Role of the chromatin remodeler ATRX in the regulation of gene expression

Julia Truch

St Edmund Hall



A thesis submitted for the degree of Doctor of Philosophy in Medical Sciences of the University of Oxford

MRC Molecular Haematology Unit
Weatherall Institute of Molecular Medicine
University of Oxford

Trinity Term 2018

To Cyril,
For his unconditional support.

Abstract

Julia Truch
St Edmund Hall

Doctor of Philosophy
Trinity Term 2018

Role of the chromatin remodeler ATRX in the regulation of gene expression

The ATR-X syndrome is an inherited, X-linked condition associated with alpha-thalassaemia and a variety of developmental abnormalities including profound intellectual disabilities. It results from mutations in a chromatin remodeling factor (called ATRX) which has been shown to perturb a wide range of nuclear activities including replication, transcription and DNA methylation. Although ATRX is known to play a role in the turnover of the histone variant H3.3, the precise role of ATRX in each of these nuclear activities and how ATRX mutations affect phenotype, are poorly understood. Previous studies have shown that ATRX is associated with repeated regions, including pericentric heterochromatin, telomeres, and rDNA repeats, some of which have the propensity to form G-quadruplexes. However, within euchromatin very little is known about ATRX target genes and how their expression is affected by this protein. The two major obstacles to answering these questions are first, the poor quality of ATRX chromatin immunoprecipitation sequencing (ChIP-seq) protocols; and second the lack of a well-defined, tractable cell model to investigate the role of ATRX on the expression of its targets in normal and affected individuals.

In this thesis I studied the role of ATRX in lymphoblastoid cell lines (LCL) as they are readily available from patients with ATR-X syndrome. First, I successfully optimised the protocol for ATRX ChIP-seq. This showed that ATRX is also found at regions of open chromatin including a subset of active promoters and enhancers. Second, using LCL from ATR-X patients and unaffected controls, I identified more than a hundred of candidate genes whose expression is affected by ATRX mutations. Intersecting these two datasets I concluded that more than fifty genes whose expression is altered by mutations in ATRX are probably primary targets of this protein including zinc finger transcription factors. I next assessed these findings using erythroid cells, which are of relevance since mutations in ATRX cause alpha-thalassaemia; a red cell phenotype caused by downregulation of alpha globin expression. Using differentiated human CD34+ cells, I confirmed the binding of ATRX at repeats within the alpha globin cluster but, reflecting the findings in LCL, I also identified ATRX enrichment at *HBA1* and *HBA2* genes and their associated enhancers suggesting that ATRX at these newly identified sites may normally contribute to the regulation of alpha globin expression.

In conclusion, these high quality ATRX-ChIP datasets allowed me to identify new ATRX targets in human samples. The new set of genes affected by ATRX mutations, within their associated chromatin environment, offers the opportunity to further understand the role of ATRX in heath and disease. The findings show for the first time that ATRX is significantly bound at active euchromatic regions with rapid nucleosome turnover, including active enhancers, in human cells.

Acknowledgement

I would like to express my gratitude to my supervisors Doug Higgs and Richard Gibbons for giving me the opportunity to work in the lab and their support and guidance throughout the project over the last four years. I am also indebted to Jim Hughes who guided my unexpected transition from banch to bioinformatics and for the discussions and advice. I also acknowledge Marella de Bruijn who mentored me to improve my presentation skills. This work would not have been possible without funding from the Medical Research Council UK.

I am grateful to Chris Fisher and David Clynes for their help and advice in the lab and also to Yavor Bozhilov and Anthony Cheong for their contribution to data in Chapter 6 and for the discussions. Thanks also go to Caroline Scott for her contribution to data in chapter 7 and her support.
This thesis has been informed by numerous suggestions from the members of the Gibbons and Higgs laboratories, both past and present: Barbara Xella, Diu Nguyen, Sachith Mattananda, Martin Larke, James Davies, Bryony Graham, Anna Rose, Danuta Jeziorska, Rob Beagrie, Tom Kent to name but a few. I am grateful to Sue Butler for her invaluable assistance with tissue culture. I would like to thank the Hughes group who hosted me and guided me through many experiments, especially Jelena Telenius and Ron Schwessinger for their support in bioinformatics and Damien Downes for his help and advice with the Capture-C experiment.

I would like to thank the Centre for Computational Biology in particular Steve Tailor for his help with the human repeatome analysis and Emmanuela Repapi for the analysis of the microarray experiment. I am also grateful to the Klose laboratory for providing me with the Bio-CAP system and the Buckle and Milne laboratories for discussions and advice.

To my friends I have made during my time in Oxford and to those that have been there even before I applied for this DPhil who I feel lucky to have had through the course of my DPhil.

Finally, my heartfelt gratitude goes to my partner Cyril who has been by my side over the last nine years and always finds a way to make me smile. None of this would have been possible without his support, caring and love.

Merci

List of abbreviations

aa	Amino acid
ADD	ATRX-DNMT3-DNMT3L
ALT	Alternative lengthening of telomere
aPAR	Ancestral pseudoautosomal region
ATAC-seq	Assay for transposase-accessible chromatin - high-throughput sequencing
ATMDS	Acquired alpha-thalassemia myelodysplastic syndrome
ATP	Adenosine triphosphate
ATRX	Alpha-thalassemia mental retardation X-linked
ATRXt	ATRX truncated form
Bio-CAP	Biotinylated CxxC affinitty purification
BSA	Bovine serum albumin
C2H2-ZNF	C2H2 zinc finger
ChIP-CHIP	Chromatin immunoprecipitation - microarray hybridization
ChIP-seq	Chromatin immunoprecipitation - high-throughput sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
D10 dCD34+	Day 10 differentiated CD34+ cells
DAXX	Death-associated protein 6
DBM	DAXX binding domain
DDR	DNA damage response
DEG	Differentially expressed genes
DIP	DNA immunoprecipitation
DMR	Differentially methylated regions
DMSO	Dimethyl-Sulphoxide
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
ds	double stranded
EGS	Ethylene glycol bis(succinimidylsuccinate)
ENCODE	Encyclopedia of DNA elements
ERV	Endogenous retroviral elements

ES	Embryonic stem
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
G4	G-quadruplexes
gDNA	Genomic DNA
GO	Gene ontology
HbH	Hemoglobin H
HMM	Hidden Markov model
HR	Homologous recombination
HUDEP	Human umbilical cord blood-derived erythroid progenitor
IAP	Intracisternal A-type particles
ICR	Imprinting control region
iPSC	Induced pluripotent stem cells
KD	Knock-down
KO	Knock-out
KRAB	Kruppel-associated box
LCL	Lymphoblastoid cell lines
LCR	Locus control region
logFC	Log2 fold change
LTR	Long terminal repeat
MEF	Mouse embryonic fibroblast
NLS	Nuclear localisation signals
NMD	Nonsens mediated decay
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PCH	Pericentromeric heterochromatin
PHD	Plant homeodomain
PML-NB	Promyelocytic leukemia nuclear bodies
PQS	Putative quadruplex sequences
PRC	Polycomb repressive complex

PrimPol	Primase-polymerase
PTM	Post-translation modifications
qPCR	Quantitative polymerase chain reaction
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
RT	Reverse transcription
SEM	Standard error of the mean
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphisms
ss	Single stranded
SV40	Simian virus 40
TAD	Topologically associated domains
TERRA	Telomeric Repeat-containing RNA
TES	Transcription end site
TFBS	Transcription factor binding sites
TIF	Telomere induced dysfunction
TSS	Transcription start site
UCSC	University of California Santa Cruz
VNTR	Variable number tandem repeat
WT	Wild type
ZNF	Zinc finger

Table of contents

Chapter 1. Introduction 1
1.1. Eukaryotic chromatin architecture and regulation of gene expression 1
1.1.1. Histone variants 2
1.1.2. Histone modifications 3
1.1.3. DNA methylation 5
1.1.4. Blurry territories 5
1.1.5. Nucleosome occupancy 6
1.1.6. Regulation of transcription 6
1.1.7. Writers, readers and erasers 8
1.1.8. Histone chaperones and chromatin remodeling factors 9
1.2. The chromatin remodeling factor ATRX 10
1.2.1. Gene and protein structures of the chromatin remodeler ATRX 10
1.2.2. Remodeling and translocation activity of ATRX and ATRX-DAXX complex 14
1.2.3. Localisation and recruitment of ATRX beyond chromatin 16
1.2.4. Localisation and recruitment of ATRX within chromatin 16
1.2.5. ATRX enrichment at pericentromeric heterochromatin (PCH) 17
1.2.6. ATRX association with telomeres, rDNA and other GC-rich sequences 19
1.2.7. ATRX occupancy at the 3' exon of Zinc Fingers 23
1.3. ATRX functions in health and disease 24
1.3.1. ATRX-associated disorders related to alpha-thalassaemia 24
1.3.2. Deposition of H3.3 28
1.3.3. ATRX-DAXX-H3.3 role at retrotransposons 33
1.3.4. Maintenance of H3K9me3 level at the 3' exon of ZNF genes 34
1.3.5. ATRX-mediated chromatin association of macroH2A1 35
1.3.6. ATRX and regulation of ancestral pseudoautosomal region gene expression 36
1.3.7. ATRX effects on gene expression, DNA replication and genomic stability 36
1.3.8. ATRX deficiency and alternative lengthening of telomere associated cancers 41
1.4. Summary, aim and objectives 43

Chapter 2. Materials and Methods 45
2.1. Cell culture 45
2.1.1. Lymphoblastoid cell lines 45
2.1.2. CD34 separation 46
2.1.3. Culturing of CD34+ 46
2.2. DNA extraction 47
2.3. RNA extraction 48
2.4. Protein extraction and Western blot 48
2.5. Southern blot 49
2.6. Primeview Affymetrix Microarray 49
2.7. Power Calculation and Correction for the Batch effects in microarray experiment analyses 50
2.8. High-throuput qRT-PCR by Fluidigm Biomark HD system 50
2.9. Cell line validation 52

2.10. Quantitative-PCR 53
2.11. Reverse transcription 53
2.12. ATRX ChIP-seq 54
2.13. Single step cross-linking ChIP 56
2.14. Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) 58
2.15. Biotinylated CxxC affinity purification (Bio-CAP) 59
2.16. RNA-DNA immunoprecipitation (DIP) 59
2.17. G4 ChIP 61
2.18. Capture-C 62
2.19. Library preparation 64
2.20. Library quantification and High throughput sequencing 64
2.21. CRISPR-Cas9 genome editing 64
2.22. Analysis of raw sequencing data from ChIP-seq, ATAC-seq and Bio-CAP-seq experiments 65
2.23. Capture-C probes designes and data analysis 65
2.24. Peak calling 65
2.25. Gene ontology and motif analyses 66
2.26. Differential enrichment analysis 66
2.27. Graphical analysis 67
2.28. Repeatome Analysis 67
2.29. Protein sequence analysis 67
2.30. RefSeq and CpG island annotation 67
2.31. GenoSTAN analysis 68
2.32. Transcription factor binding site analysis 68
2.33. Visualisation of gene expression in LCL 68

Chapter 3. ATRX ChIP-seqs and peak calling 69

3.1. Introduction 69
3.2. Results 71
3.2.1. Choice of the system and validation of the cell lines 71
3.2.2. Choice and validation of the anti-ATRX antibodies 75
3.2.3. Optimisation of the ATRX ChIP-seq protocol 76
3.2.4. Generation of robust ATRX ChIP-seq datasets in LCL 79
3.2.5. Genome-wide ATRX distribution; a challenge for peak calling 84
3.3. Discussion 88

Chapter 4. Genome-wide characterisation of the ATRX binding sites 92

4.1. Introduction 92
4.2. Results 94
4.2.1. Analysis of ATRX enrichment at repetitive elements using direct mapping of ChIP-seq data 94
4.2.2. Genome Ontology analysis of ATRX binding sites 96
4.2.3. ATRX is enriched at promoters and gene bodies 97
4.2.4. ATRX interacts with CpG island containing regions 98
4.2.5. Almost 50% of the ATRX binding sites contain putative G4 forming sequences 102

4.2.6. The majority of the ATRX binding sites contain nucleosome-depleted regions 104
4.2.7. Almost 95% of the ATRX binding sites contained at least one TFBS..... 105
4.2.8. Motif analysis conducted on ATRX binding sites identified consensus binding sequences associated with transcription factors..... 108
4.2.9. ATRX binding sites are enriched in active regulatory elements..... 111
4.2.10. ATRX binds functional enhancer clusters of the MYC gene..... 113
4.3. Discussion..... 116

Chapter 5. Identification of novel genes affected by ATRX mutations 123
5.1. Introduction..... 123
5.2. Results..... 124
5.2.1. Gene expression analysis of patient-derived LCL using microarrays..... 124
5.2.2. Assessing the microarray data in the context of published work..... 128
5.2.3. Assessing the effects of ATRX mutations based on their location..... 130
5.2.4. Identification of candidate genes potentially directly affected by ATRX mutations 133
5.2.5. Validation of the microarray..... 136
5.2.6. Downregulated ZNF genes with intragenic ATRX binding sites are enriched in members of the KRAB-containing C2H2-ZNF subfamily..... 137
5.3. Discussion..... 141

Chapter 6. Investigation of the effects of ATRX mutations on chromatin landscape 145
6.1. Introduction..... 145
6.2. Results..... 147
6.2.1. ATRX at imprinted genes..... 147
6.2.2. ATRX and aPAR gene expression..... 150
6.2.3. ATRX at the NME4 Locus..... 151
6.2.4. ATRX mutations correlate with changes in chromatin environment at regulatory elements associated with a subset of DEG..... 151
6.2.5. Chromatin interactions help discriminate between DEG that are affected directly or indirectly by ATRX..... 161
6.2.6. ATRX mutations are associated with changes in chromatin accessibility..... 163
6.2.7. DNA methylation is not sufficient to explain the downregulation observed at certain loci in patient-derived cells..... 175
6.2.8. The presence of R-loops is not sufficient to explain ATRX recruitment to the DEG 181
6.2.9. Assessing the presence of G4 in vivo..... 183
6.2.10. The ATRX affinity for ZNF genes..... 185
6.2.11. Recapitulation of the effects observed in ATRX deficient cellls via ATRX KO in control LCL and alternatives..... 186
6.3. Discussion..... 189

Chapter 7. The ATRX distribution in human differentiated CD34+ cells 198
7.1. Introduction..... 198
7.2. Results..... 200
7.2.1. Identification of the ATRX binding sites in differentiated CD34+ cells..... 200

7.2.2. ATRX is particularly enriched at certain repetitive elements 202
7.2.3. ATRX is enriched at CpG island-containing promoters and gene bodies 203
7.2.4. A third of the ATRX binding sites contains putative G4 forming sequences. 205
7.2.5. More than half of the ATRX binding sites contain nucleosome-depleted regions 206
7.2.6. Open chromatin regions that contain ATRX binding sites are enriched in active regulatory elements 209
7.2.7. Most ATRX binding sites conserved between both cell types share a similar chromatin accessibility pattern 210
7.2.8. ATRX enrichment at ZNF genes is conserved across cell types 213
7.2.9. ATRX binding sites display distinct motifs according to the cell type 213
7.2.10. The ATRX enrichment at the globin clusters is cell type-dependent 215
7.3. Discussion 220

Chapter 8. General discussion and conclusions 225
8.1. Optimisation of the ATRX ChIP-seq protocol places ATRX in euchromatin 225
8.2. Pathogenic ATRX mutations are associated with limited changes in gene expression 227
8.3. ATRX mutations are associated with changes in chromatin environment at enhancers and promoters linked to DEG 228
8.4. ATRX is enriched at the ψζ VNTR, *HBA1* and *HBA2,* and their associated enhancers in erythroid cells 229
8.5. Does ATRX deficiency lead to the loss of epigenetic memory? 229
8.6. ATRX-dependent deposition of H3.3 at euchromatin regions? 231
8.7. ATRX recruitment: different targets, different mechanisms? 232
8.8. Conclusions 234

References 237

List of figures

Figure 1.1: Schematic representation of the human full-length ATRX protein..... 11

Figure 3.1: Family tree illustrating the case-control pairs and localisation of the associated ATRX mutations..... 72

Figure 3.2:Validation of the LCL by Sanger sequencing. 74

Figure 3.3: Characterisation of the LCL used for ChIP-seq using western blot. 74

Figure 3.4: Validation of the anti-ATRX antibodies which specifically detect the full-length ATRX isoform..... 75

Figure 3.5: Sonication of double cross-linked LCL samples using the Bioruptor. 77

Figure 3.6: Sonication of double cross-linked LCL samples using the Covaris..... 78

Figure 3.7: Screenshot of the hg19 UCSC genome browser showing the ATRX ChIP-seq experiments performed in LCL..... 81

Figure 3.8: Screenshot of the hg19 UCSC genome browser showing the improvements of the ATRX ChIP-seq experiments performed in LCL.. 82

Figure 3.9: PCA analysis of the ATRX ChIP-seq experiments.. 83

Figure 3.10: Validation of ATRX targets by ChIP-qPCR..... 84

Figure 3.11: Examples of heterogeneity of the ATRX enrichment across the genome and challenges of peak calling in ATRX ChIP-seq datasets..... 87

Figure 4.1: Analysis of ATRX binding at repetitive elements using direct mapping of the ATRX ChIP-seqs and related inputs. 95

Figure 4.2:Distribution of the ATRX binding sites relative to genes..... 98

Figure 4.3: ATRX and annotated CpG islands.. 99

Figure 4.4: Analysis of the CpG clusters using Bio-CAP. 101

Figure 4.5: Analysis of putative G4 forming sequences at ATRX binding sites. 103

Figure 4.6: Analysis of chromatin accessibility at ATRX binding sites. 105

Figure 4.7: Analysis of TFBS relative to ATRX binding sites..... 107

Figure 4.8: Motif analysis carried out on all ATRX binding sites to identified enrichment in know motifs..... 108

Figure 4.9: Motif analysis carried out on ATRX binding sites containing at least one open chromatin region and using the ATAC-seq peaks not associated with ATRX as background. 109

Figure 4.10: Motif analysis carried out on ATRX binding sites, which did not contain open chromatin regions, and using random sequences from the human genome as background. 110

Figure 4.11: Identification of regulatory elements at ATRX binding sites. 113

Figure 4.12: ATRX enrichment around the MYC locus..... 115

Figure 5.1: Schematic representation of the full-length ATRX protein depicting the position of the ATRX mutations present in the selected LCL derived from ATR-X patients..... 125

Figure 5.2: Power calculation analysis to detect 1.5-fold change..... 126

Figure 5.3: Analysis of the DEG obtained by microarray-based experiment..... 127

Figure 5.4: Analysis of the microarray data associated with ATRX, NME4 and DECR2 genes 129

Figure 5.5: Comparison of ATRX mutations based on their occurrence in patients and their location. 132

Figure 5.6: Identification of DEG potentially directly affected by ATRX mutations in LCL. 135

Figure 5.7: Example of inaccurate annotation of the gene position in the Affymetrix probe 11738632_s_at. 136

Figure 5.8: Validation of the microarray experiment 137

Figure 5.9: Characterisation of the differentially expressed ZNF genes. 139

Figure 5.10: Differentially expressed ZNF genes not belonging to the C2H2-ZNF superfamily 140

Figure 5.11: Variation in ATRX enrichment between differentially expressed ZNF containing, or not, KRAB and C2H2 domains. 140

Figure 6.1: Example of imprinted genes (NDM and MKRN3) that are dysregulated in ATR-X patient derived LCL showing an absence of significant ATRX enrichment. 148

Figure 6.2: Analysis of ATRX localisation and chromatin interaction associated with the LILRB4 locus. 149

Figure 6.3: Expression analysis at a subset of aPAR genes in LCL using RT-qPCR. 150

Figure 6.4: Changes in chromatin environment at the NME4 locus in LCL. 151

Figure 6.5: Comparison of the chromatin environment at the CCR1 locus between LCL derived from unaffected controls and ATR-X patients 154

