Kugel, J.F., and Goodrich, J.A. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Molecular cell *29*, 499-509.
Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome research *24*, 496-510.
Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome biology *14*, R131.
Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome biology *10*, R124.
Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., and Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature *445*, 666-670.
Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome biology *5*, R101.
Masuho, I., Mototani, Y., Sahara, Y., Asami, J., Nakamura, S., Kozasa, T., and Inoue, T. (2008). Dynamic expression patterns of G protein-regulated inducer of neurite outgrowth 1 (GRIN1) and its colocalization with Galphao implicate significant roles of Galphao-GRIN1 signaling in nervous system. Developmental dynamics : an official publication of the American Association of Anatomists *237*, 2415-2429.
Matsumoto, S., Banine, F., Struve, J., Xing, R., Adams, C., Liu, Y., Metzger, D., Chambon, P., Rao, M.S., and Sherman, L.S. (2006). Brg1 is required for murine neural stem cell maintenance and gliogenesis. Developmental biology *289*, 372-383.
Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic acids research *34*, D108-110.
Matzke, M.A., and Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature reviews Genetics *15*, 394-408.
Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y.*, et al.* (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253-257.
McEvilly, R.J., de Diaz, M.O., Schonemann, M.D., Hooshmand, F., and Rosenfeld, M.G. (2002). Transcriptional regulation of cortical neuron migration by POU domain factors. Science *295*, 1528-1532.
McIsaac, R.S., Oakes, B.L., Wang, X., Dummit, K.A., Botstein, D., and Noyes, M.B. (2013). Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. Nucleic acids research *41*, e57.
McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.
McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC bioinformatics *11*, 165.
Medvedeva, Y.A., Khamis, A.M., Kulakovskiy, I.V., Ba-Alawi, W., Bhuyan, M.S., Kawaji, H., Lassmann, T., Harbers, M., Forrest, A.R., and Bajic, V.B. (2014). Effects of cytosine methylation on transcription factor binding sites. BMC genomics *15*, 119.
Mehler, M.F., and Mattick, J.S. (2007). Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiological reviews *87*, 799-823.

Melo, C.A., Drost, J., Wijchers, P.J., van de Werken, H., de Wit, E., Oude Vrielink, J.A., Elkon, R., Melo, S.A., Leveille, N., Kalluri, R.*, et al.* (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. Molecular cell *49*, 524-535.
Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M.*, et al.* (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333-338.
Meola, N., Pizzo, M., Alfano, G., Surace, E.M., and Banfi, S. (2012). The long noncoding RNA Vax2os1 controls the cell cycle progression of photoreceptor progenitors in the mouse retina. RNA *18*, 111-123.
Mercer, A.C., Gaj, T., Sirk, S.J., Lamb, B.M., and Barbas, C.F., 3rd (2013). Regulation of Endogenous Human Gene Expression by Ligand-Inducible TALE Transcription Factors. ACS synthetic biology.
Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proceedings of the National Academy of Sciences of the United States of America *105*, 716-721.
Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nature structural & molecular biology *20*, 300-307.
Milet, C., Maczkowiak, F., Roche, D.D., and Monsoro-Burq, A.H. (2013). Pax3 and Zic1 drive induction and differentiation of multipotent, migratory, and functional neural crest in Xenopus embryos. Proceedings of the National Academy of Sciences of the United States of America *110*, 5528-5533.
Mili, S., and Steitz, J.A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. RNA *10*, 1692-1694.
Miloso, M., Scuteri, A., Foudah, D., and Tredici, G. (2008). MAPKs as mediators of cell fate determination: an approach to neurodegenerative diseases. Current medicinal chemistry *15*, 538-548.
Ming, G.L., and Song, H. (2011). Adult neurogenesis in the mammalian brain: significant answers and significant questions. Neuron *70*, 687-702.
Mohammad, F., Mondal, T., Guseva, N., Pandey, G.K., and Kanduri, C. (2010). Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. Development *137*, 2493-2499.
Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schubeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular cell *30*, 755-766.