Figure 6.6: Capture-C experiment at the MEIS2 locus 154

Figure 6.7: Capture-C experiments showing the interactions associated with the CCR1 promoter 155

Figure 6.8: Characterisation of the chromatin environment at the UST locus in controls and cases. 157

Figure 6.9: Complex chromatin interactions occurred across the UST locus 159

Figure 6.10: Chromatin environment at the PBX4 locus. 160

Figure 6.11: Chromatin environment at the BAIAP2L1 locus. 162

Figure 6.12: Diffbind analysis assessing for the difference in chromatin accessibility genome-wide between cases and controls. 163

Figure 6.13:ATF7IP2 chromatin environment in LCL. 165

Figure 6.14: ChIP-qPCR analysing the changes in chromatin environment observed by ChIP-seq at the ATF7IP2 promoter. 166

Figure 6.15: Chromatin interactions involving the ATF7IP2 promoter region. 167

Figure 6.16: Chromatin environment at the ZNF595-ZNF718 subtelomeric region. 169

Figure 6.17: Analysis of the chromatin interactions associated with the ZNF718 promoter at the subtelomeric regions of the short arm of chromosome 4. 171

Figure 6.18: Chromatin environment at the ZNF555 - ZNF57 region 173

Figure 6.19: Analysis of the interactions associated with the promoters of ZNF555 and ZNF57. 174

Figure 6.20: Southern blot experiment illustrating the DNA methylation status of the CpG island-containing promoter of ATF7IP2 in LCL. *176*

Figure 6.21: Southern blot experiments assessing the DNA methylation status at promoter regions enriched in ATRX at ZNF555 gene (a) and ZNF57 gene (b) *178*

Figure 6.22: Analysis of the DNA methylation status of CpG island-containing promoters by Bio-CAP experiments in LCL. *180*

Figure 6.23: Analysis of R-loop formation based on DIP-qPCR at a subset of DEG promoters enriched in ATRX. *182*

Figure 6.24: Analysis of G4 in vivo. *184*

Figure 6.25: Gene expression analysis using RT-qPCR in iPSC either WT or KO for ATRX. *186*

Figure 6.26: Analysis of gene expression and chromatin environment in HUDEP cells with ATRX WT or KO. *188*

Figure 7.1: Original ATRX ChIP-seq performed in human erythroid cells by Law and colleagues at the alpha globin cluster. *200*

Figure 7.2: Illustration of the different resolutions of the ATRX peak calls obtained between D10 dCD34+ and LCL. *202*

Figure 7.3: Selected representative repetitive elements that show ATRX enrichment using direct mapping of the ATRX ChIP-seqs and related inputs. *203*

Figure 7.4: Distribution of the ATRX binding sites in D10 dCD34+ cells. *204*

Figure 7.5: Analysis of putative G4-forming sequences at ATRX binding sites. *206*

Figure 7.6: Analysis of chromatin accessibility at ATRX binding sites. *208*

Figure 7.7: Identification of regulatory elements at ATRX binding sites in D10 dCD34+ cells using GenoSTan. *210*

Figure 7.8: Analysis of the ATRX binding sites conserved between LCL and D10 dCD34+ cells. *212*

Figure 7.9: Variations in chromatin environment between D10 dCD34+ cells and LCL at the NME4 locus. *216*

Figure 7.10: ATRX enrichment across the alpha globin cluster in D10 dCD34+ cells and LCL. *218*

Figure 7.11: ATRX enrichment across the beta globin cluster in D10 dCD34+ cells and LCL. *219*

Figure 8.1: Hypothesis of ATRX interactions and dynamics depending on the change in chromatin environment. *236*

List of tables

Table 2.1: List of the LCL used in this study. *46*

Table 2.2: list of the FACS markers used to monitor the CD34+ differentiation *47*

Table 2.3: Taqman assays used in this study *51*

Table 2.4 list of primers and their localisation in the ATRX gene used in this study to validate the different cell lines. *52*

Table 2.5: Primers used for qPCR based on SYBR green assays. *53*

Table 2.6: List of the antibodies used for single cross-link ChIP protocol *57*

Table 4.1: Top terms of genome ontology highlighting a strong association between the ATRX binding sites and ZNF genes. *96*

Table 4.2: Top terms of genome ontology associated with biological processes *96*

Table 4.3: Top terms of genome ontology associated with molecular function *96*

Table 4.4: Top terms of genome ontology associated with repetitive regions *97*

Table 5.1: DAVID gene ontology analysis of biological processes. *128*

Table 5.2: Characteristics of the differentially expressed ZNF genes belonging to the C2H2-ZNF superfamily *139*

Table 6.1: List of imprinted genes identified as DEG with |logFC| > 0.5. ..... *148*

Table 6.2: List of downregulated DEG sharing similar characteristics including ATRX enrichment over their promoter region. *168*

Table 7.1: Motif analysis of ATRX binding sites to identify enrichment in know motifs. *214*

Publication

TRUCH, J., TELENIUS, J., HIGGS, D. R. & GIBBONS, R. J. 2018. How to Tackle Challenging ChIP-Seq, with Long-Range Cross-Linking, Using ATRX as an Example. *Histone Variants.* Springer.

Chapter 1. Introduction

1.1. Eukaryotic chromatin architecture and regulation of gene expression

With more than 3 billion base pairs (bp), representing approximately two metres of deoxyribonucleic acid (DNA), the human genome needs to be packed and organised to accommodate this vast amount of DNA in micrometre sized cell nuclei (Iyer et al., 2011). Therefore, the DNA is condensed into chromatin using nucleosomes as basic packaging unit structures. Each nucleosome consists of the approximately 146 bp of DNA wrapped around a core histone octamer comprised of two copies of histone H2A, H2B, H3 and H4 proteins (Richmond et al., 1984). These small proteins are composed of a globular histone-fold region and an unstructured N-terminal tail. Whereas the histone-fold domains interact with each other to form the core globular domain of the nucleosome, the tails extend beyond this boundary (Luger and Richmond, 1998). Prior to octamer assembly, histone H3 and H4 form a tetramer $[H3\text{-}H4]_2$ and H2A and H2B associate to form two heterodimers [H2A-H2B]. Then, $[H3\text{-}H4]_2$ are subsequently assembled with two H2A-H2B dimers to form the octamer. Successions of 11 nm nucleosome core particles that are connected by linker DNA form a beads-on-string structure. This structure is further folded with the help of other proteins, including the linker histone H1, to achieve thousands of fold compaction as reviewed in (Cutter and Hayes, 2015).

The beauty of this process is that it not only allows the compaction of the long double stranded (ds) DNA but it is also an integral part of the regulation of gene expression. Indeed, it is now clear that the sole analysis of the linear sequence is not sufficient to understand how a single genetic code gives rise to complex individuals composed of

highly organised specialised cells that display diverse forms and functions depending on their gene expression. Far from begin a uniform structure, the chromatin is highly dynamic and basically, subdivided in heterochromatin and euchromatin. Heterochromatin is usually associated with condensed, gene-poor and transcriptionally silent regions. Conversely, the euchromatin displays a less condensed structure containing more gene-rich and transcriptionally permissive regions (Gilbert et al., 2004, Huisinga et al., 2006). Histone exchange and turn over, including removal or exchange of part or all histone components, are essential for chromatin fluidity, composition, structure and function of different parts of the genome.

1.1.1. Histone variants

Beyond the canonical histones involved in nucleosome formation, chromatin structure also contains a number of histone variants that alter the physical properties and chemical nature of the nucleosomes and facilitate diverse processes such as DNA accessibility, repair, transcription, recombination and replication.

The canonical histones H3.1 and H3.2 are the main H3 histone supply during the DNA replication process. Indeed, they are mainly expressed during the S-phase of the replication and are incorporated into chromatin in a replication-dependent manner. By contrast, the variant H3.3, which differs from canonical H3.1 and H3.2 by only 5 and 4 amino acids (aa), respectively, is constitutively expressed and can be deposited or exchanged throughout the cell cycle (Franklin and Zweidler, 1977, Szenker et al., 2011). Thus, H3.3 deposition is replication-independent. This variant has been associated with active euchromatin and is particularly enriched at transcriptionally active genes and regulatory elements. Nevertheless, it has also been found in heterochromatic regions such as telomeres (Goldberg et al., 2010). Compared to H3.1,

nucleosomes containing H3.3 are intrinsically less stable (Jin et al., 2007). H3.3 is more sensitive to disruption and thus, less energy may be required to evict H3.3-containing nucleosome that may facilitate the transcription.

Another example of an H3 variant is the histone H3-like centromeric protein A or CENP-A, which has been associated with centromeres and is essential for mitosis (Palmer et al., 1987, Black and Cleveland, 2011).

H3 is not the only core histone with variants. Actually, H4 is the only core histone that displays only one form so far as reviewed in (Maze et al., 2014). Similar to H3.3, the H2A histone variant H2A.X can be incorporated in a replication-independent manner. This variant possesses a SQ[E/D] Φ motif (Φ: hydrophobic aa) at its C-terminus and has been associated with DNA damage and genome remodeling after fertilisation (Nashun et al., 2010, Bönisch and Hake, 2012).

Histone variants can greatly vary in size. Indeed, the histone variant macroH2A, which is characterised by the presence of a C-terminal macro-domain, is three times larger than the canonical H2A (Fried, 1992). This variant has been associated with gene repression and is enriched on the inactivated X chromosome in mammalian females (Bönisch and Hake, 2012).

1.1.2. Histone modifications

The incorporation of histone variants is not the only nucleosomal factor that contributes to the chromatin plasticity. Indeed, the unstructured N-terminal tail of histones which protrudes from the globular core of the nucleosome offers an array of possibilities for post-translation modifications (PTM) including methylation, acetylation and phosphorylation as reviewed in (Zentner et al., 2013).

Histone acetylation is a dynamic modification, associated with active chromatin, neutralises the positive charge of lysine residues. Charge neutralisation weakens the charge-dependent interaction between histones and DNA resulting in an increase in DNA accessibility (Phillips, 1963, Pogo et al., 1966, Hong et al., 1993). One of the most studied histone acetylation marks is the acetylation of lysine 27 of the histone 3 or H3K27ac. This modification has been used to differentiate putative active enhancers from poised enhancers (Creyghton et al., 2010).

Histone methylation is another well studied PTM. Interestingly, methylation of lysine residues can be either associated with active or silenced chromatin states depending on the position of the methylated residue. Thereby, trimethylated lysine 36 of histone 3 or H3K36me3, which has been associated with diverse processes such as DNA mismatch repair (Huang et al., 2018), is also considered as a mark of active transcription. Trimethylation of H3K4 is associated with active promoters, whereas monomethylation of H3K4 has been linked with enhancers (Hyun et al., 2017). By contrast, methylation of H3K9 and H3K27 are often associated with heterochromatin and silencing (Fischle et al., 2003, Zentner et al., 2013).

Phosphorylation is widely used in cells for signal transduction. Similar to histone acetylation, phosphorylation of histone residues affects their charge. Indeed, by adding a negative charge, phosphorylation may result in charge repulsion with negatively charged DNA that potentially weakens the histone-DNA interaction (Mirsky et al., 1972, Zentner et al., 2013). Phosphorylation of the histone variant H2A.X is a good illustration of a combination of histone variant and PTM. The characteristic SQ[E/D] Φ motif of the H2A.X variant is essential for the phosphorylation of its serine residue (S139 in human) giving rise to γH2A.X, which is an important mark involved in cellular response to DNA damage (Nashun et al., 2010, Maze et al., 2014).

1.1.3. DNA methylation

The range of epigenetic modifications is not restricted to histones; DNA methylation is also an important player in chromatin regulation. In higher mammals, methylation occurs at CpG dinucleotides and is catalysed by DNA methyltransferase (DNMT) enzymes (Ehrlich et al., 1982, Okano et al., 1999). Interestingly, CpG dinucleotides are globally depleted in mammals and their distribution is not even (Lander et al., 2001). A subset of CpG dinucleotides is concentrated in CG-dense regions known as CpG islands. CpG islands are often present at promoters and tend to be unmethylated at active genes (Saxonov et al., 2006, Long et al., 2013). By contrast, the methylation of promoter containing CpG islands has been associated with gene silencing (e.g. genes located on the inactivated X chromosome) (Venolia and Gartler, 1983, Jones, 2012). However, the absence of methylation at these promoters is not necessarily indicative of active genes as illustrated in the alpha globin cluster in non-erythroid cells where the gene silencing is not dependant of promoter CpG island methylation. At this cluster in human the silencing relies on the presence of H3K27me3 and polycomb complex (Garrick et al., 2008). GpC islands are not exclusive to promoters. They are also found within gene bodies and intergenic regions. Interestingly, gene body containing CpG islands of active genes can be methylated without blocking transcription elongation (Hellman and Chess, 2007).

1.1.4. Blurry territories

Although patterns can be associated with different epigenetic modifications and chromatin states, the boundaries are not so well defined. Indeed, constitutive heterochromatin and its associated marks such as H3K9me3 are traditionally associated with silencing. However, active transcription has been detected at such regions. For

instance, pericentromeric heterochromatin (PCH) in mouse embryonic stem (ES) cells displayed some level of transcriptional activity (Lehnertz et al., 2003). Similarly, telomeric heterochromatin has been associated with presence of Telomeric Repeat-containing RNA (TERRA) transcription (Oliva-Rico and Herrera, 2017).

1.1.5. Nucleosome occupancy

Beyond histone variants and histone modifications, chromatin fluidity and dynamism are also associated with nucleosome occupancy and turnover. The presence of open and closed chromatin regions regulate the DNA accessibility to other proteins such as transcription factors and RNA polymerase (Lai and Pugh, 2017). Open chromatin regions are more accessible and sensitive to DNase digestion. It can result from nucleosome-depletion or active turnover. By contrast, regions of heterochromatin tend to be closed and therefore less permissive to DNA accessibility by other proteins.

Regions of active chromatin and marks, such as H3K4me3, at unmethylated promoters are often associated with open chromatin regions whereas repressed chromatin with marks, as H3K9me3, are most likely closed.

1.1.6. Regulation of transcription

Encoding approximately 20,000 genes (Pertea and Salzberg, 2010), only 1.2% of the human genome is composed of protein coding sequences. However, the ENCODE project has assigned biochemical function to 80% of the genome (Consortium, 2012).

The regulation of gene expression is a complex interplay between chromatin dynamics, genomic cis-regulatory elements such as promoters, enhancers and insulators and additional trans-acting factors such as transcription factors.

Enhancer elements were first described in *in vitro* assays, as genomic sequences able to upregulate the expression of a gene from distance and independently of its orientation (Banerji et al., 1981, Benoist and Chambon, 1981). The discovery of enhancers highlighted the presence of an additional layer of flexibility and control for the regulation of gene expression. Surprisingly, the upregulation activity of a virus enhancer sequence (SV40) could be effective as well to promote the transcription of a human gene (beta globin) (Banerji et al., 1981). These observations implied that the "enhancer activity" was, at least to some extent, intrinsic to the sequence itself. Indeed, short and specific DNA sequences, or motifs, were characteristics of transcription factor binding sites (TFBS). Transcription factors are DNA binding proteins that have been associated with the regulation of gene expression in a general or in a tissue specific context. Transcription factors can act as activators as well as repressors of transcription (Yuh et al., 1994, Stumpf et al., 2006, Nolis et al., 2009, Peukert et al., 1997, Spitz and Furlong, 2012).

Enhancers usually possess several TFBS, which allow multiple possible combinatorial bindings. Chromatin dynamics is involved in modulating the environment and accessibility to the TFBS. The transcription factors occupancy at enhancers is often associated with DNA hypersensitivity and presence of H3K4me1 and H3K27ac, two markers of active chromatin (Shlyueva et al., 2014). However, even if specific enrichment in H3K4me1 is often defined as characteristic of enhancers, the presence of H3K27ac is a more indicative mark of an active enhancer. By contrast presence of repressive marks such as the polycomb protein-associated mark H3K27me3 has been associated with poised enhancers (Shlyueva et al., 2014).

The dynamic 3-dimentional chromatin configuration, including long range interactions between a promoter and its spatio-temporal regulatory elements, has also been

associated with transcription regulation (Zhang et al., 2013). Strikingly, long-range interactions displayed cell type specific patterns and promoter can interact with several regulatory elements whereas enhancers do not necessarily interact with their closest active gene (Zhang et al., 2013, Hay et al., 2016). Long-range interactions are not limited to promoter-enhancers and the topology of loci with interactions between CTCF bound insulators. In fact, broad structural domains termed topologically associated domains or TAD has been observed. These domains, of roughly 1Mb in size in mammals, display higher numbers of interactions within their boundary than across. Similar phenomena were observed at a smaller scale within TAD and were therefore termed sub-TAD (Phillips-Cremins, 2014). Interestingly, disruption of insulator elements has been associated with mis-regulation of transcription (Hanssen et al., 2017).

1.1.7. Writers, readers and erasers

Ensuring proper histone turnover, use of histone variants and PTM, specialised proteins are essential for chromatin dynamics allowing compaction as well as DNA accessibility for processes such as gene transcription, DNA replication and repair.

Histone PTM are intrinsically linked to groups of proteins known as writers, readers and erasers. Writers are enzymes that modify the histones during the PTM process, as the histone methyltransferase SETDB1, which is involved in the methylation of H3K9 (Schultz et al., 2002). Reader proteins recognise specific PTM. This is the case of the heterochromatin Protein 1 (HP1) that directly interacts with methylated H3K9, which is involved in heterochromatin-mediated gene silencing (Lachner et al., 2001, Ayyanathan et al., 2003). Finally, erasers are enzymes that remove specific PTM, such as the demethylase JHDM2A (jumonji domain-containing histone demethylase-2 A) that removes the methyl group in H3K9me1 and H3K9me2 (Yamane et al., 2006).

1.1.8. Histone chaperones and chromatin remodeling factors

Chromatin dynamics would not be possible without the actions of histone chaperone and chromatin remodeling proteins. These proteins are involved in the establishment of specific chromatin states throughout development and cell differentiation. They are also essential for histone turnover to regulate DNA accessibility during transcription, replication and DNA repair.

Histone chaperones play a critical role in the correct formation of nucleosome structure and prevent nonspecific interactions between the positively charged histones and the negatively charged DNA (Ransom et al., 2010).

Moreover, they are essential factors acting in the incorporation of specific histone variants. Deposition of nucleosomes can occur through different pathways implicating specific sets of proteins. Indeed, whereas H3.1 and H3.2 are associated with replication-dependent deposition involving the CAF1 (Chromatin Assembly Factor) complex, the deposition of the histone variant H3.3 is replication-independent via HIRA or DAXX-ATRX complexes (Lewis et al., 2010).

The actions of histone chaperones are intimately associated with chromatin remodeling proteins. The ATP-dependent chromatin remodeling factors use the energy of ATP hydrolysis to change nucleosome structure and position hence modulating DNA accessibility. These proteins contain a region of homology to the helicase-related Snf2 family ATPase. This domain displays conserved helicase-related motifs associated with helicase-like proteins (Ryan and Owen-Hughes, 2011). However, the presence of this domain does not make them helicases per se. Indeed, these chromatin remodeling factors lack the wedge-like DNA-duplex-destabilizing domain of helicase which is

essential for the ability to separate nucleic acid strands (Saha et al., 2006). However, ATP-dependent chromatin remodelers display translocase activity.

ATP-dependent chromatin remodelers are often split in four main families, SWI /SNF, ISWI, CHD and INO80, based on their unique flanking domains. Nevertheless, due to the presence of a unique ATRX-DNMT3-DNMT3L (ADD) domain, the chromatin remodeler alpha-thalassemia mental retardation X-linked (ATRX) is sometimes consider to belong to a fifth family named after it instead of being classified as a member of the SWI/SNF family (Bartholomew, 2014).

1.2. The chromatin remodeling factor ATRX

1.2.1. Gene and protein structures of the chromatin remodeler ATRX

Located on the long arm of chromosome X (Xq13), human *ATRX* is a large gene spanning approximately 300kb and containing 36 exons. This gene encodes for two major isoforms: a full-length protein of ~280 kDa and a truncated form (ATRXt) of ~180 kDa (Garrick et al., 2004). Additional isoforms have been observed including a 267 kDa one.

The full-length protein is a large protein of 2492 aa containing two main domains, a N-terminal ADD domain and a C-terminal helicase/ATPase domain (Garrick et al., 2004). The truncated form, however, results from an alternative splicing event at intron 11 allowing the use of a proximal intronic poly-A signal (Garrick et al., 2004). Thus, the latter protein isoform does not contain the ATPase domain and is not characteristic of the member of the Snf2 family.

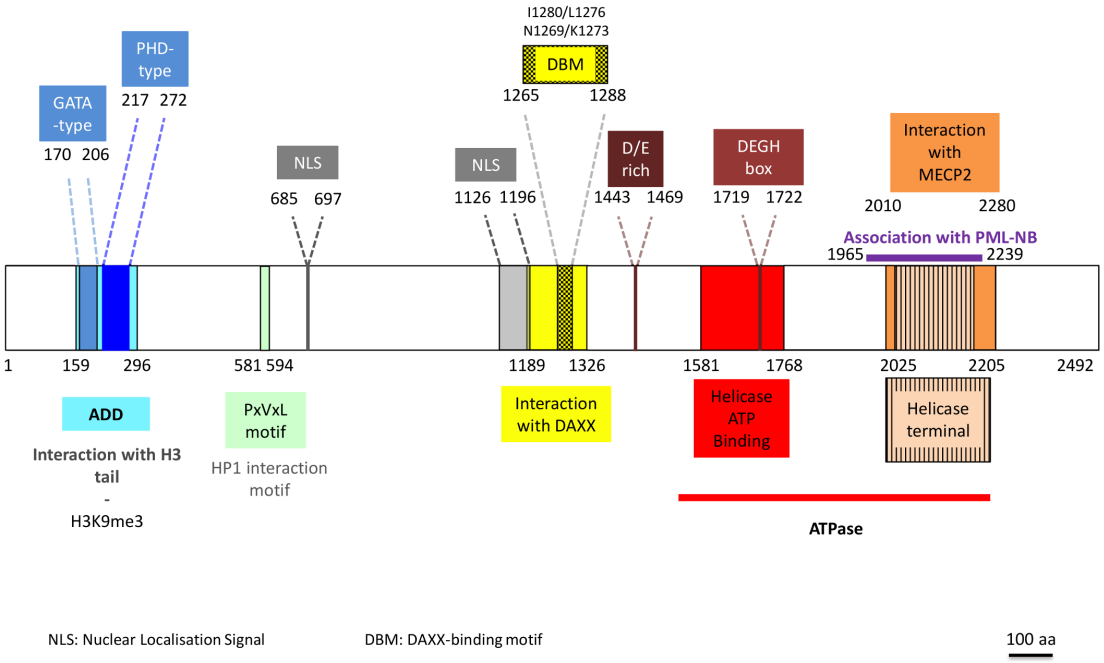


Figure 1.1: Schematic representation of the human full-length ATRX protein.

The ADD domain is cysteine-rich and comprises two subdomains, a GATA-like zinc finger and a plant homeodomain finger (PHD) (Argentaro et al., 2007).

The characteristic features of the ATRX ADD domain were only identified in three other proteins, DNMT3A, DNMT3B, and DNMT3L, all involved in DNA methylation. The presence the ADD domain in ATRX orthologs has been so far restricted to vertebrates and ATRX orthologs (Xnp) in *Drosophila* and *C. elegans* are deprived of it. Interestingly, the latter two species show rare or absent DNA methylation (Argentaro et al., 2007).

The human ADD domain forms a globular structure via hydrophobic interactions. The N-terminal GATA-like subdomain interacts with one Zinc ion through four cysteines. Its structure is very similar to the zinc finger binding protein GATA-1, which is an erythroid transcription factor. The homology between the GATA-like subdomain and the GATA-1 may confer some DNA binding properties to ATRX. More distal to the

N-terminus, the ADD domain contains a second zinc finger binding motif (***Figure 1.1***). This motif is structurally related to the PHD finger subdomain with an additional peculiarity as it is linked with a long C-terminal α-helix. It is the interaction between this helix and the GATA-like module that is responsible for the globular organisation of the ADD domain (Argentaro et al., 2007). The PHD module is one of the most common components in nuclear proteins and is often associated with reader proteins able to sense specific histone PTM (Li and Li, 2012). This is for instance the case of the human bromodomain and PHD domain transcription factor (BPTF). BPTF, a subunit of the ATP-dependent chromatin-remodeling complex, NURF (nucleosome remodeling factor), is able to discriminate between unmodified and monomethylated H3K4 and preferentially interacts with H3K4me2/3 via its PHD module (Li et al., 2006).

Further analysis of the ADD domain revealed a unique interplay between the GATA-like and the PHD subdomains resulting in a combinational readout of both H3K4me0 and H3K9me3 marks (Eustermann et al., 2011, Iwase et al., 2011). By contrast with the BPTF, the ATRX PHD subdomain recognises unmethylated H3K4 whereas the H3K9me3 recognition involved the GATA-like subdomain as well as the linker connecting the module to the PDH finger. Therefore, the ADD structure provides a scaffold coupling the two recognition events through a rigid mutual orientation of the two recognition sites (Eustermann et al., 2011). The combinational readout of the ATRX ADD domain is not shared by the DNMT3 ADD domains which only offer a H3K4me0 recognition (but not the H3K9me3) (Iwase et al., 2011).

Downstream of the ADD domain, ATRX contains a PxVxL motif, which is involved in the interaction with the HP1 protein (***Figure 1.1***). HP1 is a transcriptional repressor which binds directly to methylated H3K9 via its C-terminal 'chromatin organisation modifier' (chromo) domain (Zeng et al., 2010) and with several PxVxL motif containing proteins, such as ATRX, via its N-terminal chromoshadow domain (Lechner et al., 2005). HP1 is mainly known to be a major constituent of heterochromatin and is involved in gene silencing. Consistent with its role as a chromatin remodeler, ATRX possesses two functional nuclear localisation signals (NLS) necessary for its translocation into the nucleus (***Figure 1.1***) (Bérubé et al., 2008).

Upstream of its ATPase domain, ATRX displays a DAXX-binding motif containing the residues N1269/K1273 and L1276/ I1280, that are essential for its interaction with the histone chaperone Death-associated protein 6 (DAXX) (***Figure 1.1***) (Tang et al., 2004, Hoelper et al., 2017). Interestingly, the C-terminal part of ATRX contains the ATPase domain characteristic of the chromatin remodeling factors and protein interaction sequences embedded within it (***Figure 1.1***). It is the case of the Methyl-CpG-binding protein 2 (MeCP2) interacting sequence. MeCP2 protein interacts with methylated CpG sites and is essential for normal brain function (Shah and Bird, 2017). Another sequence spanning the helicase terminal domain of ATRX is involved in the ATRX association with promyelocytic leukemia nuclear bodies (PML-NB) (***Figure 1.1***) (Bérubé et al., 2008). PML-NB are speckles of 0.1 to 1 μm in diameter that have been observed in the nuclear matrix of most cell lines and tissues (Lallemand-Breitenbach, 2010). The nuclear matrix is a non-chromatin framework fulfilling a structural role and being responsible for the maintenance of the nucleus shape and spatial organisation of chromatin (Wasąg and Lenartowski, 2016). PML proteins are the key organisers of

PML-NB, which recruit numerous nuclear proteins and are involved in various processes such as transcription regulation, apoptosis and anti-viral responses (Dellaire and Bazett-Jones, 2004). Pml-null mice develop normally despite their inability to form these structures (Lallemand-Breitenbach, 2010) but they showed impaired immune function and sensitivity to carcinogens (Dellaire and Bazett-Jones, 2004).

1.2.2. Remodeling and translocation activity of ATRX and ATRX-DAXX complex

ATRX forms a chromatin remodeling complex with the histone chaperone DAXX. Several *in vitro* studies have analysed the remodeling and translocation activity of ATRX alone or in complex with DAXX. Similar to the SWI/SNF chromatin remodeler complex (member of the SWI/SNF family), ATRX-DAXX complex showed an ATPase activity that was stimulated by DNA or nucleosomes (Xue et al., 2003). ATRX alone also displayed an ATPase activity stimulated by these two factors (Mitson et al., 2011). Yet, another study found only mononucleosomes stimulated ATRX's ATPase activity in the absence of DAXX (Tang et al., 2004).

The mononucleosome disruption assay is another common experiment used to investigate the properties of chromatin remodeling complexes. This assay is based on a fragment of DNA including a nucleosome positioning sequence assembled into a rotationally phased nucleosome that is subsequently digested by DNAseI and analysed on denaturing polyacrylamide gel electrophoresis. DNAseI digestion produces a characteristic pattern of 10bp ladders. The chromatin remodeling activity of a complex is thereby assessed based on the ability of the complex to alter this digestion pattern in the presence or absence of ATP. As a SWI/SNF complex, ATRX complex could alter the digestion pattern in an ATP-dependent manner (Xue et al., 2003). Nevertheless, the

results suggested that chromatin remodeling activity of the ATRX complex did not alter nucleosome phasing as observed for a SWI/SNF complex. Actually, it was mainly involved in the disruption of the DNA-histone interaction at the entry of the nucleosome (Xue et al., 2003). Subsequent studies showed that DAXX/ATRX complex can assemble H3.3-containing nucleosomes along a DNA template and catalyse the formation of extended nucleosome arrays (Lewis et al., 2010).

Furthermore, ATRX alone, or in complex with DAXX, displays ATP-dependent translocase activity and is able to displace the third strand of a DNA triple-helix (Xue et al., 2003, Mitson et al., 2011). It is nevertheless suspected that the ATRX complex first interacts with flanking double stranded DNA before third strand displacement. Indeed, the blunt triple helix was unaffected by ATRX even in presence of ATP.

Despite its translocase activity, the ATRX complex remains a chromatin remodeler complex and fails to show helicase activity. By contrast with the BLM helicase, ATRX displayed no detectable activity during duplex DNA displacement assay (Xue et al., 2003) and was unable to unwind four-stranded DNA G-quadruplex structures (G4) (Clynes et al., 2014). Nevertheless, the G4 unwinding assay was performed using recombinant ATRX alone in the absence of DAXX and histone proteins. Therefore, it remains unknown if ATRX could participate to G4 unwinding in an *in vivo* context.

Taken together, these results confirmed the status of ATRX as an ATP-dependent chromatin remodeler and highlighted both similarities and differences, between the remodeling activities of a characteristic member of the SWI/SNF family, the SWI/SNF complex and ATRX complex. There are many reasons to debate whether ATRX should be classified among the SWI/SNF family or in a separate family.

1.2.3. Localisation and recruitment of ATRX beyond chromatin

The localisation of ATRX within the nucleus is not limited to chromatin. Thus, speckled nuclear staining observed by immune-fluorescence highlights the co-localisation of ATRX, DAXX and PML-NB (Xue et al., 2003). The C-terminal sequence of ATRX is implicated in the recruitment of ATRX to PML-NB. Mutations in this sequence impaired ATRX-PML-NB interaction and mutant ATRX proteins displayed a more diffuse pattern throughout the nucleus (Bérubé et al., 2008). Interestingly, ATRX recruitment to PML-NB has also been shown to be DAXX dependent (Tang et al., 2004). Indeed, ATRX co-localisation with PML-NB was impaired in presence of a DAXX mutant lacking either its ATRX or PML interacting motif (Tang et al., 2004). This result was further supported by work carried out in human cells showing that DAXX knock out (KO) abolished ATRX accumulation at PML-NB (Morozov et al., 2017).

1.2.4. Localisation and recruitment of ATRX within chromatin

As a chromatin remodeler, ATRX is expected to be enriched at chromatin regions. Indeed, ATRX has been observed to predominantly be localised to repetitive heterochromatin regions including pericentromeric, telomeric and ribosomal DNA and to some interspersed repeats of endogenous retroviral elements (ERV) (McDowell et al., 1999, Wong et al., 2010, Voon et al., 2015, Sadic et al., 2015). Surprisingly, its method of recruitment differs between its different targets.

1.2.5. ATRX enrichment at pericentromeric heterochromatin (PCH)

Whereas most of the metazoan telomeric heterochromatin is constituted of repeats of a short conserved DNA motif (5′-TTAGGG-3′), centromeric and pericentromeric heterochromatins do not display strong sequence conservation and repeat sequences. Indeed, the organisation can be highly variable between organisms (and sometimes even between chromosomes of the same species) (Saksouk et al., 2015). Even mouse and human display different centromeric and pericentromeric compositions (Saksouk et al., 2015). Epigenetic signatures of PCH are characterised by methylation of the CpG sites, and high level of H3K9me3 and HP1.

Study in mouse cells have greatly contributed to our understanding of ATRX recruitment to pericentromeric regions. The mouse centromere is composed of minor satellites flanked by A/T rich pericentromeric regions, which are made of 'major' satellites. Studies in mouse cells showed that the ATRX location to PCH mainly depends on H3K9me3. Indeed, in mouse embryonic fibroblast (MEF) cells, upon double KO of the H3K9me3 methyltransferase Suv39h1/h2 and depletion of H3K9me3, ATRX localisation to PCH was impaired (Iwase et al., 2011). By contrast, it had no effect on ATRX localisation to DAXX-positive PML bodies. Furthermore, re-expression of Suv39h1 associated with the recovery of H3K9me3 is sufficient to recover ATRX signal at PCH. As the localisation of MeCP2 at PCH is retained in Suv39h1/h2 double KO, these results suggested that MeCP2 was not the main influence on the recruitment of ATRX to PCH (Iwase et al., 2011).

Nevertheless, experiments in MeCP2-null mice showed that ATRX recruitment to heterochromatin foci was impaired by MeCP2 loss only in tissues normally expressing a high level MeCP2 such as brain (Nan et al., 2007). In that case, HP1 and H3K9me3 were not sufficient to ensure ATRX localisation in such MeCP2 null tissues. By

contrast, ATRX localisation appeared unaffected in absence of MeCP2 in tissues that normally are not expressing a high level of MeCP2 such as heart (Nan et al., 2007).

Loss of H3K9me3 in Suv39h1/h2 double KO was also associated with depletion of HP1 at PCH, leaving the possibility that the interaction with HP1 may be essential for ATRX recruitment to this region (Iwase et al., 2011). Further experiments disrupting the HP1 interacting motif of ATRX, had less dramatic effects on ATRX localisation to PCH than Suv39h1/h2 double KO. By contrast, knock down of ATRX did not affect HP1 enrichment at PCH (Iwase et al., 2011). Complementary experiments have supported the predominant role of H3K9me3 and HP1 in ATRX localisation to PCH. Thus a combination of two ATRX mutations, one destabilising the ADD domain and one abolishing the HP1 interaction, abolished the ATRX enrichment at PCH. Nevertheless, single mutation destabilising the ADD domain had more effect than the single mutation impairing the interaction with HP1 on the ATRX localisation at PCH. Even though ATRX association with HP1 is involved in its recruitment to PCH, interactions of ATRX and HP1 with H3K9me3 are mutually exclusive suggesting that ATRX and HP1 binds to adjacent H3 tails (Eustermann et al., 2011).

Strengthening the combinatorial readout of the ATRX ADD domain and the role of H3K9me3 for ATRX localisation to PCH, mutations affecting either H3K4me0 or H3K9me3 recognition without affecting the stability of the ADD domain, impaired the location of ATRX to PCH (Eustermann et al., 2011). Completing the picture, work in DAXX null cells showed that ATRX still localises to PCH suggesting that DAXX is not essential for ATRX recruitment to PCH (Ishov et al., 2004).

1.2.6. ATRX association with telomeres, rDNA and other GC-rich sequences

Located at each end of linear chromosomes, telomeres are nucleoprotein structures, which play essential roles in genome stability and chromosome replication. In vertebrates, telomeres contain highly repetitive sequences composed of several kb arrays of G-rich short tandem repeat sequences (TTAGGG). ATRX localisation at telomeres varies between cells types and across the cell cycle (Wong et al., 2010). Indeed, whereas ATRX accumulates at PCH in both mouse ES and somatic cells, and in human cells, ATRX enrichment at telomeres was mainly observed in mouse ES cells (Wong et al., 2010). These results raised the possibility that ATRX localisation at telomere may be affected by the pluripotent state of the cell. This hypothesis was further supported by the decrease of ATRX signals at telomeres across cell differentiation (Wong et al., 2010). Moreover, timeline analysis of ATRX localisation using synchronised mouse ES cells revealed that ATRX is predominantly recruited at telomeres during the S-phase and dissociated at mitosis. By contrast, ATRX localisation at PCH remained stable (Wong et al., 2010). Further contrasting with PCH, ATRX localisation to GC rich telomeric and rDNA repeats was not reduced in MEF Suv39h1/h2 double KO (Nguyen et al., 2017). However, ATRX localisation to telomeres is affected by PML-NB. Indeed, in PML-depleted mouse ES cells, ATRX localisation at telomeres during S-phase was reduced (Chang et al., 2013).

The association between ATRX localisation at telomeres and PML-NB was further supplemented by the analysis of the DEK oncoprotein. DEK, like DAXX, is an H3.3 histone chaperone. In mouse ES cells, it co-localises with ATRX and PML at telomeres. Similar to ATRX, DEK localisation to telomeres appeared linked to cellular pluripotency and telomeric DEK enrichment observed in ES cells is decreased upon differentiation (Ivanauskiene et al., 2014). The depletion of DEK in mouse ES cells is

associated with a delocalisation of ATRX and PML from telomeres and an increased enrichment of these factors at PCH. Thus, DEK is also involved in ATRX localisation to telomeres (Ivanauskiene, 2014). Interestingly, DEK also displays DNA binding properties for non-B form DNA structures such as cruciform DNA instead of sequence specificity (Waldmann et al., 2003).

Represented by the Watson and Crick model, the canonical B-form DNA structure is a right-handed double helix. This conformation is common in cells. However, a wide range of non-canonical structures can be adopted by DNA *in vitro* and *in vivo* based on its sequence and environment (Zhao et al., 2010, Brázda et al., 2011, Rhodes and Lipps, 2015, Bochman et al., 2012). A growing number of proteins can bind to these structures and an increasing number of studies have highlighted the importance of these structures including in the regulation of gene expression (Bochman et al., 2012, Waldmann et al., 2003, Brázda et al., 2011).

Interestingly, genome-wide analyses of ATRX binding sites in mouse and human have showed that in addition to telomeres and rDNA, ATRX was also enriched at a subset of GC-rich tandem repeats and CpG islands (Law et al., 2010, Voon et al., 2015). Comparison of ATRX enrichment in wild-type (WT) and Suv39h1/h2 double KO MEF cells suggested that the requirement of H3K9me3 for the localisation of ATRX was inversely proportional to the GC content of its targets at GC-rich repeat sequences (Nguyen et al., 2017). GC-rich repeats are prone to form non B-form structures including G-quadruplexes (Rhodes and Lipps, 2015). Typically, G4 are stable four-stranded DNA structures that resulted from the intramolecular or intermolecular folding of single-stranded guanine rich sequences. G4 are held by planar quartets of guanine

(G-quartets) bound via Hoogsteen hydrogen interactions. Monovalent cations such as K+ and Na+ can further stabilise the structure (Rhodes and Lipps, 2015). G4 can adopt a variety of topologies. Putative G4 forming sequences are predicted based on the presence of a motif containing at least four runs of minimum three guanine residues with eventual presence of other bases in between the G-runs: $G_{3+}N_{1-n}G_{3+}N_{1-n}G_{3+}N_{1-n}G_{3+}$ (Hänsel-Hertsch et al., 2017, Sahakyan et al., 2017).

In vitro analysis revealed that, once formed, most G4 were more thermodynamically stable than duplex DNA and they displayed much slower unfolding kinetics than those of DNA or RNA hairpins (Lane et al., 2008). These observations suggested that G4 may act as a block which may require specific proteins to be unfolded. Such proteins indeed exist and include BLM, WRN, FANCJ and Pif1 helicases (Paeschke et al., 2013, Rhodes and Lipps, 2015).