Monnier, P., Martinet, C., Pontis, J., Stancheva, I., Ait-Si-Ali, S., and Dandolo, L. (2013). H19 lncRNA controls gene expression of the Imprinted Gene Network by recruiting MBD1. Proceedings of the National Academy of Sciences of the United States of America *110*, 20693-20698.
Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G.L., and Sartorelli, V. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. Molecular cell *51*, 606-617.
Mutai, H., Nagashima, R., Sugitani, Y., Noda, T., Fujii, M., and Matsunaga, T. (2009). Expression of Pou3f3/Brn-1 and its genomic methylation in developing auditory epithelium. Developmental neurobiology *69*, 913-930.
Myers, K.A., and Baas, P.W. (2007). Kinesin-5 regulates the growth of the axon by acting as a brake on its microtubule array. The Journal of cell biology *178*, 1081-1091.
Nadar, V.C., Lin, S., and Baas, P.W. (2012). Microtubule redistribution in growth cones elicited by focal inactivation of kinesin-5. The Journal of neuroscience : the official journal of the Society for Neuroscience *32*, 5783-5794.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717-1720.
Nakai, S., Sugitani, Y., Sato, H., Ito, S., Miura, Y., Ogawa, M., Nishi, M., Jishage, K., Minowa, O., and Noda, T. (2003). Crucial roles of Brn1 in distal tubule formation and function in mouse kidney. Development *130*, 4751-4759.
Narayanan, R., and Tuoc, T.C. (2014). Roles of chromatin remodeling BAF complex in neural differentiation and reprogramming. Cell and tissue research *356*, 575-584.
Natoli, G., and Andrau, J.C. (2012). Noncoding transcription at enhancers: general principles and functional models. Annual review of genetics *46*, 1-19.
Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635-640.
Ng, S.Y., Bogu, G.K., Soh, B.S., and Stanton, L.W. (2013). The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. Molecular cell *51*, 349-359.
Ng, S.Y., Johnson, R., and Stanton, L.W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. The EMBO journal *31*, 522-533.
Nguyen, V.T., Kiss, T., Michels, A.A., and Bensaude, O. (2001). 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. Nature *414*, 322-325.
Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E.M. (2004). Megabase deletions of gene deserts result in viable mice. Nature *431*, 988-993.
Ogura, T., Yokoyama, T., Fujisawa, H., Kurashima, Y., and Esumi, H. (1993). Structural diversity of neuronal nitric oxide synthase mRNA in the nervous system. Biochemical and biophysical research communications *193*, 1014-1022.
Ohrt, T., and Schwille, P. (2008). siRNA modifications and sub-cellular localization: a question of intracellular transport? Current pharmaceutical design *14*, 3674-3685.
Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H*., et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563-573.
Olmsted, J.B., Carlson, K., Klebe, R., Ruddle, F., and Rosenbaum, J. (1970). Isolation of microtubule protein from cultured mouse neuroblastoma cells. Proceedings of the National Academy of Sciences of the United States of America *65*, 129-136.
Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q*., et al.* (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46-58.
Palmer, A.C., Ahlgren-Berg, A., Egan, J.B., Dodd, I.B., and Shearwin, K.E. (2009). Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter. Molecular cell *34*, 545-555.
Palmer, A.C., Egan, J.B., and Shearwin, K.E. (2011). Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors. Transcription *2*, 9-14.
Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Molecular cell *32*, 232-246.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes & development *10*, 1991-2002.
Paoletti, L., Elena, C., Domizi, P., and Banchio, C. (2011). Role of phosphatidylcholine during neuronal differentiation. IUBMB life *63*, 714-720.
Pauler, F.M., Barlow, D.P., and Hudson, Q.J. (2012). Mechanisms of long range silencing by imprinted macro non-coding RNAs. Current opinion in genetics & development *22*, 283-289.
Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A*., et al.* (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome research *22*, 577-591.
Payer, B., and Lee, J.T. (2008). X chromosome dosage compensation: how mammals keep the balance. Annual review of genetics *42*, 733-772.
Peco, E., Escude, T., Agius, E., Sabado, V., Medevielle, F., Ducommun, B., and Pituello, F. (2012). The CDC25B phosphatase shortens the G2 phase of neural progenitors and promotes efficient neuron production. Development *139*, 1095-1104.