Interestingly, a subset of ATRX targets, including telomeric regions, have been predicted to form G4 (Law et al., 2010). Moreover *in vitro* experiments have showed that ATRX can directly bind to G4 even though the remodeling activity of ATRX alone was insufficient to unwind it (Clynes et al., 2014). An appealing hypothesis to consider is the possibility that ATRX may be targeted to a subset of its targets via its interaction with G4 structures *in vivo*.

In addition, transcribed GC-rich regions are also prone to form a three-stranded RNA:DNA hybrid structures known as R-loops (Sanz et al., 2016). In this structure, one DNA strand interacts with a complementary RNA strand while the complementary DNA strand is displaced and single stranded. Several ATRX targets have the potential to form R-loops including the telomeric repeats. By analysing ATRX recruitment at an ectopic locus that contains various numbers of telomeric repeats (TTAGGG), recent work from our group showed that ATRX recruitment at this region was not only

dependent on the repeat number and orientation but also transcription (Nguyen et al., 2017). Furthermore, the abundance of R-loops was correlated with ATRX enrichment at this region (Nguyen et al., 2017). Transcription is also a mechanism potentially favourable to the formation of G4 as it induces the transient formation of single stranded DNA that may subsequently fold into G4 structures (Hänsel-Hertsch et al., 2016). It is therefore conceivable that transcription could be intimately associated to ATRX recruitment at some loci.

Another type of GC-rich regions that is enriched in ATRX is the CpG clusters or islands. Among the CpG island regions bound by ATRX, a subgroup is associated with differentially methylated regions (DMR) observed at imprinting loci.

Imprinting genes have the peculiarity of being monoallelically expressed in a parental of origin manner. This phenomenon is associated with the methylation of only one allele of the DMR, depending of their parental origin. Several studies in mouse highlighted the enrichment of ATRX at such regions (Kernohan et al., 2010, Kernohan et al., 2014, Voon et al., 2015). Interestingly, ATRX localisation at DMR appears to vary dependent on the cell types and developmental stage. Thus, in mouse ES cells, ATRX localises preferentially to methylated alleles of the imprinted DMR that are enriched in H3K9me3 modifications (Voon et al., 2015). By contrast, in mouse forebrain, ATRX predominantly binds to the unmethylated allele of H19 imprinting control region (ICR), which is involved in the imprinting regulation of H19 and Igf2 (Kernohan et al., 2014). In parallel, this study emphasised the absence of ATRX at the same site, pre-birth (E13.5). At this region, ATRX co-localises with MeCP2 in mouse forebrain. This protein is known to recruit ATRX to heterochromatin regions in brain (Nan et al., 2007). Analysis of ATRX enrichment in MeCP2-null forebrain showed that

ATRX occupancy at H19 ICR was also dependent on MeCP2 (Kernohan et al., 2014). Interestingly, ATRX is enriched at the inactive X (iX) chromosome and co-localised with macroH2A1.2 at that regions (Baumann and De La Fuente, 2009).

1.2.7. ATRX occupancy at the 3' exon of Zinc Fingers

In human cell lines, ATRX localisation has recently been reported at another group of gene regions corresponding to the 3' exon of a subset of Zinc Finger (ZNF) genes (Valle-García et al., 2016). With more than 700 ZNF genes identified in the human genome, it is the largest family of putative transcriptional factors. In particular, ATRX enrichment was found at ZNF genes encoding C2H2 zinc finger (C2H2-ZNF) amino acid motifs (Valle-García et al., 2016). The number of C2H2 motifs is variable across the C2H2-ZNF proteins, with an average of nine C2H2 motifs per protein, which are mostly encoded in the 3' exon of the gene. ATRX occupancy at the 3' exon of the C2H2-ZNF encoding genes was furthermore correlated with the presence of an 'atypical chromatin signature': an enrichment of both H3K9me3 and H3K36me3 (Valle-García et al., 2016). Even though C2H2 motifs are repeated, the corresponding coding sequences are not tandem repeats per se. Furthermore, the 3' exon of ZNF genes enriched in ATRX was found to negatively correlated with G-content and putative G4 forming sequences. It is therefore unlikely that ATRX could be recruited to these regions through its interaction with G4 (Valle-García et al., 2016). In addition to ATRX, the 3' exon of ZNF genes is also enriched in ZNF274 (Frietze et al., 2010). ZNF274 is a putative transcription factor and contains a Kruppel-associated box (KRAB) domain, which is often associated with transcriptional repressors (Lupo et al., 2013). Depletion of ZNF274 correlated with the reduction of ATRX enrichment at some of the 3' exon of ZNF genes suggesting that ZNF274 may be involved in ATRX occupancy at this

subset of sites (Valle-García et al., 2016). Furthermore, ZNF274 is suspected to recruit the H3K9 methyltransferase SETDB1 at the 3' end of ZNF genes via its association with the KRAB-associated protein TRIM38 (Frietze et al., 2010). ZNF274 knock down correlated with a decreased level of TRIM38 and SETDB1 at these sites (Frietze et al., 2010). Strikingly, the reduction of ATRX occupancy in ZNF274 deficient cells was also associated with a reduced level of H3K9me3 at that sites (Valle-García et al., 2016). It is therefore plausible to suspect that H3K9me3 depletion may be due to the reduced level of SETDB1 recruitment to these sites due to ZNF274 deficiency. This observation raised the possibility that it is H3K9me3, as observed at PCH for instance, rather than ZNF274, which may be directly responsible for ATRX localisation at the 3' exon of some ZNF genes.

1.3. ATRX functions in health and disease

1.3.1. ATRX-associated disorders related to alpha-thalassaemia

In 1995, congenital mutations in the *ATRX* gene were identified as the cause of ATR-X syndrome (Gibbons et al., 1995). ATR-X syndrome is an X-linked disorder, which almost exclusively affects males. Interestingly, female carriers display a preferential X inactivation of the chromosome containing the mutated *ATRX* allele and are usually physically and intellectually unaffected (Gibbons, 2006). Surprisingly, the first female that has been diagnosed with ATR-X syndrome was born after *in vitro* fertilisation. As female carriers, she showed a skewed pattern in X inactivation. However, she has a preferential inactivation of the WT ATRX allele. Therefore, the mutated ATRX allele has been preferentially expressed (Badens et al., 2006).

ATR-X syndrome is a very rare disease characterised by developmental delay, facial dysmorphism, genital abnormalities and a deficient production of alpha globin protein,

which gives rise to the alpha-thalassaemia phenotype (Gibbons, 2006). Additional features include skeletal abnormalities, reduced muscle function (muscle hypotonia), microcephaly and short stature. These symptoms vary in each case and can range from moderate to severe.

In 2008, over 200 cases were identified in less than 200 independent families. As the symptoms may be confused with other mental retardation syndromes, it is suspected that the ATR-X syndrome is under-diagnosed and, although unknown, its prevalence has been estimated between 1/30,000-40,000 neonate boys (Wada et al., 2013). The distribution of pathogenic mutations identified in ATR-X patients is not random. Thus, most of these mutations fall into two domains, ADD domain which contains 50% of them and the ATPase domain which account for 30% of the identified mutations; most mutations are missense (Gibbons et al., 2008).

ATR-X patients with alpha-thalassaemia display intact alpha globin loci. In contrast, patients with standard alpha-thalassaemia harbour mutations within the alpha globin cluster including deletions of the alpha globin genes (Harteveld and Higgs, 2010). This peculiarity and the complexity of the phenotypes associated with ATR-X syndrome suggest that ATRX plays a role in the regulation of gene expression including the alpha globin genes.

The role of ATRX in the regulation of the alpha globin was further supported through the analysis of samples from patients with a myelodysplasia syndrome (clonal disorder affecting myeloid progenitor) associated with alpha-thalassaemia (ATMDS) (Gibbons et al., 2003). These patients have acquired a *de novo* form of alpha-thalassaemia and could exhibit such severe downregulation of the alpha globin that would have been lethal during development. However, as for the ATR-X syndrome, no abnormality was detected within the alpha globin cluster (Gibbons et al., 2003). Microarray analysis of

ATMDS patients together with unaffected controls revealed a significant downregulation of *ATRX* gene expression. By contrast, patients with myelodysplasia syndrome without alpha-thalassaemia did not showed significant changes in *ATRX* expression. Further analyses identified acquired ATRX mutations that were present only in affected cells but absent in cells unrelated to the myeloid lineage within the same patient. Taken together these observation strongly suggest an essential role of ATRX for the normal expression of alpha globin *in vivo* (Gibbons et al., 2003).

Originally, the diagnosis of ATR-X syndrome required the presence of alpha-thalassaemia. It is now clear that not all the patients present this phenotype which is detected in 85% of the cases (Gibbons et al., 2008). However, the detection of alpha-thalassaemia is often indirect and do not assess the alpha globin expression itself. One common method is to use the detection of HbH inclusions in cells obtained from a peripheral blood sample. These inclusions result from an imbalance between the production of alpha globin and beta globin chains, both part of tetrameric haemoglobin molecule composed of two chains of alpha globin and two chains of beta globin. The insufficient production of alpha globin chains is associated with an excess of beta globin chains that cannot associate with alpha globin chains to form normal haemoglobin, which therefore aggregate and form tetramers of beta globin seen as HbH inclusions. However, the detection of HbHs inclusion requires at least a 50% of reduction in alpha globin synthesis. Consequently, the absence of HbH inclusions does not necessarily mean that the alpha globin expression is unaffected (Gibbons, 2006).

Strikingly, the severity of the alpha-thalassaemia in ATR-X patients did not appear to be dependent of the specific mutation in ATRX. Indeed, even within patients with the same ATRX point mutation, the severity of the alpha-thalassaemia, measured as a

percentage of cells containing HbH inclusion, ranges from 0 to 14%, suggesting that individual genetic backgrounds have an effect (Gibbons, 2006). Although still unclear, the first insight into the underlying mechanism linking ATRX mutations, genetic background and the alpha globin regulation came from the first ATRX ChIP-seq and ChIP-on-CHIP experiments performed using human cells (Law et al., 2010). This study focused on the subtelomeric region of the short arm of chromosome 16, which includes the alpha globin cluster. ATRX binding sites associated with G-rich tandem repeats were identified in proximity or within genes downregulated in ATR-X patients (Law et al., 2010). Thus, the alpha globin cluster displayed an ATRX enrichment at the (GCGGG)n variable number tandem repeat (VNTR) ψζVNTR approximately 1kb upstream of the HBM promoter. Interestingly, ATRX occupancy at this site appeared to correlate with the alpha globin expression, and the ATRX peak was only detected in erythroid cells. By contrast, an ATRX peak observed at a VNTR located in the intron of the widely expressed Nucleoside Diphosphate Kinase *NME4* gene was present in both erythroid and non-erythroid cells. As with alpha globin, NME4 is significantly downregulated in ATR-X patients. Both the ψζVNTR and the NME4 VNTR fall in G-rich ATRX binding sites predicted to form G4 and are very polymorphic between individuals. Surprisingly, the comparison between the level of HbH inclusions and the average size of the ψζVNTR repeats in a group of more than forty ATR-X patients revealed a correlation between the size of the ψζVNTR and the severity of the alpha-thalassaemia. Thus, the larger the ψζVNTR repeat was, the more severe the alpha-thalassaemia appeared to be (Law et al., 2010). Similarly, the larger the VNTR at NME4 was, the more downregulated the allele *in cis* was.

Which ATRX function could explain its role on the regulation of gene expression?

1.3.2. Deposition of H3.3

ATRX has been intimately linked to the histone chaperone DAXX and the deposition of the H3.3 variant in a replication independent manner. Interestingly, the presence of an ATRX-binding-deficient DAXX mutant or loss of DAXX have been shown to be associated with a reduced level of ATRX protein without impacting the *ATRX* gene expression (Lewis et al., 2010, Hoelper et al., 2017).

DAXX is an H3.3-specific histone chaperone. Only five amino acids differ between H3.1 and H3.3, one located in the unstructured N-terminal tail and the other four in the histone fold domain. The affinity of DAXX for H3.3 over H3.1 is dependent of the specific amino acid present in the histone fold domain and single mutation of one of the amino acids is not sufficient to abolish the H3.3/DAXX interaction (Lewis et al., 2010). Nevertheless, conversion of H3.3 specific Glycine (G90) residue, located in the histone fold domain, into a methionine residue as found in H3.1 and H3.2 had the largest effect and results in 50 % reduction of H3.3-DAXX interaction and H3.3 association with the H3.1/2-specific histone chaperone CAF1 (Elsässer et al., 2012). Even though DAXX is identified as a H3.3-specific histone chaperone as it can discriminate H3-variants, it is actually involved in the deposition of H3.3-H4 dimers and tetramers (Elsässer et al., 2012, Drané et al., 2010).

The ATRX-DAXX complex, as HIRA complex, is involved in the replication-independent deposition of H3.3. Interestingly, HIRA and ATRX-DAXX complexes showed different contribution to H3.3 deposition depending on the chromosomal location. The deletion of either HIRA or ATRX results in embryonic death in mice at E11 and E9.5, respectively (Roberts et al., 2002, Goldberg et al., 2010, Garrick et al., 2006). The analysis of the deposition of tagged-H3.3 in mouse ES cells by ChIP-seq

revealed that HIRA KO is correlated with H3.3 depletion at active and repressed genes. By contrast, ATRX deletion was associated with H3.3 depletion at telomeres. Regarding the TFBS enriched in H3.3, most of them were unaffected in either HIRA null or ATRX null (Goldberg et al., 2010).

The role of ATRX in the deposition of H3.3 has been a fascinating area of research for the past decade and further work has supported ATRX-dependent deposition of H3.3 at diverse spatio-temporal localisations. Given the repetitive nature of many ATRX binding sites, it is not surprising that ATRX plays a role in the replication independent histone replacement. Indeed, repetitive sequences have been associated with rapid histone turnover as they are often predicted to be an unfavourable substrate for nucleosomes (Schneiderman et al., 2009).

The role of ATRX in H3.3 deposition at telomeres was further investigated using both ATRX and DAXX KO mouse ES cells. These experiments emphasised that even though ATRX was not required for H3.3 incorporation by DAXX *in vitro*, the impairment of H3.3 incorporation *in vivo* at telomeres was similar in ATRX KO and DAXX KO (Lewis et al., 2010). Notably, in ATRX null cells, DAXX was not identified at telomeres by ChIP. Taken together, these results suggested that ATRX is probably required for DAXX recruitment at telomeres and subsequent H3.3 deposition (Lewis et al., 2010).

Additionally, the DEK depletion is associated with the delocalisation of ATRX from telomeres and results in reduced deposition of H3.3 at those sites. At the same time, a redistribution of H3.3 outside of the telomeres, probably in a HIRA-dependent manner, was observed. In addition, an increase in the DNA damage mark γH2A.X, enhanced

transcription of the telomere-encoded TERRA non-coding RNA and fragile telomere phenotypes were also observed (Ivanauskiene et al., 2014). The upregulation of TERRA was also noticed in ATRX null mouse ES cells (Goldberg et al., 2010).

Furthermore both H3.3 and ATRX depletion by RNAi affect telomere integrity associated with an increased number of telomere dysfunction induced foci (TIF) observed by microscopy (Wong et al., 2010, Wong et al., 2009).

Interestingly, the increase of TIF in ATRX depleted cells was particularly noted if in the WT cells, ATRX was normally abundant at telomeres. In ES cells, where ATRX is particularly enriched at telomeres, the increase in TIF was significant, which was different in somatic mouse cells where ATRX did not accumulate at telomeres (Wong et al., 2010).

ATRX dependent deposition of H3.3 is not limited to telomeric heterochromatin as it is also the case at PCH. Indeed, MEF cells depleted in ATRX or DAXX showed a reduced deposition of H3.3 at PCH associated with a downregulation of pericentric repeats transcription. Similar effects was observed upon H3.3 depletion (Drané et al., 2010).

A recent study has also suggested an ATRX-mediated H3.3 deposition at rDNA tandem repeats (Udugama et al., 2018). ATRX KO mouse ES cells showed a decrease H3.3 enrichment at rDNA that is associated with a loss in rDNA copy number compare to control. Reduction of H3K9me3 and H3K20me3, two marks that co-localise with heterochromatin, were also observed at a subset of loci within the rDNA repeats in ATRX KO cells. Additionally, these cells had an increased in γH2A.X enrichment at the rDNA repeat loci. Taken together, these results suggested that ATRX is required for the maintenance of the heterochromatin integrity at rDNA (Udugama et al., 2018).

Interestingly, reduced level of DNA methylation has been observed at rDNA repeats in ATR-X patients (Gibbons et al., 2000).

The interplay between ATRX and H3.3 deposition was further associated with imprinted loci. In mouse EC cells, in which ATRX binds preferentially to the methylated allele of the DMR, H3.3 enrichment also appeared to be biased toward the methylated allele of DMR (Voon et al., 2015). These DMR were also enriched in H3K9me3 modifications. In ATRX KO cells, H3.3 enrichment at these regions was lost and so was the H3K9me3 marks. In addition, DNA methylation was also impaired and partial de-repression of the normally silenced allele was observed (Voon et al., 2015). Thus, in the context of mouse ES cells, ATRX appeared to be required for the deposition of H3.3, the maintenance of H3K9me3 and gene silencing at the methylated allele of DMR (Voon et al., 2015).

However, ATRX role and localisation at imprinted loci may differ depending on the context. Thus, in mouse forebrain, in which ATRX predominantly binds to the unmethylated H19 imprinting control region (maternal), the role of ATRX is possibly not restricted to H3.3 deposition. ATRX conditional KO does not affect DNA methylation at *H19* ICR and *Igf2* DMR in mouse forebrain. The absence of ATRX correlated with increase in histone acetylation and decrease level of H3K20me3 and H3K9me3. In addition, an increase in H19 (maternally expressed) and Igf2 (paternally expressed) expression was observed after birth but not during development (Kernohan et al., 2010). The increase in H19 transcripts was due to the upregulation of the maternal allele but not due to a reactivation of the paternal allele. However, ATRX KO only occurs at E8.5. At that stage, imprinting establishment had already occurred at H19 and

Igf2 loci. Thus, the upregulation of H19 was more likely due to a defect in the maintenance of the imprinting than of its establishment (Kernohan et al., 2010).

Post-natal silencing of H19 has been observed after birth with an increase in histone repressive marks such as H3K9m3 and H3K20me3. Additionally, *H19* ICR is occupied by ATRX after birth whereas it is not the case at E13.5 (Kernohan et al., 2014). This study suggested that ATRX is involved in the post-natal silencing of the active maternal H19 gene (Kernohan et al., 2010, Kernohan et al., 2014). In addition to ATRX, CTCF and MeCP2 were present at *H19* ICR in mouse forebrain. The absence of ATRX resulted with a reduced occupancy of CTCF at that site after birth. Nevertheless, even in absence of ATRX, the CTCF occupancy remained undisturbed at E13.5. Thus, impairment in CTCF enrichment was correlated with ATRX recruitment suggesting that the CTCF localisation in embryonic brain was independent of ATRX whereas, it becomes ATRX dependent after birth (Kernohan et al., 2014).

H19 ICR does not contain tandem repeats and is not predicted to form G4. In absence of ATRX, an increase of nucleosome protection in the H19 ICR was observed suggesting that ATRX remodeling activity may be important for nucleosome positioning allowing CTCF binding (Kernohan et al., 2014). By using DNA-FISH to analyse the specific proximity of the imprinted domains in the nucleus of brain cells in presence or absence of ATRX, showed that the transient co-localisation of the tested imprinted domains was not significantly affected by ATRX loss (Kernohan 2014). However, the *in cis* long-range interactions across the H19/Igf2 domain were reduced. This abnormal chromosomal looping was suggested to be due to the decrease in CTCF occupancy (Kernohan et al., 2014).

1.3.3. ATRX-DAXX-H3.3 role at retrotransposons

Recent studies have suggested a role of ATRX, DAXX and H3.3 for the silencing and the maintenance of the heterochromatin integrity at a subset of retrotransposons including intracisternal A-type particles (IAP) (Sadic et al., 2015, Elsässer et al., 2015). However, a recent study put into question some of these findings (Hoelper et al., 2017). Indeed, using mouse ES cells depleted for ATRX or DAXX, they found that only DAXX, and not ATRX, leads to dysregulation of these retrotransposons. These results were further supported by the expression of ATRX-binding-deficient DAXX that was sufficient to repress the IAP elements similarly to WT DAXX. By contrast with non-ERV repeats, H3.3 deposition at these elements was also dependent of DAXX and not ATRX. Additionally, it appeared that DAXX and TRIM28 co-operated to repress ERV and DAXX or TRIM28 depletion had similar effects on ERV expression. By contrast, genes differentially expressed upon loss of either DAXX or ATRX were unaffected by depletion of SETDB1 or TRIM28 (Hoelper et al., 2017).

Surprisingly, DAXX formed a complex with SETDB1-TRIM28 independently from its complex with ATRX. Moreover, tagged SETDB1 did not co-immunoprecipitated with ATRX. DAXX associates with SETDB1-TRIM28 via a different interaction than ATRX and ATRX-binding-deficient DAXX could still co-immunoprecipitated with SETDB1-TRIM28 as well as with H3.3 (Hoelper et al., 2017).

DAXX co-folds with H3.3/H4. Several studies highlighted that depletion of H3.3 was associated with a reduced overall nuclear level of DAXX (DeNizio et al., 2014, Elsässer et al., 2015, Hoelper et al., 2017). Based on these observations, Hoelper et al. used different H3.3 mutants, either unable to interact with DAXX or unable to be stably

deposited into nucleosome, to further investigate the role of both DAXX and H3.3 in the regulation of ERV expression. ERV expression was stabilised by mutant H3.3 that is unable to be stably deposited into nucleosomes. Whereas ERV expression was not stabilised by DAXX-binding-deficient H3.3 suggesting that DAXX stabilisation/level was the main factor involved in ERV repression, but not H3.3 incorporation (Hoelper et al., 2017). Further experiments suggested that ERV repression might be more likely to be associated with the recruitment of histone deacetylase, such as HDAC1, than with H3.3 deposition. Strengthening these findings, HDAC1 interacts with DAXX and co-immunoprecipitated TRIM28 in a DAXX dependent manner. These results suggest that DAXX, stabilised via its interaction with H3.3, may act as a binding factor at ERV between HDACs and SETDB1-TRIM28 therefore coupling lysine deacetylation with methylation of H3K9me3 (Hoelper et al., 2017).

1.3.4. Maintenance of H3K9me3 level at the 3' exon of ZNF genes

ATRX has been proposed to maintain H3K9me3 level at the 3' exon of the ZNF genes that are enriched in ATRX, H3K9me3 and H3K36me3. Indeed, in ATRX depleted cells, these sites showed a decrease in H3K9me3 level. However, H3K36me3 level remained unaffected. By contrast with ERV, ATRX function at ZNF 3' exons is presumably DAXX-independent as no significant effect on H3K9me3 enrichment at ZNF genes was noticed in DAXX deficient cells. Therefore, H3.3 deposition may not be involved in H3K9me3 maintenance at these sites. The authors suggested an association between ATRX and the complex ZNF274/TRIM38/SETDB1 to fulfil this role. However, ATRX KO has no effect of ZNF274 and TRIM38 localisation and an uncertain effect on SETDB1 localisation (Valle-García et al., 2016).

1.3.5. ATRX-mediated chromatin association of macroH2A1

ATRX functions are not limited to its association with DAXX and H3.3 deposition. In addition to H3.3, ATRX has been associated with another histone variant, macroH2A1 (Ratnakumar et al., 2012). Along with macroH2A2, macroH2A1 is one of the two isoforms of the H2A histone variant macroH2A (Costanzi and Pehrson, 2001). In contrast with H3.3, macroH2A is associated with transcriptional repression. Interestingly the macroH2A-ATRX association appeared to be DAXX independent and no H3 was detected in the macroH2A-ATRX complex suggesting that ATRX-DAXX-H3.3 complex and ATRX-macroH2A are two distinct complexes with different regulation (Ratnakumar et al., 2012). In human cell lines, depletion of ATRX using shRNA was associated with a global increase of macroH2A in chromatin without affecting its total cellular level. Therefore, ATRX may act as a negative regulator of the incorporation of macroH2A1 into chromatin (Ratnakumar et al., 2012). ATRX knock down is correlated with a macroH2A1 increase at telomeres and at the alpha globin cluster further associated with a reduced expression of the alpha globin genes. So, these observations supported the role of macroH2A1 as a repressive variant and its ATRX-dependent association with chromatin (Ratnakumar et al., 2012). However, other studies have shown that depletion in ATRX is correlated with an increase in TERRA expression (Ivanauskiene et al., 2014, Goldberg et al., 2010). These results question the repressive role of macroH2A at those sites. Consequently, further studies are required to understand the mechanisms associating ATRX and macroH2A1 with chromatin as well as their functions.

1.3.6. ATRX and regulation of ancestral pseudoautosomal region gene expression

Experiments in mouse forebrain identified ancestral pseudoautosomal region (aPAR) genes among the most downregulated genes in absence of ATRX. In contrast to human, in which equivalent genes are found in sexual chromosomes, aPAR genes are located on autosomal chromosomes in mouse. Some of the mouse aPAR genes are located at the subtelomeric regions or at repetitive sites (Levy et al., 2008). The effects of ATRX depletion on the aPAR gene expression were observed during the replication and in terminally differentiated tissues. Therefore, it was thought that the impact of ATRX on the regulation of these genes is probably independent of DNA replication (Levy et al., 2008). ATRX was enriched within the gene body of several aPAR genes at sites predicted to form G4 (Levy et al., 2014). Loss of ATRX was correlated with reduced H3.3 occupancy at these sites and impeded the progression of the RNA polII at these G-rich regions. However, no change in TERRA expression, DNA methylation or histone modifications was noticed. It was therefore proposed that ATRX facilitates the transcription elongation through G-rich templates via deposition of H3.3 to support the passage of the transcription machinery (Levy et al., 2014).

1.3.7. ATRX effects on gene expression, DNA replication and genomic stability

It is now clear that ATRX depletion is associated with the dysregulation of a subset of coding and non-coding genes. Beyond the regulation of gene expression, ATRX depletion additionally affects other essential cellular processes including mitosis, meiosis and DNA replication (Clynes et al., 2014, De La Fuente et al., 2004, Ritchie et al., 2008). ATRX-depleted human HeLa cell lines display an increasing number of

abnormal nuclear shapes and intranuclear DNA bridges (Ritchie et al., 2008). Furthermore, a mitotic delay, mainly due to a prolonged prometaphase to metaphase transition, is associated with congression of the condensed chromosome toward the metaphase plate. Increased numbers of misaligned chromosomes, chromosome decondensation and compromised cohesion of the sister chromatids are linked to ATRX depletion (Ritchie et al., 2008). Interestingly, conditional ATRX deletion in developing mouse forebrain results in newborn pups exhibiting reduced cortical and hippocampal size illustrating the essential role of ATRX for normal brain development in mouse (Berube et al., 2005). Mice brain sections showed that in absence of ATRX, apoptotic cells, cells with micronuclei and misaligned chromosomes increased significantly (Ritchie et al., 2008). Similar defects in chromosome alignment were observed during meiosis in ATRX depleted mouse oocytes (De La Fuente et al., 2004).

Although involved in the replication-independent deposition of the histone variant H3.3, work in mice revealed that ATRX is also required for the replication-dependent maintenance of chromatin. ATRX deficient myoblasts showed accumulation of p53 upon treatment with hydroxyurea, a DNA replication inhibitor, implying the activation of p53 checkpoint during S-phase (Huh et al., 2012). Additionally, ATRX deficient myoblasts had a delayed S phase completion and showed marks of genomic instability with activation of DNA damage response (DDR) and fragile telomeres. The presence of both DDR and chromatin instability support the hypothesis that DDR was not properly functioning (Huh et al., 2012, Lovejoy and Cortez, 2009). In WT myoblasts, ATRX enrichment was mainly observed at late-replicating chromatin domains, often associated with heterochromatin, implying a putative role of ATRX in supporting this process. Replication stress, resulting from stalled or collapsed DNA replication forks, activates ATM protein kinases. Activation of ATM should target downstream effectors

including p53 and recruit repair proteins such as Rad51 involved in the homologous recombination (HR) repair pathway. Efficacy of this process is a keystone for proper S phase progression and maintenance of the genomic stability. Although an increase of p53 was observed in ATRX deficient cells, the level of HR repair protein Rad51 was decreased compare to control (Huh et al., 2012).

Work performed in human cell lines and mouse ES cells further supported the relationship between ATRX, DNA replication and DDR. Indeed, ATRX deficiency is associated with various DNA replication defects such as increased sensitivity to hydroxyurea, prolonged S-phase, stalling of the replication fork and double strand breaks (Leung et al., 2013, Clynes et al., 2014). Furthermore, ATRX is localised to DNA damaged sites after lazer-induced microirradition (Leung et al., 2013). Probably, the most exciting discovery of these studies related to ATRX and the DDR is the DNA-independent interaction and co-localisation of ATRX with the MRN complex. Indeed, this complex, composed of the Rad50, MRE11 and NBS1 proteins, is known to play a role in double strand break repair (through HR or non-homologous end joining pathways) and to restart of the stalled replication forks (Clynes et al., 2014, Leung et al., 2013). Altogether, these findings suggest that in addition to gene expression, ATRX may play a role in DNA replication and damage response.

Supporting this hypothesis, ATRX deficiency has been associated with many defects in various cell types and tissues, sometimes mimicking the phenotypes observed in ATR-X patients including testicular defects and compromised muscle function (Bagheri-Fam et al., 2011, Huh et al., 2012). Mice with conditional ATRX KO in Sertoli cells have small testis associated with apoptosis of proliferating Sertoli cells (Bagheri-Fam et al., 2011). Although viable, skeletal muscle conditional ATRX KO in mouse resulted in smaller littermate. The body mass discrepancy was still significant

three weeks after birth but not in adulthood. Nevertheless, conditional ATRX KO adult mice showed abnormal recovery after muscle injury associated with deficient expansion of myoblast cells and a reduction of regenerative muscle fibers. By contrast, conditional ATRX KO occurring in terminally differentiated cells had no significant effect (Huh et al., 2012).

Replicative stress can result in the formation of double strand breaks. Yet, this damage has been shown to silence a reporter gene even 4kb away from the break sites (Shanbhag et al., 2010). G4 structures are potentially a common factor that bridges gene expression, DNA replication and DDR. Indeed, *in silico* and *in vivo* analyses using G4 ChIP-seq have highlighted the presence of G4 forming sequences at regulatory elements whereas *in vitro* G4-induced polymerase stalling experiments have identified approximately 700,000 G4 forming sequences in human (Hegyi, 2015, Hänsel-Hertsch et al., 2016, Chambers et al., 2015). Actually, the latter study used the ability of G4 forming sequences to disrupt replication to identify them (Chambers et al., 2015). Experiments focusing on G4-unwinding helicases such as Pif1 deepened our understanding of G4 effects on genome integrity. Thus, experiments in yeast containing a human G-rich sequence, able to form G4 *in vitro,* showed a genetic instability of the human sequence in the absence of Pif1 (Ribeyre et al., 2009). The data suggested that the genetic rearrangements were supported by homologous recombination. After mutation of the human sequence to impair its G4 forming potential, the sequence remained stable in the absence of Pif1 (Ribeyre et al., 2009).

To form G4, the sequence needs to be partially single stranded. Thus, these structures are more likely to form during processes such as DNA replication, transcription or

repair during which DNA duplex is transiently melted into single strands. Studies from the Sale group in chicken DT40 cells have highlighted a link between G4 associated defects in DNA replication and the perturbation of local gene expression. Thus, the low fidelity polymerase REV1, involved in translesional DNA synthesis during replication has been shown to be important for the maintenance of the epigenetic memory during the replication of G4 forming sequences (Sarkies et al., 2010). The maintenance of epigenetic marks, through cell division, is supported by the tight coupling between DNA synthesis processivity and the recycling of the modified parental histones. During replication, a mixture of newly synthesised histones and recycled parental histones are incorporated into the daughter strands. During this process, the re-deposition of the parental histones carrying PTM from the parental chromatin into the newly synthesised daughter strands is supposed to be essential to then propagate the PTM to the newly synthesised histones. By disrupting the DNA synthesis processivity and the histone recycling, the parental histones cannot be correctly incorporated and therefore their information cannot be transmitted to the newly synthesised histones; thus, resulting in the loss of the epigenetic memory. This phenomenon has been observed in absence of REV1 at G4 forming sequences where gap filling occurs to complete replication of the damaged DNA templates. Depending on the locus, this phenomenon resulted in de-repression of the repressed genes at originally silenced loci or loss of activation at transcribed loci (Sarkies et al., 2012, Sarkies et al., 2010). Similar effects associated with G4-induced epigenetic instability were observed upon mutations of helicases that are able to unwind G4 such as FANCJ, BLM and WRN (Sarkies et al., 2012).

The association between G4 stabilisation, replicative stress, epigenetic instability and aberrant gene expression was further supported by two studies focussing on the effects of nucleotide pool depletion and impairment of DNA Primase-Polymerase (Primpol)

on loci containing G4 forming sequences. Thus, depletion of the nucleotide pool using hydroxyurea also resulted in epigenetic instability at G4 forming sequences (Papadopoulou et al., 2015). Then, a recent study focussing on Primpol highlighted the importance of efficient by-passing of the G4 block for the maintenance of the epigenetic integrity. This primase is able to bind G4 and promote the repriming of DNA synthesis, very closely coupled to it. Despite its inability to unwind G4, it was able to prevent epigenetic instability at G4 forming sequences (Schiavone et al., 2016).

Given the G4 forming potential of a subset of ATRX binding sites, including at telomeres and ψζVNTR, it is legitimate to suspect that such phenomena could occur at some of them in the absence of ATRX. Indeed, in embryonic mouse brain, ATRX loss results in accumulation of replicative damages at telomeres and PCH. In neuroprogenitors, ATRX deficiency is associated with an increased sensitivity to G4 stabilising drug treatment suggesting that ATRX is playing a role in the replication of the G4 forming sequences (Watson et al., 2013).

1.3.8. ATRX deficiency and alternative lengthening of telomere associated cancers

In addition to ATMDS, ATRX mutations have been recently associated with a subset of cancers using the alternative lengthening of telomere (ALT) pathway, which maintains telomeres (Schwartzentruber et al., 2012, Heaphy et al., 2011, Lovejoy et al., 2012). Critical for genomic stability and immortality, the maintenance of telomeres in cancer cells is often associated with the reactivation of the telomerase enzyme. This enzyme is usually expressed only during development and is restricted to germ cells and tissue stem cells in adults. However, in approximately 10% to 15% of the cancers,

cells use ALT pathways, rather than telomerase reactivation, to maintain their telomeres and achieve immortalisation (Lovejoy et al., 2012). Although still unclear, the ALT pathway is thought to be dependent on recombination between telomeric sequences. Strikingly, recent studies have identified defects in ATRX in 90% cells lines using the ALT pathway for survival as well as in ALT cancers such as neuroblastomas and pediatric glioblastomas (Heaphy et al., 2011, Lovejoy et al., 2012, Schwartzentruber et al., 2012). Regarding the effects of ATRX deficiency on the maintenance of the genomic integrity, it is tempting to speculate that ATRX may act as a suppressor of the ALT pathway. However, studies in mouse and Hela cells showed that ATRX loss alone is insufficient to trigger ALT pathway (Clynes et al., 2014). Yet, in U2OS cells lines that are ALT positive and ATRX deleted, ectopic expression of ATRX results in the suppression of this pathway in a DAXX-dependent manner (Clynes et al., 2015). Supporting the role of ATRX in DNA replication, including at G4 forming sequences, the re-expression of ATRX is correlated with a decrease in replication fork stalling, whereas, stabilisation of G4 structures using pyridostatin impedes the suppression of the ALT pathway by ATRX re-expression (Clynes et al., 2015). Furthermore, ATRX re-expression led to a relocalisation of the MRN complex. Thus, a loss of the MRN complex member MRE11 from the telomeres suggested that ATRX may sequester MRN complex and therefore interfere with MRN-mediated telomeric recombination (Clynes et al., 2015). Besides, yeast mre11 shows a high binding affinity for G4 forming sequences (Ghosal and Muniyappa, 2005).

In mouse model used to study human glioblastoma, loss of ATRX was associated with faster tumour growth and decreased median survival (Koschmann et al., 2016).

Furthermore, experiments based on ALT positive tumours support observations in mouse ES cells and showed a lower rDNA copy number than ALT negative ones

(Udugama et al., 2018). Interestingly, the ATRX complex is not the only chromatin remodeler complex/histone chaperone that has been associated with ALT pathways as illustrated by Asf1 and HIRA (O'sullivan et al., 2014, Jiang et al., 2011).

1.4. Summary, aim and objectives

Associated with severe diseases including the ATR-X syndrome, ATMDS and some cancers using the ALT pathway, ATRX has been showed to be important for fundamental cellular processes as diverse as regulation of the gene expression, DNA replication, cell division and DDR. Over two decades, ATRX has been highly associated with heterochromatin regions such as centrometric, pericentric and rDNA hetorochromatin. Its localisation to heterochromatin is further supported by its structure that contains an ADD domain, which binds H3K9me3/H3K4me0, and heterochromatin associated interacting motifs such as HP1.

However, outside of the constitutive heterochromatin environment, only a very limited subset of targets has been identified in human. Actually, at the beginning of the current study, only one ATRX ChIP-seq dataset performed in human was published (Law et al., 2010). Almost four years later, a single additional study has investigated the distribution of human ATRX in a genome-wide manner (Valle-García et al., 2016) and the role of ATRX at euchromatin regions remains poorly understood. This may be partly due to the difficulty in producing high quality ATRX ChIP-seq datasets.

The overall goal of this study has been to investigate in greater depth the role of ATRX in the regulation of gene expression in health and ATR-X syndrome.

Therefore, I first optimised the ATRX ChIP-seq protocol and produced high quality ChIP-seq datasets. Second, I analysed the genome-wide signatures of the ATRX

binding sites to uncover the presence of potential subgroups. Third, I identified novel candidate genes that displayed aberrant expression pattern in ATR-X patients derived lymphoblastoid cell lines (LCL). I narrowed down the list of candidate genes by selecting the most likely direct ATRX targets based on ATRX binding sites identified by ChIP-seq. Then, I investigated the effects of ATRX deficiency on the chromatin environment of the candidate genes. Finally, I assessed the relevance of these findings by further analysing the ATRX distribution in human erythroid cells.

Chapter 2. Materials and Methods

2.1. Cell culture

2.1.1. Lymphoblastoid cell lines

LCL derived from ATR-X patients and healthy controls (***Table 2.1***) were stored in liquid nitrogen in 10% Dimethyl-Sulphoxide (DMSO) .Upon recovery, LCL thrawed and washed with RPMI medium supplemented with 15% FCS, 1% PenStrep and 1% glutamine. Then, LCL were incubated with supplemented RPMI media at 37°C and 5% CO_2 until reaching the desidered number which depend on the experiment.