Pennisi, E. (2014). Cell biology. Lengthy RNAs earn respect as cellular players. Science *344*, 1072.
Persengiev, S.P., Li, J., Poulin, M.L., and Kilpatrick, D.L. (2001). E2F2 converts reversibly differentiated PC12 cells to an irreversible, neurotrophin-dependent state. Oncogene *20*, 5124-5131.
Petryniak, M.A., Potter, G.B., Rowitch, D.H., and Rubenstein, J.L. (2007). Dlx1 and Dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. Neuron *55*, 417-433.
Pinter, S.F., Sadreyev, R.I., Yildirim, E., Jeon, Y., Ohsumi, T.K., Borowsky, M., and Lee, J.T. (2012). Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations. Genome research *22*, 1864-1876.
Politis, P.K., Thomaidou, D., and Matsas, R. (2008). Coordination of cell cycle exit and differentiation of neuronal progenitors. Cell Cycle *7*, 691-697.
Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R*., et al.* (2006a). Forces shaping the fastest evolving regions in the human genome. PLoS genetics *2*, e168.
Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A*., et al.* (2006b). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167-172.
Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS genetics *5*, e1000617.
Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome research *17*, 556-565.
Powell, W.T., Coulson, R.L., Crary, F.K., Wong, S.S., Ach, R.A., Tsang, P., Alice Yamada, N., Yasui, D.H., and Lasalle, J.M. (2013). A Prader-Willi locus lncRNA cloud modulates diurnal genes and energy expenditure. Human molecular genetics *22*, 4318-4328.
Qureshi, I.A., Gokhan, S., and Mehler, M.F. (2010). REST and CoREST are transcriptional and epigenetic regulators of seminal neural fate decisions. Cell Cycle *9*, 4477-4486.
Qureshi, I.A., and Mehler, M.F. (2012). Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nature reviews Neuroscience *13*, 528-541.
Racanelli, A.C., Turner, F.B., Xie, L.Y., Taylor, S.M., and Moran, R.G. (2008). A mouse gene that coordinates epigenetic controls and transcriptional interference to achieve tissue-specific expression. Molecular and cellular biology *28*, 836-848.

Ramos, A.D., Diaz, A., Nellore, A., Delgado, R.N., Park, K.Y., Gonzales-Roybal, G., Oldham, M.C., Song, J.S., and Lim, D.A. (2013). Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. Cell stem cell *12*, 616-628.
Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 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.
Rapicavoli, N.A., Poth, E.M., and Blackshaw, S. (2010). The long noncoding RNA RNCR2 directs mouse retinal cell specification. BMC developmental biology *10*, 49.
Rapicavoli, N.A., Poth, E.M., Zhu, H., and Blackshaw, S. (2011). The long noncoding RNA Six3OS acts in trans to regulate retinal development by modulating Six3 activity. Neural development *6*, 32.
Rapicavoli, N.A., Qu, K., Zhang, J., Mikhail, M., Laberge, R.M., and Chang, H.Y. (2013). A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. eLife *2*, e00762.
Rassoulzadegan, M., Magliano, M., and Cuzin, F. (2002). Transvection effects involving DNA methylation during meiosis in the mouse. The EMBO journal *21*, 440-450.
Reed, R., and Chiara, M.D. (1999). Identification of RNA-protein contacts within functional ribonucleoprotein complexes by RNA site-specific labeling and UV crosslinking. Methods *18*, 3-12.
Reeves, F.C., Burdge, G.C., Fredericks, W.J., Rauscher, F.J., and Lillycrop, K.A. (1999). Induction of antisense Pax-3 expression leads to the rapid morphological differentiation of neuronal cells and an altered response to the mitogenic growth factor bFGF. Journal of cell science *112 (Pt 2)*, 253-261.
Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends in genetics : TIG *16*, 276-277.
Riley, K.J., and Steitz, J.A. (2013). The "Observer Effect" in genome-wide surveys of protein-RNA interactions. Molecular cell *49*, 601-604.
Riley, K.J., Yario, T.A., and Steitz, J.A. (2012). Association of Argonaute proteins and microRNAs can occur after cell lysis. RNA *18*, 1581-1585.
Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annual review of biochemistry *81*, 145-166.
Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E*., et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.
Rishi, V., Bhattacharya, P., Chatterjee, R., Rozenberg, J., Zhao, J., Glass, K., Fitzgerald, P., and Vinson, C. (2010). CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. Proceedings of the National Academy of Sciences of the United States of America *107*, 20311-20316.
Robertson, K.D., Ait-Si-Ali, S., Yokochi, T., Wade, P.A., Jones, P.L., and Wolffe, A.P. (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. Nature genetics *25*, 338-342.
Roesler, W.J., Vandenbark, G.R., and Hanson, R.W. (1988). Cyclic AMP and the induction of eukaryotic gene transcription. The Journal of biological chemistry *263*, 9063-9066.
Rollins, R.A., Haghighi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., and Bestor, T.H. (2006). Large-scale structure of genomic methylation patterns. Genome research *16*, 157-163.
Rolls, A., Shechter, R., London, A., Ziv, Y., Ronen, A., Levy, R., and Schwartz, M. (2007). Toll-like receptors modulate adult hippocampal neurogenesis. Nature cell biology *9*, 1081-1088.

Roussa, E., Gaedicke, S., and Krivokuca, D. (2013). Transforming growth factor betas (TGF-beta s) are novel extrinsic determinants in the induction and survival of mouse rostral hindbrain serotonergic neurons. Faseb Journal *27*.
Sahara, S., Yanagawa, Y., O'Leary, D.D., and Stevens, C.F. (2012). The fraction of cortical GABAergic neurons is constant from near the start of cortical neurogenesis to adulthood. The Journal of neuroscience : the official journal of the Society for Neuroscience *32*, 4755-4761.
Sahoo, T., del Gaudio, D., German, J.R., Shinawi, M., Peters, S.U., Person, R.E., Garnica, A., Cheung, S.W., and Beaudet, A.L. (2008). Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nature genetics *40*, 719-721.
Sakurai, T. (2012). The role of NrCAM in neural development and disorders--beyond a simple glue in the brain. Molecular and cellular neurosciences *49*, 351-363.
Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353-358.
Sanjana, N.E., Cong, L., Zhou, Y., Cunniff, M.M., Feng, G., and Zhang, F. (2012). A transcription activator-like effector toolbox for genome engineering. Nature protocols *7*, 171-192.
Santoro, F., and Barlow, D.P. (2011). Developmental control of imprinted expression by macro non-coding RNAs. Seminars in cell & developmental biology *22*, 328-335.
Santoro, F., Mayer, D., Klement, R.M., Warczok, K.E., Stukalov, A., Barlow, D.P., and Pauler, F.M. (2013). Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. Development *140*, 1184-1195.
Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M.*, et al.* (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.
Saxonov, S., Berg, P., and Brutlag, D.L. (2006). 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.
Schmitt, S., Prestel, M., and Paro, R. (2005). Intergenic transcription through a polycomb group response element counteracts silencing. Genes & development *19*, 697-708.
Schmitz, K.M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. Genes & development *24*, 2264-2269.
Seidl, C.I., Stricker, S.H., and Barlow, D.P. (2006). The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. The EMBO journal *25*, 3565-3575.
Seisenberger, S., Peat, J.R., and Reik, W. (2013). Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. Current opinion in cell biology *25*, 281-288.
Sen, G.L., Reuter, J.A., Webster, D.E., Zhu, L., and Khavari, P.A. (2010). DNMT1 maintains progenitor function in self-renewing somatic tissue. Nature *463*, 563-567.
Serganov, A., and Patel, D.J. (2007). Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. Nature reviews Genetics *8*, 776-790.
Sessa, L., Breiling, A., Lavorgna, G., Silvestri, L., Casari, G., and Orlando, V. (2007). Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster. RNA *13*, 223-239.
Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., and Ko, M.S. (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 : an international journal for rapid publication of reports on genes and genomes *16*, 45-58.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V.*, et al.* (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.
Shevtsov, S.P., and Dundr, M. (2011). Nucleation of nuclear bodies by RNA. Nature cell biology *13*, 167-173.