Sample	Mutation	Batch
JC-K1650N	4950G>T; K1650N	1
MF-D2035V	6104A>T; D2035V	1
RS-P190S	568C>T; P190S	1
RH-R246C	736C>T; R246C	1
FF-Ctrl	Control	1
FC-Ctrl	Control	1
BH-Ctrl	Control	1
AF-V1538G	4613T>G; V1538G	2
GH-176_177insQ	528insCAA; 176_177insQ	2
AH-S_insertion	4317G>A; S insertion 53bp (fs)	2
SB-R246C	736C>T; R246C	2
ND-R246C	736C>T; R246C	2
CB-Ctrl	Control	2
GD-Ctrl	Control	2
GF-Ctrl	Control	2
CN-C1614R	4840T>C; C1641R	3
MS-Y2084H	6250T>C; Y2084H	3
JJ-C200S	599G>C; C200S	3
PG-R246C	736C>T; R246C	3
SG-R246C	736C>T; R246C	3
PN-Ctrl	Control	3
KG-Ctrl	Control	3
AS-Ctrl	Control	3

Sample	Mutation	Batch
AN-S_insertion	6218-12574G>A; S insertion 128bp (fs)	4
DL-delex1	del exon 1	4
CK-L192F	576G>C; L192F	4
MC-Y2163C	6488A>G; Y2163C	4
LL-R246C	736C>T; R246C	4
RN-Ctrl	Control	4
JL-Ctrl	Control	4
SK-Ctrl	Control	4
MtF-D2035V	6104A>T; D2035V	5
RA-S576X	1727C>A; S576X	5
SW-G249D	746G>A; G249D	5
AN-H1609R	4826A>G; H1609R	5
TS-A238P	712G>C; A238P	5
DM-V178_K198del	536A>G; S c532-594del; V178-K198del	5
OS-R246C	736C>T; R246C	5
JMK-R246C	736C>T; R246C	5
KT-R246C	736C>T; R246C	5
TA-ctrl	Control	5
PH-Ctrl	Control	5
FP-Ctrl	Control	5
BT-Ctrl	Control	5
AJ-Ctrl	Control	5
HC-Crtl	Control	5
SOG-Ctrl	Control	5
SS-Ctrl	Control	5

Table 2.1: List of the LCL used in this study. LCL derived from healthy controls (green), LCL derived from ATR-X patients displaying the common mutation R246C (red) and ATR-X patients covering a representative range of ATRX mutations (yellow). The batch numbers of associated with the microarray experiment are indicated in the right column. Samples from a same batch were cultured together and RNA was extracted at the same time.

2.1.2. CD34 separation

CD34+ separation was performed by Caroline Scott. CD34+ cells were isolated from either 50mL of healthy adult peripheral blood or leucocyte reduction filters (obtained from the NHSBT) using Ficoll-Histopaque (Sigma) density centrifugation and positive bead selection (Miltenyi). The CD34+ fraction was cryopreserved in freezing media consisting of 90% FBS (Gibco) and 10% DMSO.

2.1.3. Culturing of CD34+

CD34+ culture and differentiation were performed by Caroline Scott. Frozen CD34+ cells were recovered into base medium (IMDM Source BioScience UK Limited containing 3% (v/v) AB Serum, 10 μg/mL insulin, 3 U/mL heparin (all from Sigma-Aldrich, Poole, UK), 2% (v/v) fetal bovine serum (Gibco), 200 μg/mL Human holo-transferrin (R&D Systems Minneapolis, MN, USA). For the expansion phase of our three-stage erythroid culture system the media was supplemented with 3 U/mL EPO (Eprex Janssen Cilag UK), 10 ng/mL stem cell factor (SCF) and 1 ng/mL IL-3 (Peprotech, London UK). Cells were maintained at density of $2x10^5$ cells until day 7 when they were transferred to phase 2 differentiation media consisting of base media with 3 U/mL EPO and 10 ng/mL SCF. From this point cells were staged using cytospins and FACS using a panel of 6 flurophore-conjugated monoclonal antibodies (***Table 2.2***). Staining was conducted in FACS buffer containing 10% FBS and 90% PBS (both from

Gibco, UK) and samples analysed using the Attune NxT (Life Technologies, California USA). Gates were set using fluorescence-minus-one controls run for each differentiation and data analysed using FlowJo software v 10.0.8r1. Cells were collected at day 10.

Marker	Protein	Channel	Supplier	Catalogue Number
CD235a	Glycophorin A	PE	BD Bioscience	555570
CD71	Transferrin	PerCP Cy5.5	Biolegend	334114
CD49D	α-Integrin	APC	BD Bioscience	561892
CD34	CD34	PE/Cy7	Biolegend	343616
CD233	Band3	FITC	IBGRL	9439FI
CD36	Platelet glycoprotein	APC/Cy7	Biolegend	336213
Hoechst 33258	viability dye	violet	Invitrogen	H3569

Table 2.2: list of the FACS markers used to monitor the CD34+ differentiation

2.2. DNA extraction

Genomic DNA (gDNA) was extracted using Gentra Puregene Cell Kit (Qiagen) following manufacturer's protocol. gDNA quality was checked on agarose gel or using Agilent Genomic DNA ScreenTape System (Agilent).

Polymerase chain reaction (PCR) product DNA was purified with QIAquick PCR purification kit (Qiagen) or MinElute PCR Purification kit (Qiagen) following manufacturer's protocol.

PCR product DNA fragments from agarose gel band were purified using the QIAquick gel extraction kit (Qiagen) following manufacturer's protocol.

gDNA extraction protocol for DIP and BIO-CAP experiments will be described in the related sections.

2.3. RNA extraction

RNA was extracted using RNeasy Mini kit from Qiagen. 4 x 10^6 cells were used as starting material. Lysates were homogenised using QIAsherdder spin column (Qiagen). Samples were submitted to on-column DNase digestion using RNAse-Free set DNase Set (Cat no. 79254) and RNA quality assessed by Nanodrop and integrity was assessed using Agilent RNA ScreenTape assay system (Agilent). Only samples displaying a RINe score higher than 8 were selected. Note that RNA extractions for the microarray experiments were carried out by batches of 7 to 16 cell lines resulting in potential batch effect for further analysis (***Table 2.1***). All the samples present in a same batch were processed at the same time from the cell culture to the microarray experiment.

2.4. Protein extraction and Western blot

5x10^6 cells per sample were washed twice with PBS and centrifiged at 1,000 rpm for 5 min to remove supernatant. Cell pellets were placed on ice and resuspended in 4°C high salt RIPA buffer (1 M NaCl, 1% NP-40, 1% Na-Deoxycholate, 0.2% SDS, 50 mM Tris-HCl ph 7.4) supplemented with 1X of with protease inhibitor (Roche). Samples were incubated for 25 min with occasional mixing by pipetting. Samples were syringe 30 times using a 25G gauge needle and spin at 13,000 rpm for 10 min at 4°C. The supernatant was transfer to a fresh tube and stored at -80°C or immediately prepared for the electrophoresis.

For electrophoresis, equal volume of lysate and 2x loading buffer with 0.2M DTT were mixed and heated at 95°C for 5 min. Electrophoresis was performed in a 4-12% Bis-Tris gel (ThermoFisher Scientific) with 1X MOPS buffer. The sizes of the bandes were scaled using PageRuler Plus ladder (ThermoFisher Scientific). Gels were run a constant voltage of 120V. Size separated protein extracts were transferred from the gel to a

polyvinylidene difluoride membrane (ImmobilonP; Millipore) overnight. Then, it was blocked with 5% milk in TBS-T (Tris buffered saline + 0.1% Tween-20) for 1 hour. The membrane was incubated at RT for 1 hour in TBS-T with 2.5% milk with primary antibody anti-alpha-tubulin (abcam-ab7291) or anti-ATRX (39f prepared by Sue Butler, Santa Cruz-H300 sc-15408 or Abcam-ab97508). Subsequently, the membrane was washed three times with TBS-T and incubated for 1hour in TBS-T with 2.5% milk with appropriated horseradish peroxidase-conjugated secondary antibody. After, three washes with TBS-T, the bands were visualised on radiographic film.

2.5. Southern blot

15 µg of gDNA samples were digested with BlpI, SacII, MslI, MluI-HF, PsiL and/or PmlI restriction enzymes (NEB) at 37°C for at least 6 hours. Digested samples were run on a 1% agarose gel at 40V overnight. After gel visualisation to validate the digestion efficiency, the gel was incubated 1 hour in 1M NaOH solution and subsequently neutralised with Neutralisation buffer (1M Tris, 3M NaCl, 80 mL/L concentrated HCl ph 7.4) for 2 hours. DNA was transferred onto Zeta-probe nylon membrane (Biorad). Then the membrane was hybridized at 42°C over night with appropriated probes labelled with ^{32}P. The membrane was washed with 2X SSC (0.03 M Sodium citrate and 0.3 M NaCl) and then with 0.1% SSC and developed.

2.6. Primeview Affymetrix Microarray

Microarray experiment was done using PrimeView™ Human Gene Expression Array Plate and following the 100 or 81/4-format 3' IVT PLUS Reagent Kit User Manual (Affymetrix). 250ng of RNA were used as starting material. In short, RNA samples were used for cDNA synthesis following by the production of Biotinilated-labbled

cRNA. Labbeled cRNA was then purified, fragmented and hybridised on arrays. Quality controls were performed during the process using Agilent Technologies Agilent RNA ScreenTape System. After staining, the arrays were scanned. Results were further analysed using Affymetrix® Transcriptome Analysis Console (TAC) and Affymetrix® Expression Console™ Softwares (Affymetrix).

2.7. Power Calculation and Correction for the Batch effects in microarray experiment analyses

The power calculation and correction for the batch effects and subsequent analysis of the microarray data were performed by Emmanouela Repapi. The power calculation was determined using samr package and the ssize package from R. In addition to the microarray analysis done using the manufacturer's instruction (described above), a parallel analysis has been done allowing the correction of the batch effect generated during RNA extraction using R script (with limma (Ritchie et al., 2015), affy (Gautier et al., 2004) and genefilter (Gentleman et al., 2016) packages. To increase the power, half of the probes were removed after normalisation based on their very weak signal or weak variation across samples (Bourgon et al., 2010).

2.8. High-throuput qRT-PCR by Fluidigm Biomark HD system

High-throughput qRT-PCR was performed with preliminary extracted RNA using a Fluidigm 192:24 Gene Expression Chip and Taqman assay described in ***Table 2.3***. Data were analysed using the Fluidigm Real-time PCR analysis 3.1.3 software and were normalised to the mean of two housekeeping genes (RPL13A and GAPDH).

Gene Symbol	Assay ID	Fluidigm Biomark	qPCR
ACTB	Hs01060665_g1	x	x
ADNP2	Hs01108772_m1		x
AHI1	Hs01053982_m1	x	x
ASMTL	Hs00992779_m1		x
ATF7IP2	Hs01101353_m1	x	x
C9orf57	Hs01370396_m1	x	x
CCR1	Hs00928897_s1	x	x
CD200R1	Hs00793597_m1	x	
CD99	Hs00908458_m1		x
CEP70	Hs00225806_m1	x	
CSTA	Hs00193257_m1	x	
DHRSX	Hs00376948_m1		x
GAPDH	Hs03929097_g1	x	x
GDA	Hs00188737_m1	x	
GPAT2	Hs00999810_g		x
LPAR6	Hs00271758_s1		x
MEIS2	Hs00542636_m1	x	x
MYC	Hs00153408_m1	x	
MYOM2	Hs01556320_m1	x	
NME4	Hs00359037_m1	x	x
PBX4	Hs00257935_m1	x	x
PDE6B	Hs01557860_m1		x
RPL13A	Hs04194366_g1	x	x
TAB2	Hs00248373_m1	x	
UST	Hs01053964_m1	x	
ZNF506	Hs00869072_g1	x	x
ZNF554	Hs01014440_m1	x	
ZNF555	Hs00328296_s1	x	x
ZNF57	Hs00380549_m1	x	x
ZNF595	Hs00745818_s1	x	x
ZNF718	Hs00799585_sH	x	x

Table 2.3: Taqman assays used in this study

2.9. Cell line validation

PCR experiments to validate the mutations present in each cell lines were carried out using for each reaction: 5 µL of 10X Applied Biosystems Buffer II, 3 µL of 25mM MgCl2, 0.4 µL of 25mM dNTP, 0.2 µL of AmpliTaq Gold DNA polymerase (5U/µl), 0.05 µL of Pwo (5U/µl), 1 µL of 10 µM Forward oligo, 1 µL of 10 µM Reverse oligo (see table below for oligo sequences), 38.35 µL of water and 1µL of template (100ng/µl). The cycle conditions were set up as following; an initial denaturation at 95°C for 10 minutes, followed by 35 cycles of denaturation temperature sets at 95°C for 1 minute, annealing temperature for 1 minute (***Table 2.4***) and the extension temperature (72°C) for 1 minute. Finally, a final extension at 72°C for 5 minutes was performed. Samples were run on a 2% agarose gel to check amplification.

Exon	Oligonucleotide sequence 5' to 3'	Annealing Temperature (°C)
8	GGAATTTGCCAAGGTTGTC	58
	TCTTCAAGACTGTGCCCTCA	
9	TTGCCAATTTTTGCTTTGGT	58
	TCCAGGGTTTTATGGAAAAGAA	
10A	CAGTTTCCTGAAAGAAGGGAAT	58
	CAGAGCCAGAACAGGAATCA	
10B	GCAGCAGTGGAACTGAACAA	60
	TCATTTTCAACCAAATGCTCA	
15	GGAATGCATACAGAGGTAGAACA	58
	TTCAAGCATGTGGTAAATGTCA	
17	TTTTGGAAAATCTCCCTTGG	58
	AGCTGCCTCACACCCAAATA	
19	AACGAAGGTGACAATCTGTTTTC	58
	TTCCCACTGAAATATGCATCAC	
27	CCCCATGGGTAGGTCTTTTT	58
	GGAAGGAAGGAAAAGCAACA	
29	TGAGCAAGGTGGAAAATCTG	57
	TGACTGTACCTTGCATTTGCT	
30	AAGAAATGAATTCTCTGAACTCTTGA	58
	CCAACTTTGTTTCCCTCTCTG	

Table 2.4 list of primers and their localisation in the ATRX gene used in this study to validate the different cell lines.

2.10. Quantitative-PCR

qPCR based on Taqman assay were performed using Taqman probes with a 2x Taq mastermix (Applied Biosystems) and Taqman expression Assay (ThermoFisher Scientific) detailed in ***Table 2.3***. qPCR experiments based on SYBR green assay were performed using SYBR green master mix (Applied Biosystems). The primers used for these experiments are listed in ***Table 2.5***.

Name	Forward	Reverse
AHI1	CCCACTCCTCAGGTTTCAAG	GGCCTTTGGAATAGGAGTTG
ATF7IP2	GCTGCCATTGGTCCATAATC	CCTTGGCTGCTTGTGATTTC
DIST	GAGATGCTGGAGTCAGGACCAT	AGGAGTCAGGAGCAGCAGTCA
G4 negative/CpG island negative control	GCAAAAGCGTTTGGGTATTC	TGTCCTTTCAAAAGGCCTACC
GAPDH promoter	TATAAATTGAGCCCGCAGCCT	AAAGAAGATGCGGCTGACTGT
GAPDH gene body	ACCTGTGCTCCCACTCCTGATTTC	TGCCAAGTTGCCTGTCCTTCC
HBB Promoter	AACTGTGTTCACTAGCAACCTCAAA	ACAGGGCAGTAACGGCAGACT
rDNA	CGGGCGTGGAATGCGAGTG	GGGGTCTGATGAGCGTCGGC
UST	CGCCCACTTTCTTTTCTCAG	AACAGAGCAAGGCAACACAG
ZNF506	AGCTGGGCAATGAGAACTTG	AGAGTGCCAGTTCGACATCC
ZNF555	GGAACTGCCCAATTTACAGC	TGATTAGACAGAGCTCGGGAAG
ZNF57	GTCCGGAAAACTGTCCAATC	TTGGACACTACTGCAAACCTG
ZNF718	AGGCCCATCAAAACATGG	GATTTGCTTCGCTTCCTCAC

Table 2.5: Primers used for qPCR based on SYBR green assays.

2.11. Reverse transcription

cDNA was produced using High Capacity RNA-to-cDNA Kit (Applied Biosystems) using 1.5µg of RNA as starting material and following the manufacturer's instructions.

2.12. ATRX ChIP-seq

All ATRX ChIPs have been carried out following the optimised protocol described in (Truch et al., 2018) at the exception of only one ChIP- seq experiment. This ChIP-seq, for which results were displayed in Chapter 3, was performed by Helena Ayyub following the protocol described in (Law et al., 2010) and has been sequenced by the Wellcome Trust Centre using Hi-seq (Illumina).

Cell cross-linking: 10^8 cells were washed with PBS and spun down. Cells were fixed with 2 mM EGS in 20 mL of PBS for 45 minuntes with constant mixing. A second cross-linking step was performed by adding formaldehyde to the sample to a final concentration of 1% and incubated for 20 minutes with constant mixing. Sample was quenched by adding glycine to a final concentration of 125 mM, followed by a 2 minutes incubation with constant mixing. Sample was then spun down at 4°C and washed twice with 4°C PBS. The washed pellet was then snap frozen on dry ice and stored at -80°C (after optimisation, the samples were stored for less than 72 hours).

Cell lysis and sonication: Cell pellet was thrawed on ice and re-suspended in 10 mL of 4°C Lysis Buffer 1 (100mM Hepes-KOH, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40 and 0.25% Triton X-100) complemented with protease inhibitor cocktail (1X final). Sample was incubated for 10 minutes at 4°C under constant mixing. After centrifugation, the pellet was re-suspended in 10 mL of 4°C Lysis buffer 2 (200mM NaCl, 1mM EDTA, 0.5mM EGTA and 10mM Tris-HCl) complemented with protease inhibitor cocktail (1X final). The sample was then spun down at 2,500 rpm for 2 minutes at 4°C. Pellet was re-suspended in 3 mL of 4°C Lysis buffer 3 (1mM EDTA, 0.5mM EGTA, 10mM Tris-HCl at pH 8, 100mM NaCl, 0.1% Na-Deoxycholate and 0.5% N-lauroyl sarcosine) complemented with protease inhibitor cocktail (1X final) and syringe at least 10 times each sample with a G27 needle.

Sample was either sonicated using the Diagenode Bioruptor (during optimisation only) or the Covaris. For the sonication using the Covaris, samples were shared in three milliTUBE 1 mL tubes with AFA fiber. After optimisation, each milliTUBE 1 mL tube was sonicated for 25 min with a 12.5% duty and 5 intensity at 4°C to yield DNA fragments 100 and 250 bp. After sonication, 1/10 volume of 10% Triton X-100 was added to each sample. Samples were centrifuged at 14,000 rpm for 10 minutes at 4°C. The supernatants were collected. The three aliquots of 1mL originated from the same sample were pooled and 50 µL aliquots were collected as input and stored at -20°C.

Pre-clearing and chromatin immunoprecipitation: pre-clearing and chromatin immunoprecipitation were performed at 4°C. Sonicated samples were pre-cleared with a mix of Protein A and G Dynabeads (1:1 ratio) pre-blocked in 0.5% BSA, for 1 hour. Pre-clearing beads were precipitate the beads from the chromatin using a magnetic stand. The supernatant was divided in two. On half was mixed with Protein A and G Dynabeads (1:1 ratio) pre-coated with 10µg of anti-ATRX antibody (Santa Cruz-H300 sc-15408 or Abcam-ab97508) and the other half with Dynabeads pre-coated with anti-rabbit IgG control. Samples were incubated overnight under constant mixing.

Washing, eluting and reverse cross-linking and DNA purification: samples were washed five times with 4°C Wash buffer (50mM Hepes at pH 7.6, 1mM EDTA, 0.7% Na-Deoxycholate, 1% NP-40 and 0.5M LiCl) complemented with protease inhibitor cocktail (1X final) and once with 4°C PBS complemented with protease inhibitor cocktail (1X final). Alternatively, the washes were done using the wash buffers from the Chromatin Immunoprecipitation Assay Kit (Merck). Subsequently, the chromatin was eluted with 250 µL of Elution buffer (0.5% SDS and 100mM NaHCO3) at room temperature for 15 min. The elution was repeated for 20 min at 65°C. 450 µL of Elution buffer was added to the input. Eluted chromatin and input were reverse-cross-linked

with 0.2 M NaCl at 65°C for 4 hours. After proteinase K treatment, the DNA was purified by phenol/chloroform and precipitated with ethanol, sodium acetate and carrier glycogen or GlycoBlue. Samples were assessed by qPCR for positive and negative enrichment.