Siegenthaler, J.A., and Miller, M.W. (2005). Transforming growth factor beta 1 promotes cell cycle exit through the cyclin-dependent kinase inhibitor p21 in the developing cerebral cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience *25*, 8627-8636.
Simon, M.D. (2013). Capture hybridization analysis of RNA targets (CHART). Current protocols in molecular biology / edited by Frederick M Ausubel [et al] *Chapter 21*, Unit 21 25.
Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465-469.
Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proceedings of the National Academy of Sciences of the United States of America *108*, 20497-20502.
Smallwood, S.A., and Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. Trends in genetics : TIG *28*, 33-42.
Smith, A.G. (1991). Culture and differentiation of embryonic stem cells. Journal of Tissue Culture Methods *13*, 89-94.
Smith, M.A., Gesell, T., Stadler, P.F., and Mattick, J.S. (2013). Widespread purifying selection on RNA structure in mammals. Nucleic acids research *41*, 8220-8236.
Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nature reviews Genetics *14*, 204-220.
Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology *3*, Article3.
Sproul, D., and Meehan, R.R. (2013). Genomic insights into cancer-associated aberrant CpG island hypermethylation. Briefings in functional genomics *12*, 174-190.
Srisawat, C., and Engelke, D.R. (2001). Streptavidin aptamers: affinity tags for the study of RNAs and ribonucleoproteins. RNA *7*, 632-641.
Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A., and Cedar, H. (1982). Clonal inheritance of the pattern of DNA methylation in mouse cells. Proceedings of the National Academy of Sciences of the United States of America *79*, 61-65.
Steshina, E.Y., Carr, M.S., Glick, E.A., Yevtodiyenko, A., Appelbe, O.K., and Schmidt, J.V. (2006). Loss of imprinting at the Dlk1-Gtl2 locus caused by insertional mutagenesis in the Gtl2 5' region. BMC genetics *7*, 44.
Stoltenburg, R., Reinemann, C., and Strehlitz, B. (2007). SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomolecular engineering *24*, 381-403.
Sumazin, P., Yang, X., Chiu, H.S., Chung, W.J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J.*, et al.* (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell *147*, 370-381.
Sun, B.K., Deaton, A.M., and Lee, J.T. (2006). A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. Molecular cell *21*, 617-628.
Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y., and Lee, J.T. (2013). Jpx RNA activates Xist by evicting CTCF. Cell *153*, 1537-1551.

Sunwoo, H., Dinger, M.E., Wilusz, J.E., Amaral, P.P., Mattick, J.S., and Spector, D.L. (2009). MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome research *19*, 347-359.
Taft, R.J., Pheasant, M., and Mattick, J.S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays : news and reviews in molecular, cellular and developmental biology *29*, 288-299.
Takayama, K., Horie-Inoue, K., Katayama, S., Suzuki, T., Tsutsumi, S., Ikeda, K., Urano, T., Fujimura, T., Takagi, K., Takahashi, S*., et al.* (2013). Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. The EMBO journal *32*, 1665-1680.
Talkowski, M.E., Maussion, G., Crapper, L., Rosenfeld, J.A., Blumenthal, I., Hanscom, C., Chiang, C., Lindgren, A., Pereira, S., Ruderfer, D*., et al.* (2012). Disruption of a large intergenic noncoding RNA in subjects with neurodevelopmental disabilities. American journal of human genetics *91*, 1128-1134.
Tang, X., Hou, M., Ding, Y., Li, Z., Ren, L., and Gao, G. (2013). Systematically profiling and annotating long intergenic non-coding RNAs in human embryonic stem cell. BMC genomics *14 Suppl 5*, S3.
Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome research *22*, 947-956.
Tanno, B., Cesi, V., Vitali, R., Sesti, F., Giuffrida, M.L., Mancini, C., Calabretta, B., and Raschella, G. (2005). Silencing of endogenous IGFBP-5 by micro RNA interference affects proliferation, apoptosis and differentiation of neuroblastoma cells. Cell death and differentiation *12*, 213-223.
Terashima, M., Kobayashi, M., Motomiya, M., Inoue, N., Yoshida, T., Okano, H., Iwasaki, N., Minami, A., and Matsuoka, I. (2010). Analysis of the expression and function of BRINP family genes during neuronal differentiation in mouse embryonic stem cell-derived neural stem cells. Journal of neuroscience research *88*, 1387-1393.