2.13. Single step cross-linking ChIP

The single step cross-linking ChIP-seq was performed using the ChIP Assay kit (Millipore).

Cell cross-linking: $5x10^6$ to 10^7 cells were used to carry out ChIP experiments. Cells were washed with PBS and fixed with 1% formaldehyde for 10 minutes with constant mixing. Samples were quenched by adding glycine to a final concentration of 125 mM, followed by a 2 minutes incubation with constant mixing. Samples were then spun down at 4°C and washed twice with 4°C PBS. The washed pellets were then snap frozen on dry ice and stored at -80°C.

Cell lysis and sonication: Cell pellets were thrawed on ice and re-suspended in 4°C SDS Lysis Buffer complemented with protease inhibitor cocktail (1X final). Samples were incubated for 10 minutes at 4°C and sonicated with the Covaris using either the milliTUBE 1 mL tubes with AFA fiber or pre-split snap-cap 6x16mm microTube. Using the milliTUBE 1 mL tube samples were sonicated for 8 minutes with a 5% duty and 4 intensity at 4°C. Alternatively, the microTube samples were sonicated for 8 minutes with a 2% duty and 3 intensity at 4°C. Then, Samples were centrifuged at 14,000 rpm for 10 minutes at 4°C. The supernatants were collected and diluted 1 in 10 with the ChIP dilution buffer complemented with protease inhibitor cocktail (1X final). 10 to 50 µL aliquot was collected as input and stored at -20°C.

Pre-clearing and chromatin immunoprecipitation: pre-clearing and chromatin immunoprecipitation were performed at 4°C. Sonicated samples were pre-cleared with 60 µL of protein A/Salomon sperm DNA Agarose beads (Millipore Kit) or ChIP-grade protein G agarose beads (Cell Signaling Technology), for 1 hour. The pre-cleared samples were then incubated with the appropriate antibody (***Table 2.6***) and incubated overnight under constant mixing. Subsequently, 60 µL of protein A/Salomon sperm DNA Agarose beads (Millipore Kit) or ChIP-grade protein G agarose beads (Cell Signaling Technology) was added to the samples and incubated for 1 hour.

Target	Reference	Supplier
NTD PolII	sc899	Santa Cruz
CTD RNA polII	ab817	Abcam
H3	ab1791	Abcam
H3.3	09-838	Millipore
H3K27ac	ab4729	Abcam
H3K27me3	ab6002	Abcam
H3K36me3	ab9050	Abcam
H3K4me1	ab8895	Abcam
H3K4me1	ab195391	Abcam
H3K4me3	07-473	Merck
H3K4me3	ab8580	Abcam
H3K9me3	ab8898	Abcam
Rabbit IgG	X0903	Dako

Table 2.6: List of the antibodies used for single cross-link ChIP protocol

Washing, eluting and reverse cross-linking and DNA purification: the samples were sequentially washed with Low Salt Buffer, High salt buffer, LiCL Buffer at 4°C. Then, two washes were performed with TE buffer at room temperature. All the buffers were complemented with protease inhibitor cocktail (1X final). Subsequently, the chromatin was eluted twice with 250 µL of Elution buffer (0.5% SDS and 100mM NaHCO3) at room temperature for 15 min. 450 µL of Elution buffer was added to the input. Eluted chromatin and input were reverse-cross-linked with 0.2 M NaCl at 65°C for 4 hours. The samples were sequentially treated with to RNase A and then with proteinase K. The DNA was purified by phenol/chloroform and precipitated with ethanol, sodium acetate and carrier glycogen or GlycoBlue. Samples were assessed by qPCR for positive and negative enrichment.

2.14. Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)

ATAC-seq was carried out as described in (Buenrostro et al., 2013) using 80, 000 cells as starting material. Cell pellets were first resuspended in 4°C ATAC lysis buffer (10mM Tris-HCl, 10mM NaCl, 3mM MgCl2 and 0.1% (v/v) IGPAL) and washed once with PBS. After a 10 min centrifugation at 4°C and 500g, pellets were resuspended in transposition reaction mix (Illumina Nextera DNA Library Prep Kit FC-121-1030) for 30 min at 37°C. As control, transposition reactions were also performed on gDNA. Transposed DNA were extracted using MinElute PCR Purification kit (Qiagen) and PCR amplified and indexed. Indexed libraries were purified using QIAquick PCR purification kit (Qiagen).

2.15. Biotinylated CxxC affinity purification (Bio-CAP)

Bio-CAP experiments were performed as described in (Blackledge et al., 2011).

In brief, genomic DNA was extracted from 2.5 x 10^7 cells using 950 µL of extraction buffer (20 mM Tris HCl (pH 8.0), 10 mM EDTA, 100 mM NaCl and 0.5% SDS) followed by RNaseA/T1 (Thermo Scientific) treatment and phenol/chloroform extraction. gDNA were sonicated to an average size of 200 bp and diluted to 17.5 mg/mL in CAP100 buffer (12.5% glycerol, 0.1% Triton-x-100, 20mM HEPES pH 7.9 and 100mM NaCl). 100 µL aliquot was retained as input. NeutrAvidin Agarose Resin (Thermo Scientific) was conjugated with biotinylated hKDM2B ZF-CxxC protein (kindly provided by Rob Klose). 500 mL of diluted sonicated DNA were incubated with conjugated CxxC resin for 1 hour at 4°C. Serial washes and elutions using increasing salt concentration were performed. First, two washes with 1 mL of CAP100 buffer were carried out, followed by two elutions with 50 µL of CAP300 (12.5% glycerol, 0.1% Triton-x-100, 20 mM HEPES pH 7.9 and 300 mM NaCl) and two elutions with 50 µL of CAP500 (12.5% glycerol, 0.1% Triton-x-100, 20 mM HEPES pH 7.9 and 500 mM NaCl). Subsequently, the elution process was repeated using buffers with 700 mM and 1 M NaCl sequentially. DNA purification from eluted samples and input were performed using QIAquick PCR purification kit (Qiagen) and assessed by qPCR. Samples derived from elutions using 700 mM and 1 M NaCl were used for sequencing.

2.16. RNA-DNA immunoprecipitation (DIP)

gDNA extraction: 10^7 cells per sample were washed with 4°C PBS. Cell pellets were resuspended in 600 µL of 4°C Cell Lysis Budder (0.5% NP-40, 85 mM KCL and 5 mM PIPES) and incubated on ice for 10 minutes. Samples were centrifuged at 2400rpm for

5min at 4°C. Pellets were resuspended in 800 µL of Nuclear Lysis Buffer (1% SDS, 25 mM Tris-HCL pH8 and % mM EDTA) and incubated on ice for 10 minutes. 20 µg of Proteinase K (Thermo Fisher) were added and samples were incubated over night at 45°C. Subsequently, 400 µL of gDNA isolation buffer (3M Potassium Acetate and 11.5% Glacial Acetic Acid) were added and samples were centrifuged at 13,000 rpm for 5 minutes at 4°C. Supernatants were collected and DNA was precipitated with isopropanol. Pellets were washed with 70% ethanol and air dried before resuspension in water.

RNase H treatment: 50-100 µg of DNA was equally divided in two. Half of the sample was treated with 5 µL of recombinant RNAse H (5U/µl, NEB) in a total volume of 100 µL for 4 hours. The other half was used as control and incubated with 5 µL of water instead. Subsequently, samples were diluted with 300 µL of IP Dilution Buffer (0.001% SDS, 1.1% Triton X-100, 0.12 mM EDTA, 16 mM Tris-HCl pH8, 0.160 mM NaCl and protease inhibitor cocktail).

Sonication: Samples were sonicated for 15 minutes (30 seconds ON, 30 secondes OFF) with high power setting using the Diagenode Bioruptor.

Pre-clearing and IP: Sonicated samples were diluted with IP Dilution Buffer to a final volume of 3 ml. 100 µL of each samples were collected as input before pre-clearing of the remaining sample with Protein A Dynabeads pre-blocked with 1% BSA for 1 hour. Pre-cleared samples were collected and incubated over night at 4°C with either 10 µg of S9.6 antibody specific to RNA:DNA hydrids (produced by Sue Butler) or normal mouse IgG as control. Samples were then incubated for 2 hours at 4°C with 60 µL of Protein A Dynabeads pre-blocked with 1% BSA.

Washes, elution and DNA purification: Beads were collected and sequentially washed for 5 min at 4°C with Buffer A (20 mM Tris-HCl pH8, 2 mM EDTA, 0.1% SDS, 1%

Triton-X 100 and 0.165M NaCl protease inhibitor cocktail), Buffer B (20 mM Tris-HCl pH8, 2 mM EDTA, 0.1% Triton-X 100 and 0.55M NaCl and protease inhibitor cocktail) and Buffer C (100 mM Tris-HCl pH8, 1 mM EDTA, 1% NP-40 and 0.01% Sodium Deoxycholate, 250 mM Li and protease inhibitor cocktail). Then beads were washed with Buffer D (100 mM Tris-HCl pH8, 1 mM EDTA and protease inhibitor cocktail) twice for 5 min at 4°C and once at room temperature. Subsequently, captured DNA was eluted twice in 250 µL Elution Buffer (1% SDS and 0.1 M NaHCO3). Then, eluted DNA and input samples were purified using QIAquick PCR purification kit (Qiagen).

2.17. G4 ChIP

G4 ChIP experiments were performed following the protocol described in (Hänsel-Hertsch et al., 2016) with few modifications.

Pyridostatin treatment: cells were treated with 2 µM pyridostatin or DMSO (control) for 28 hours and subsequently harvested.

Cell cross-linking, lysis and sonication: 10^7 cells were cross-linked for 15 min in 1% formaldehyde and quenched with glycine for 10 min. Chromatin isolation and preparation were performed using the Chromatrap following manufacturer's protocol. Chromatin was sonicated using a Bioruptor Pico Diagenode for 7 cycles (30 seconds ON/ 30 seconds OFF) to achieve an average fragment size of 100 to 500 bp.

Immunoprecipitation, washes, elution and DNA purification: 2.5 µL of sonicated chromatin was diluted in 44.5 µL intracellular salt solution (25 mM HEPES pH 7.5, 10.5 mM NaCl, 110 mM KCl, 130 nM CaCl2, 1 mM MgCl2) supplemented with 1% BSA and blocked for 20 minutes at 16°C. An aliquot was stored as input control before addition of the recombinant BG4 antibody (absolute antibody). Samples were incubated for 1 hour at 16°C. Subsequently 50 µL of a mix of protein A and G dynabeads, blocked

beforehand with intracellular salt solution supplemented with 1% BSA, was added to the samples. After 1 hour incubation at 16°C, beads were captured and washed three times with 200 µL wash buffer (10 mM Tris pH 7.4, 100 mM KCl, 0.1% (v/v) Tween 20) at 4°C, twice at 37°C and once again at 4°C. Then, beads were resuspended in 75 µL of elution buffer, containing 1× TE buffer with 20 µg proteinase K, and first incubated 1 hour at 37°C and then 2 hours at 65°C. Finally, DNA was purified using the MinElute kit.

2.18. Capture-C

Capture-C experiment was performed following the protocol described in (Davies et al., 2015).

3C Library preparation: $2x10^7$ cells were cross-linked with 2% formaldehyde for 10 min and quenched 0.125% Glycine. After being washed with cold PBS, cell pellets were re-susspended in 5 mL of 4°C lysis buffer (10 mM Tris pH 8, 10 mM NaCl, 0.2% Igepal CA-630, cOmplete Protease Inhibitor Cocktail) and incubated for 20 min one ice. After centrifugation, the pellets were re-suspended in Lysis buffer and frozen at -80°C. After thawing, the lysis buffer was removed. The pellets were re-suspended in DpnII buffer and dounce homogenized on ice. After centrifugation, the pellets were re-seuspended in DpnII buffer and digested with DpnII enzyme at 37°C overnight. An undigested control was performed in parallel. Ligation reaction was performed by adding DNA ligase and ligation buffer to the reaction followed by an overnight incubation. An aliquot of unligated sample was collected before the ligation reaction as unligated control. Reverse-cross-linking and proteinase K treatment were carried out at 65°C overnight and DNA was extracted using phenol/chloroform extraction method.

Addition of Illumina Sequencing Adaptors: 6µg of 3C library was sonicated to approximately 200 bp. Ampure XP SPRI bead clean up was carried out using 1:9 ratio of material to beads and libraries were generated using the NEBNext DNA library Prep Reagent Set and NEBNext Multiplex Oligos for Illumina Primer (New England Biolabs) following the manufacturer's instructions.

Capture: Capture probe oligonucleotides (available upon request) were reconstituted to 2.9 nM total concentration. 2 µg of differentially indexed samples were pooled to a 1:1 ratio. 2 µg of the pooled library was mixed with 5 µg of COT DNA, 1 nmol of TS-HE Universal Oligo and 1 nmol of TS-HE Index Oligos (Roche). The sample was completely dried by vacuum centrifuge at 50°C. Subsequently, 7.5 µL 2x Hybridisation Buffer and 3 µL Hybridisation Component A (Roche) were added per library in the pooled sample. After 10 minutes denaturation at 95°C, the sample was added to Biotinylated oligonucleotide capture probes and incubated at 47°C for 72 hours. Then, hybridisation reactions were mixed with streptavidin beads and incubated for 45 min at 47°C. Streptavidin beads and bound DNA were washed with Nimblegen SeqCap EZ Hybridisation and wash kit (Roche). Bound DNA was then amplified using KAPA HiFi HotStart ReadyMix. Amplified DNA was then clean up using Ampure XP SPRI beads. A second capture was performed using 2µg of DNA from the first capture but using an incubation time of 24 hours at 47°C instead of 72 hours. After PCR amplification and DNA clean up, the Capture-C libraries were quantified by KAPA Library Quantification Complete Kit and sequenced.

2.19. Library preparation

Unless specified otherwise, library preparation was performed using NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) and the NEBNext Multiplex Oligos for Illumina (NEB) manufacturer's protocol.

2.20. Library quantification and High throughput sequencing

The sample size distribution was assessed using D1000 ScreenTape assays (Agilent) on the Agilent 2200 TapeStation system. Samples were quantified using the Qubit dsDNA HS Assay and KAPA Library Quantification Kits for Illumina® platforms (KAPA Biosystems). 4nM libraries with compatible indexes were pooled and sequencings were run on NextSeq 500 sequencer (Illumina) using either a 300 cycles (Capture-C only), 150 cycles or 75 cycles NextSeq 500/550 High Output v2 kit (Illumina).

2.21. CRISPR-Cas9 genome editing

Genome editing was performed using pX458 plasmid provided by Anthony Cheong. sgRNA was designed to target the ATRX exon 16 (NG_008838.2). The sequence of sgRNA was:

ATRX-sgRNA-Ex16-TOP 5' caccGTCCAATAACAACCAAGT 3'

ATRX-sgRNA-Ex16-BOTTOM 5' aaacACTTGGTTGTTATTGGAC 3'

Nucleofection was carried out on $1.5x10^6$ cells using Lonza nucleofector kit. Cells were incubated for three days before. Then, cells were sorted by single cell FACS by Kevin Clark via selection for GFP-positive, DAPI-negative cells. Alternatively pX459 plasmid provided by Anthony Cheong was used and cells were submitted to puromycin selection instead of FACS sorting.

2.22. Analysis of raw sequencing data from ChIP-seq, ATAC-seq and Bio-CAP-seq experiments

Fastq files were aligned to the human genome hg19 using the in-house pipeline provided by Jelena Telenius using Bowtie 2 (Langmead and Salzberg, 2012) (/home/molhaem2/telenius/Jelena_DNase_pipe_VS_101_090415/DnaseAndChip_pipe_1.sh). In addition, --atac flag was used for ATAC-seq analysis. Data quality was assessed using fastQC reports (Andrews, 2010). Trimming of the adapter sequences was performed using trim-galore (Krueger, 2015). PCR duplicates were removed using Samtools (Li et al., 2009). Signal artefact blacklisted regions (Kundaje, 2013) were excluded. For normalisation, the total number of mapped reads or total number of mapped reads under peak region (based on Diffbind normalisation principle) was determined after the removal of PCR duplicates and excluding the reads mapping to the mitochondrial chromosome. Reads were normalised per 100 million reads. Data were visualised on the UCSC genome browser (http://genome.ucsc.edu/) (Kent et al., 2002).

2.23. Capture-C probes designes and data analysis

Capture-C probes were designed using CapSequm (http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi). Capture-C data were analysed using an in-house pipeline provided by Jelena Telenius, James Davis and Marieke Oudelaar with mapping performed on the human genome hg19.

2.24. Peak calling

Peak calling was performed using MACS2 (Zhang et al., 2008) and HOMER (Heinz et al., 2010) using input (for ChIP-seq and Bio-CAP-seq) or ATAC-seq performed on gDNA (ATAC-seq) as background control. Each sample was independently peak

called. Peak calling quality was assessed by direct visualisation on MIG (McGowan et al., 2013) and peaks associated with artefacts were removed. An additional stringency threshold was applied by excluding any peaks associated with regions to which the related input signal was above the set threshold. A further filtering was based on the representation of each peak across replicates. Peaks were retained for the final peak call only if: (1) they were present in at least three replicates and independent of the antibody used for ATRX ChIP-seq or (2) they were present in at least two replicates for the other experiments.

2.25. Gene ontology and motif analyses

Gene ontology analysis for the differentially expressed genes has been performed using GOTERM_BP_DIRECT DAVID 6.8 (Da Wei Huang et al., 2008).

Gene and genome ontology analyses based on the ATRX binding sites were carried out using HOMER annotatePeaks.pl (Heinz et al., 2010).

Motif analysis of the ATRX binding sites was performed using HOMER findMotifsGenome.pl (Heinz et al., 2010).

2.26. Differential enrichment analysis

Differential enrichment analysis between ATAC-seq datasets was performed with DiffBind (Stark and Brown, 2011) using data experiments carried out in FF-Ctr, CB-Ctr, TA-Ctr and BT-Ctr as controls, and MF-D2035V, SB-R246C, RA-S576X and MatF-D2035V as cases.

2.27. Graphical analysis

Heat maps and profile plots analyses were performed using Deeptools (Ramírez et al., 2014). PCA analyses were performed using DiffBind (Stark and Brown, 2011). Volcano plots were plotted on R.

2.28. Repeatome Analysis

Repeat enrichment in ATRX ChIP-seq data (read one of the paired end only) was performed using a repeat analysis pipeline based on (Day et al., 2010) and adapted by Steve Taylor, and mapped using the hg18 assembly to compare a union of the fastq reads of ChIP vs Input.

2.29. Protein sequence analysis

Amino acid sequences were retrieved from the UCSC browser. Each fasta sequence was submitted to InterProScan (Finn et al., 2016) for domain determination. For the ZNF analysis the number of C2H2 repeat was subsequently manually assessed. The HMM profiles of the KRAB domain (PF01352) and ZF-C2H2 (PF00096) were retrieved from Pfam (Finn et al., 2013)

2.30. RefSeq and CpG island annotation

RefSeq annotation for hg19 was used to characterise the promoter, gene body and intergenic regions identified at ATRX binding sites (O'Leary et al., 2015). The CpG island annotation for hg19 was retrieved from the cpgIslandExt table in UCSC database.

2.31. GenoSTAN analysis

GenoSTAN analysis was performed with the help of R scripts done by Ron Schwessinger based on (Zacher et al., 2017). For the annotation in LCL, ATAC-seq and H3K4me3, H3K4me1, H3K27ac, H3K27me3 and RNA polII ChIP-seq datasets were used. Additionally, CTCF ChIP-seq (GEO accession number GSM733752) was used in the analysis. 20 states were initially used and analogue states were subsequently pooled together resulting in seven distinct states: Active promoter, active promoter containing CTCF, active enhancer, CTCF sites, poised/weak enhancer, repressed region and background. Ambiguous subsets of state annotation that were unclearly differentiated between promoter and enhancer was further correlated using Refseq annotation. Sites overlapping with TSS were associated with promoter and the others with enhancers.