Terranova, R., Yokobayashi, S., Stadler, M.B., Otte, A.P., van Lohuizen, M., Orkin, S.H., and Peters, A.H. (2008). Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. Developmental cell *15*, 668-679.
Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B*., et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.
Tian, D., Sun, S., and Lee, J.T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell *143*, 390-403.
Trowbridge, J.J., Snow, J.W., Kim, J., and Orkin, S.H. (2009). DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. Cell stem cell *5*, 442-449.
Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689-693.
Tufarelli, C., Stanley, J.A., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., and Higgs, D.R. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nature genetics *34*, 157-165.
Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537-1550.
Ullah, Z., de Renty, C., and DePamphilis, M.L. (2011). Checkpoint kinase 1 prevents cell cycle exit linked to terminal cell differentiation. Molecular and cellular biology *31*, 4129-4143.

Urano, T., Yashiroda, H., Muraoka, M., Tanaka, K., Hosoi, T., Inoue, S., Ouchi, Y., and Toyoshima, H. (1999). p57(Kip2) is degraded through the proteasome in osteoblasts stimulated to proliferation by transforming growth factor beta1. The Journal of biological chemistry *274*, 12197-12200.
Vallot, C., Huret, C., Lesecque, Y., Resch, A., Oudrhiri, N., Bennaceur-Griscelli, A., Duret, L., and Rougeulle, C. (2013). XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. Nature genetics *45*, 239-241.
van de Hoef, D.L., Bonner, J.M., and Boulianne, G.L. (2013). FKBP14 is an essential gene that regulates Presenilin protein levels and Notch signaling in Drosophila. Development *140*, 810-819.
van Dijk, E.L., Chen, C.L., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Ne, P., Loeillet, S.*, et al.* (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature *475*, 114-117.
van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., Macinnes, A.W., Cuppen, E., and Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome biology *15*, R6.
Vance, K.W., and Ponting, C.P. (2014a). Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends in genetics : TIG *30*, 348-355.
Vance, K.W., and Ponting, C.P. (2014b). Transcriptional regulatory functions of nuclear long noncoding RNAs. Trends in genetics : TIG.
Vance, K.W., Sansom, S.N., Lee, S., Chalei, V., Kong, L., Cooper, S.E., Oliver, P.L., and Ponting, C.P. (2014). The long non-coding RNA Paupar regulates the expression of both local and distal genes. The EMBO journal *33*, 296-311.
Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic acids research *35*, D88-92.
Vogel, T., Ahrens, S., Buttner, N., and Krieglstein, K. (2010). Transforming growth factor beta promotes neuronal cell fate of mouse cortical and hippocampal progenitors in vitro and in vivo: identification of Nedd9 as an essential signaling component. Cereb Cortex *20*, 661-671.
Wang, G.Z., Lercher, M.J., and Hurst, L.D. (2011a). Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. Genome biology and evolution *3*, 320-331.
Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R.*, et al.* (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome research *22*, 1680-1688.
Wang, J., Gong, C., and Maquat, L.E. (2013a). Control of myogenesis by rodent SINE-containing lncRNAs. Genes & development *27*, 793-804.
Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F., and Fan, Q. (2010). CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic acids research *38*, 5366-5383.
Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Molecular cell *43*, 904-914.
Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A.*, et al.* (2011b). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120-124.
Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126-130.

Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X., and Liu, H. (2013b). Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. Developmental cell *25*, 69-80.
Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics *10*, 57-63.
Wang, Z., Liu, D.X., Wang, F.W., Zhang, Q., Du, Z.X., Zhan, J.M., Yuan, Q.H., Ling, E.A., and Hao, A.J. (2013c). L-Cysteine promotes the proliferation and differentiation of neural stem cells via the CBS/H(2)S pathway. Neuroscience *237*, 106-117.
Watanabe, K., Kamiya, D., Nishiyama, A., Katayama, T., Nozaki, S., Kawasaki, H., Watanabe, Y., Mizuseki, K., and Sasai, Y. (2005). Directed differentiation of telencephalic precursors from embryonic stem cells. Nature neuroscience *8*, 288-296.
Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics *39*, 457-466.
Weber, M., and Schubeler, D. (2007). Genomic patterns of DNA methylation: targets and function of an epigenetic mark. Current opinion in cell biology *19*, 273-280.
West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E. (2014). The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites. Molecular cell.
White, E., Schlackow, M., Kamieniarz-Gdula, K., Proudfoot, N.J., and Gullerova, M. (2014). Human nuclear Dicer restricts the deleterious accumulation of endogenous double-stranded RNA. Nature structural & molecular biology *21*, 552-559.
Wood, A.J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C.V., Peters, J., Bourc'his, D., and Oakey, R.J. (2008). Regulation of alternative polyadenylation by genomic imprinting. Genes & development *22*, 1141-1146.
Wu, J.I., Lessard, J., Olave, I.A., Qiu, Z., Ghosh, A., Graef, I.A., and Crabtree, G.R. (2007). Regulation of dendritic development by neuron-specific chromatin remodeling complexes. Neuron *56*, 94-108.
Wu, X., and Sharp, P.A. (2013). Divergent transcription: a driving force for new gene origination? Cell *155*, 990-996.
Xiong, Z., and Laird, P.W. (1997). COBRA: a sensitive and quantitative DNA methylation assay. Nucleic acids research *25*, 2532-2534.
Xu, A.G., He, L., Li, Z., Xu, Y., Li, M., Fu, X., Yan, Z., Yuan, Y., Menzel, C., Li, N.*, et al.* (2010). Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. PLoS computational biology *6*, e1000843.
Yang, L., Lin, C., Jin, C., Yang, J.C., Tanasa, B., Li, W., Merkurjev, D., Ohgi, K.A., Meng, D., Zhang, J.*, et al.* (2013). lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature *500*, 598-602.
Yang, L., Lin, C., Liu, W., Zhang, J., Ohgi, K.A., Grinstein, J.D., Dorrestein, P.C., and Rosenfeld, M.G. (2011). ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell *147*, 773-788.
Yang, Y.W., Flynn, R.A., Chen, Y., Qu, K., Wan, B., Wang, K.C., Lei, M., and Chang, H.Y. (2014). Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. eLife *3*, e02046.
Yang, Z., Zhu, Q., Luo, K., and Zhou, Q. (2001). The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature *414*, 317-322.

Yap, K.L., Li, S., Munoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Molecular cell *38*, 662-674.
Yin, T., Cook, D., and Lawrence, M. (2012). ggbio: an R package for extending the grammar of graphics for genomic data. Genome biology *13*, R77.
Yoo, E.J., Cooke, N.E., and Liebhaber, S.A. (2012). An RNA-independent linkage of noncoding transcription to long-range enhancer function. Molecular and cellular biology *32*, 2020-2029.
Yoon, J.H., Abdelmohsen, K., and Gorospe, M. (2013). Posttranscriptional gene regulation by long noncoding RNA. Journal of molecular biology *425*, 3723-3730.
Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., and Ponting, C.P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. Genome biology and evolution *4*, 427-442.
Young, T.L., Matsuda, T., and Cepko, C.L. (2005). The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. Current biology : CB *15*, 501-512.
Youngman, E.M., and Green, R. (2005). Affinity purification of in vivo-assembled ribosomes for in vitro biochemical analysis. Methods *36*, 305-312.
Yu, D.H., Ware, C., Waterland, R.A., Zhang, J., Chen, M.H., Gadkari, M., Kunde-Ramamoorthy, G., Nosavanh, L.M., and Shen, L. (2013). Developmentally programmed 3' CpG island methylation confers tissue- and cell-type-specific transcriptional activation. Molecular and cellular biology *33*, 1845-1858.
Yu, G., Zerucha, T., Ekker, M., and Rubenstein, J.L. (2001). Evidence that GRIP, a PDZ-domain protein which is expressed in the embryonic forebrain, co-activates transcription with DLX homeodomain proteins. Brain research Developmental brain research *130*, 217-230.
Zampieri, M., Guastafierro, T., Calabrese, R., Ciccarone, F., Bacalini, M.G., Reale, A., Perilli, M., Passananti, C., and Caiafa, P. (2012). ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. The Biochemical journal *441*, 645-652.
Zhang, B., Arun, G., Mao, Y.S., Lazar, Z., Hung, G., Bhattacharjee, G., Xiao, X., Booth, C.J., Wu, J., Zhang, C*., et al.* (2012). The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. Cell reports *2*, 111-123.
Zhang, Q., Chen, C.Y., Yedavalli, V.S., and Jeang, K.T. (2013). NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. mBio *4*, e00596-00512.
Zhang, Q., Wang, H.Y., Marzec, M., Raghunath, P.N., Nagasawa, T., and Wasik, M.A. (2005). STAT3- and DNA methyltransferase 1-mediated epigenetic silencing of SHP-1 tyrosine phosphatase tumor suppressor gene in malignant T lymphocytes. Proceedings of the National Academy of Sciences of the United States of America *102*, 6948-6953.
Zhang, X., Lian, Z., Padden, C., Gerstein, M.B., Rozowsky, J., Snyder, M., Gingeras, T.R., Kapranov, P., Weissman, S.M., and Newburger, P.E. (2009). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood *113*, 2526-2534.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W*., et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.
Zhao, F.Q. (2013). Octamer-binding transcription factors: genomics and functions. Front Biosci (Landmark Ed) *18*, 1051-1071.
Zhao, H., Luo, Y., Liu, X., Wang, R., Yan, F., Li, S., Leak, R.K., and Ji, X. (2013). Ischemic post-conditioning partially reverses cell cycle reactivity following ischemia/reperfusion injury: a genome-wide survey. CNS & neurological disorders drug targets *12*, 350-359.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Molecular cell *40*, 939-953.
Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750-756.
Zhong, J., Chuang, S.C., Bianchi, R., Zhao, W., Lee, H., Fenton, A.A., Wong, R.K., and Tiedge, H. (2009). BC1 regulation of metabotropic glutamate receptor-mediated neuronal excitability. The Journal of neuroscience : the official journal of the Society for Neuroscience *29*, 9977-9986.
Zhou, X., and O'Shea, E.K. (2011). Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. Molecular cell *42*, 826-836.
Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. Trends in genetics : TIG *24*, 481-484.
Ziller, M.J., Muller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C.B., Bernstein, B.E., Lengauer, T*., et al.* (2011). Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS genetics *7*, e1002389.

Appendix

Appendix Table 1	Transient *Dali* knockdown
Appendix Table 2	Transient *Dali* knockdown GO analysis
Appendix Table 3	Transient *Pou3f3* knockdown
Appendix Table 4	Transient *Pou3f3* knockdown GO analysis
Appendix Table 5	Genes shared between transient *Dali* and *Pou3f3* knockdowns
Appendix Table 6	Stable *Dali* knockdown - proliferating cells
Appendix Table 7	Stable *Dali* knockdown - proliferating cells (GO analysis)
Appendix Table 8	RA differentiation – control cells
Appendix Table 9	RA differentiation – *Dali* knockdown cells
Appendix Table 10	RA differentiation – control cells (GO analysis)
Appendix Table 11	RA differentiation – *Dali* knockdown cells (GO analysis)
Appendix Table 12	Stable *Dali* knockdown – RA-differentiated cells
Appendix Table 13	Stable *Dali* knockdown – RA-differentiated cells (GO analysis)
Appendix Table 14	Stable *Dali* knockdown – differential responders
Appendix Table 15	Transient *Dali* overexpression
Appendix Table 16	Genes shared between transient *Dali* knockdwon, *Pou3f3* knockdown and *Dali* over-expression
Appendix Table 17	*Dali* CHART-Seq – peaks against input
Appendix Table 18	*Dali* CHART-Seq – peaks against input and *lacZ*
Appendix Table 19	*Dali* CHART-Seq – peak-associated genes
Appendix Table 20	All genes regulated by *Dali*
Appendix Table 21	*Dali* CHART-Seq – peak-associated genes (GO analysis)

All Apendix Tables are available at http://wwwfgu.anat.ox.ac.uk/~vladislava.