2.32. Transcription factor binding site analysis

TFBS analysis was performed using the Uniform TFBS (wgEncodeAwgTfbsUniform) dataset associated with LCL (Consortium, 2012).

2.33. Visualisation of gene expression in LCL

UCSC visualisation of the gene expression in LCL was carried out using Transcription of GM12878 cells from ENCODE (wgEncodeRegTxnCaltechRnaSeqGm12878R2x75Il200SigPooled) (Consortium, 2012).

Chapter 3. ATRX ChIP-seqs and peak calling

3.1. Introduction

The development of ChIP-seq has revolutionised analysis of the genomic distribution of chromatin interacting proteins (Furey, 2012). This technique allows us to map, *in vivo,* proteins that interact with DNA, either directly or indirectly, by covalently cross-linking proteins and DNA in their native chromatin environment using agents such as formaldehyde. After shearing by sonication or nuclease digestion, cross-linked chromatin is subjected to immunoprecipitation using antibodies specific to the proteins of interest (Park, 2009). This process allows the specific enrichment of the DNA fragments associated with the pulled-down protein. The sample is then reverse-cross-linked, the proteins are digested, and the target DNA fragments are purified. Subsequently, the DNA enrichment can be assessed at specific loci by quantitative polymerase chain reaction (so called ChIP-qPCR) (Collas and Dahl, 2008) or labelled and hybridized to a microarray, so called ChIP-CHIP (Lee et al., 2006). With the development of next generation sequencing, it is now possible to look at such DNA enrichment genome-wide by ChIP-seq (Furey, 2012). The falling cost of sequencing has paved the way for establishing comprehensive analyses of the the epigenetic marks in different cell types as illustrated by the ENCODE project (Consortium, 2012).

Despite well-established protocols producing robust ChIP-seq data for many proteins, ATRX has presented particular challenges. Success of the ChIP-seq protocol is highly dependent on the specificity and avidity of the available antibodies to the target proteins. However, even using optimised antibodies, some proteins are more difficult to study due to their highly dynamic, transient or indirect associations with DNA (Zeng et al., 2006). In these cases, standard cross-linking using a single short spacer arm cross-

linking agent such as formaldehyde may not be sufficient. Even though formaldehyde cross-links not only protein-protein but also DNA-protein interactions, its range is limited to 2 Å. By contrast ethylene glycol-bis(succinimidyl succinate) (EGS) has a spacer arm of approximately 16 Å, thus facilitating the cross-linking of proteins that do not interact directly with DNA (Zeng et al., 2006). Therefore, combining these two cross-linking agents in a two-step method has been an attractive option to overcome the inherited limitations of standard ChIP-seq protocols (Zeng et al., 2006). However, using stronger fixation methods comes at a price; this may significantly affect the quality of the ChIP-seq data and their interpretation. Notably, the fixed chromatin becomes more resitant to sonication. More extensive sonication is therefore needed and this may generate artefacts such as reduce ChIP efficiency and the generation of noisy background (Meyer and Liu, 2014).

ATRX ChIP-seq presents an extra level of complexity. It requires a large number of cells as primary material but the amount of enriched DNA collected after ChIP is limited and contains a high proportion of repetitive sequences. These factors jeopardise the quality of the subsequent library due to potential increases in PCR amplification bias and the mapping efficiency of the sequenced data.

In this chapter, I describe the optimisation of the existing ATRX ChIP-seq protocols, using LCL, to generate high quality and reproducible datasets that provide an important starting point for this study. LCL are immortalised cells, derived from peripheral blood by infecting the resting B cells with Epstein Barr Virus (Sie et al., 2009) and they can provide a constant supply of patient-derived material. Therefore, these cells have been extensively used for genetic and epigenetic studies including the ENCODE project, generating extensive, publically available data (Consortium, 2012). Furthermore, the laboratory has a collection of LCL derived from ATR-X patients constituting an

invaluable resource for analysing the effects of pathogenic ATRX mutations. Consequently, this system provides not only a platform to optimise the ChIP-seq protocol and analyse the basic role of ATRX but also to study the effects of pathogenic mutations, and to compare such data with publicly available epigenetic analyses. The work presented in this chapter addressed the issue associated with the poor quality of the ChIP-seq data that has hampered the identification of ATRX targets and provides a high resolution map of the interactions of ATRX with chromatin.

3.2. Results

3.2.1. Choice of the system and validation of the cell lines

Despite all their advantages, LCL are cell lines and even though care was taken to use early passages and limit their time in culture, these cells are known to acquire genetic and epigenetic alterations (Mohyuddin et al., 2004). To address these limitations and to reduce the background effects, I used three case-control pairs. Each pair is composed of LCL derived from an ATR-X patient (case) and from the unaffected father (control) (***Figure 3.1***). As ATR-X syndrome is X-linked, unaffected fathers have a single wildtype ATRX allele. I thus used case-control pairs composed of relatives to limit the variations in the genomic background and so help to assess the general effects of pathogenic ATRX mutations in ATR-X syndrome. In addition, each group (controls and cases) was composed of three independent LCL clones to avoid clonal effects. The LCL from the control group were used to characterise ATRX targets in a genome-wide manner.

We selected case LCL with representative mutations including variants in the ADD and ATPase domains, as they contain most of the ATRX mutations identified in ATR-X patients (Gibbons et al., 2008). The first case, SB-R246C, harbours a c.736C>T,

p.R246C mutation. This is the most common mutation identified in ATR-X syndrome and it lies in the ADD domain (Gibbons et al., 2008). Its related control LCL is CB-Ctr. The second case, MF-D2035V, has a c.6104A>T, p.D2035V mutation in the ATPase domain (Gibbons et al., 2008). ATRX ChIP-seq data using LCL derived from the unaffected father, FF-Ctr, were previously generated by a former member of the laboratory and were used as a reference for the ChIP-seq optimisation. The last case, RA-S576X, has a c.1727C>A, p.S576X mutation. This mutation creates a stop codon producing a shortened form of the ATRX-truncated (ATRXt) isoform as well as reduced levels of the full-length ATRX which is associated with an alternative spliceoform which leads to "skipping" of the mutation (Gibbons et al., 2008). Its related control LCL is TA-Ctr.

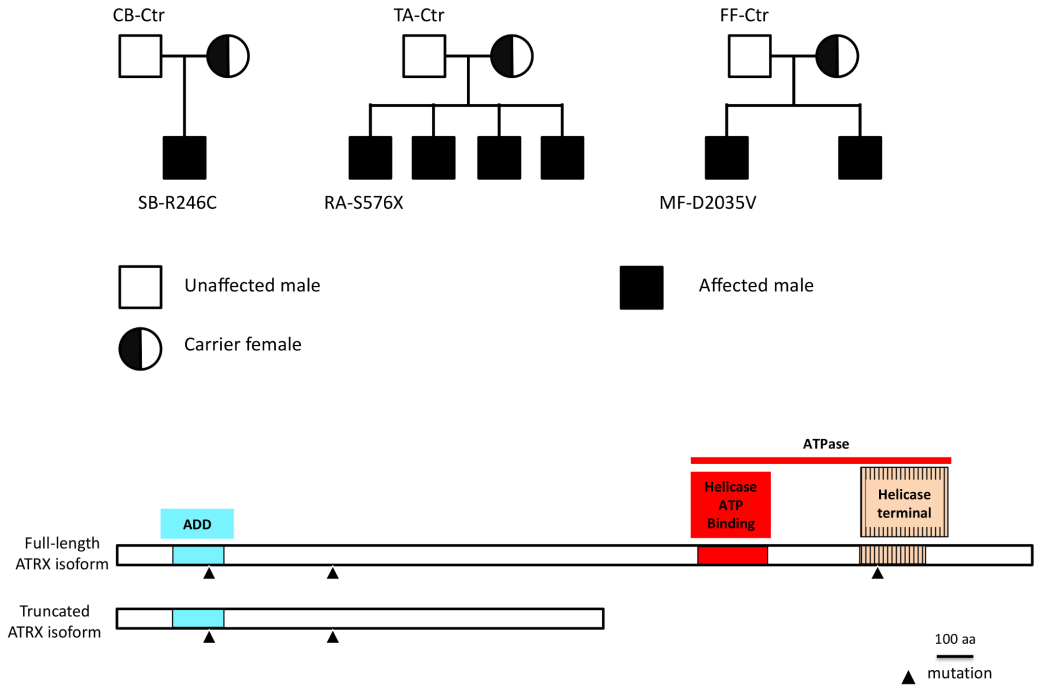


Figure 3.1: Family tree illustrating the case-control pairs and localisation of the associated ATRX mutations.

The six LCL of the three control/case pairs (CB-Ctr/SB-R246C, FF-Ctr/MF-D2035V, TA-Ctr/RA-S576X) were tested to ensure they were mycoplasma negative and I confirmed the ATRX genotype in each of the cell lines using Sanger sequencing (***Figure 3.2***). Then, the samples were analysed at the protein level by western blot using an anti-ATRX antibody (39f) targeting the N-terminal of the protein. This antibody allows the visualisation of both the truncated (~200 kDa) and the full-length ATRX (~280 kDa) isoforms (***Figure 3.3***). Controls display a relatively stronger signal for both the full-length and ATRX-truncated isoforms than cases, except for MF-D2035V which shows a decreased signal of the ATRX full-length but an ATRXt level similar to the control (***Figure 3.3***). As expected the ATRXt in MF-D2035V appears unaffected as the mutation lies in the ATPase domain and is not included in ATRXt. By contrast, the ATRX mutations present in SB-R246C and RA-S576X are also in the sequence of the truncated isoform and their signals show a knock down in expression of both the truncated and full-length proteins. A longer exposure highlighted the presence of an additional truncated ATRX isoform at ~75KDa (marked as ATRX-S576X truncated in ***Figure 3.3***) in RA-S576X. This signal represents the mutant ATRXt isoform produced in the absence of alternative splicing.

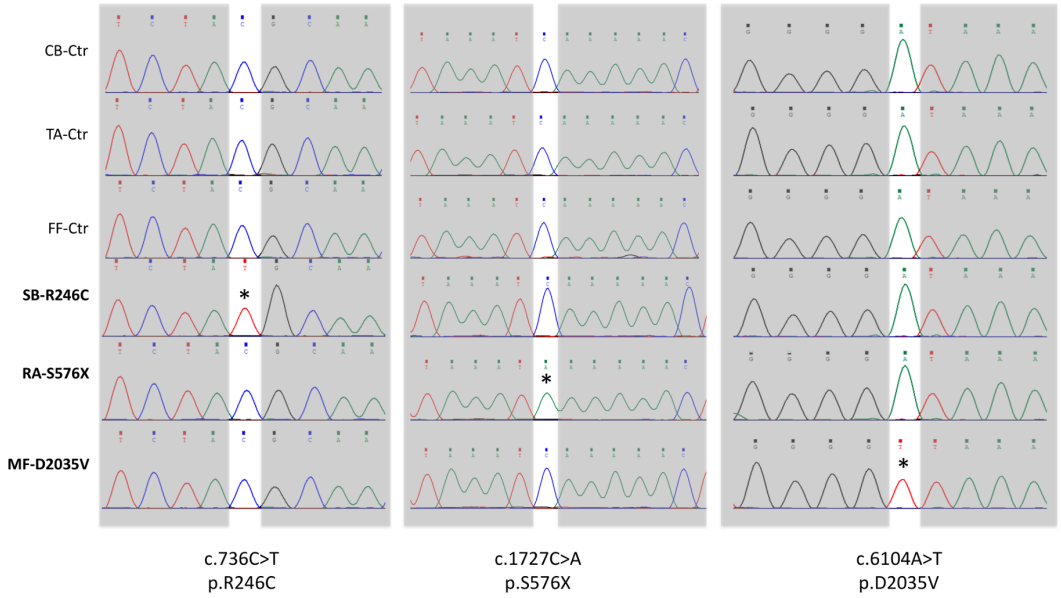


Figure 3.2:Validation of the LCL by Sanger sequencing.

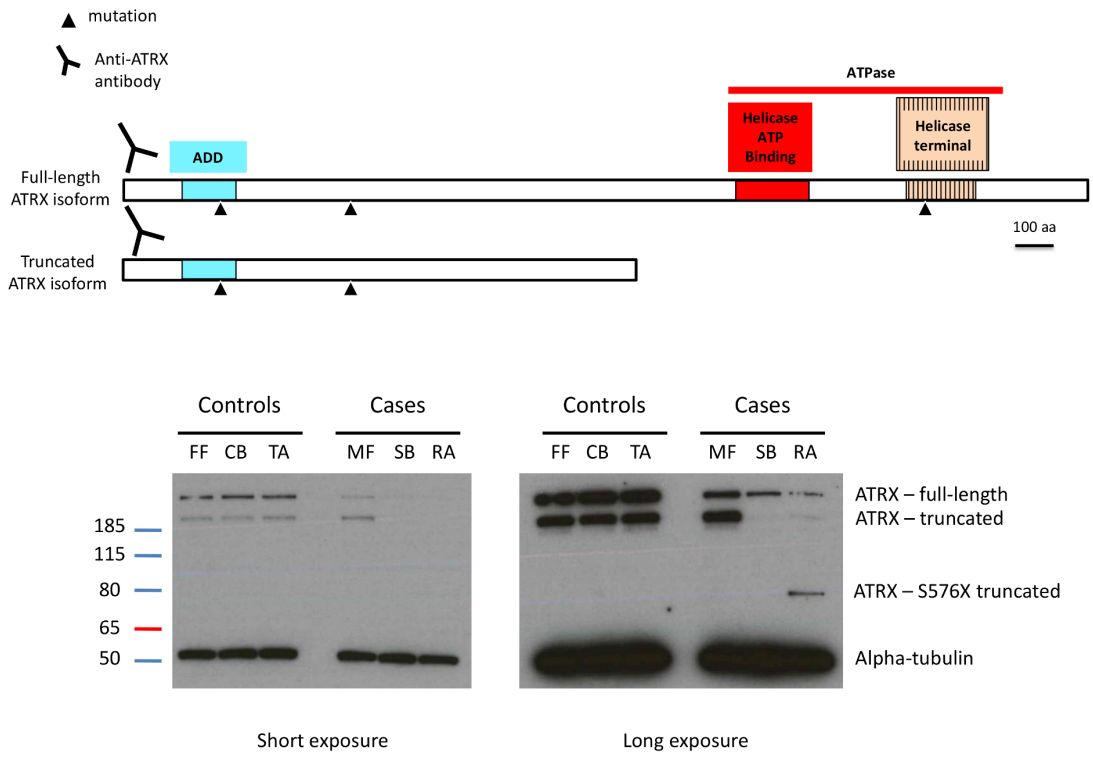


Figure 3.3: Characterisation of the LCL used for ChIP-seq using western blot. The three control LCL are labelled FF, CB and TA for FF-Ctr, CB-Ctr and TA-Ctr, respectively. The three related case LCL are labelled MF, SB and RA for MF-D2035V, SB-R246C and RA-S576X, respectively. 39f anti-ATRX antibody was used to highlight both the truncated and full-length ATRX isoforms. Alpha-tubulin is used as loading control.

3.2.2. Choice and validation of the anti-ATRX antibodies

Then, I used the anti-ATRX antibody H300 (sc-15408, Santa Cruz) to specifically analyse the full-length ATRX isoform. This antibody targets the C-terminal amino acids 2193 to 2492 of the human ATRX protein (https://www.scbt.com/scbt/product/atrx-antibody-h-300), and so, does not detect the ATRXt isoform. It has been validated and successfully used in previous studies (Voon et al., 2015, Law et al., 2010). In addition, I used a second antibody, the Abcam ab97508, sharing similar characteristics to further strengthen the data. Both antibodies were tested by western blot using protein extracts from the three healthy individuals and three ATR-X patients diplaying the different ATRX mutations described above. A strong signal was observed in the healthy individuals at the expected size of the full-length ATRX (280 kDa) isoform with both antibodies. With respect to the tubulin control, fainter ATRX signals were observed in the ATR-X patients (***Figure 3.4***).

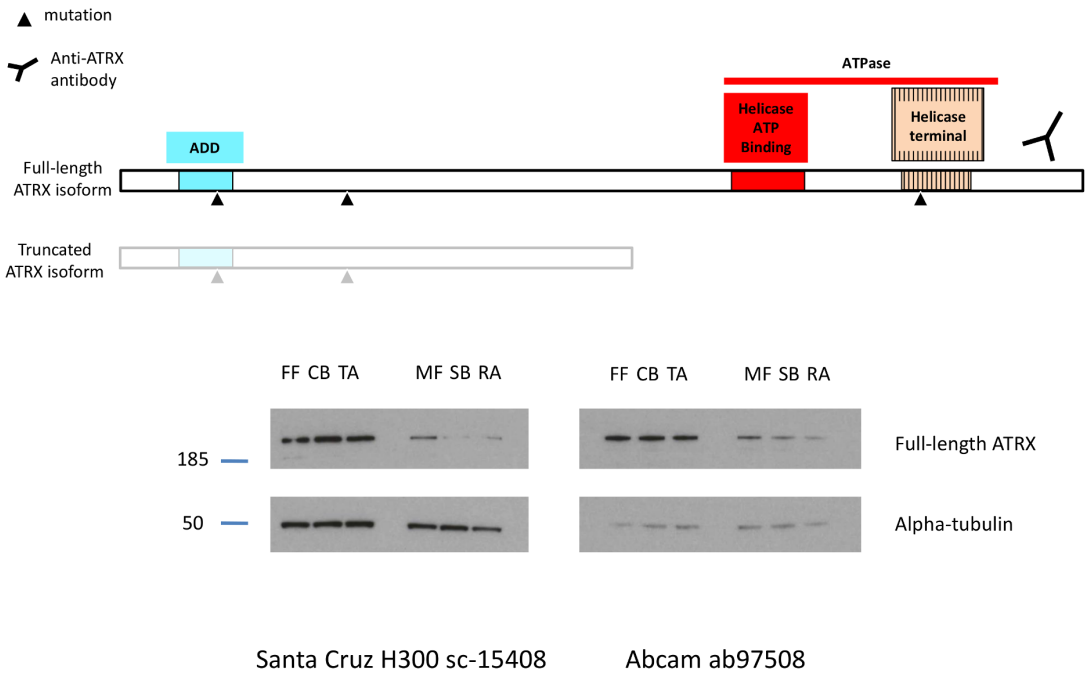


Figure 3.4: Validation of the anti-ATRX antibodies which specifically detect the full-length ATRX isoform.

3.2.3. Optimisation of the ATRX ChIP-seq protocol

Fragmentation of the cross-linked chromatin is required in ChIP-seq as it allows subsequent isolation of DNA fragments interacting with the protein of interest. If the chromatin is insufficiently sheared (or under-sonicated), the DNA fragments immunoprecipitated with the protein of interest may be significantly larger than the actual region of protein-DNA interaction resulting in a decrease in the resolution of the ChIP-seq data and information loss. As a double-edged sword, excessive sonication (or over-sonication), may also cause a drop of efficiency and create artefacts in ChIP-seq. Thus, leading to 'noisy' background and increase in false positives which confound the data interpretation.

To optimise the sonication parameters, I used the Bioruptor and the protocol described by (Law et al., 2010) as a starting point. After lysis of the double cross-linked LCL, the sample was evenly split into two tubes (A and B). An aliquot of each tube was analysed after different cycles of sonication. Despite appropriate sample homogenisation, I observed a batch effect between tubes A and B (***Figure 3.5***.a). After 30 cycles, a significant proportion of under-sonicated chromatin remained in tube A with fragments larger than 1,000 bp (***Figure 3.5***.a). Similarly, tube B contained under-sonicated chromatin after 28 cycles. After 35 cycles, the size of most fragmented chromatin was estimated between 100 and 300 bp. Beyond 35 cycles the chromatin fragmentation generated a similar pattern. Nevertheless, further sonication can also lead to the production of over-sonicated fragments. A protocol using 35 cycles was selected as a starting point for correct sonication of the two-step cross-linked chromatin in LCL. Yet, subsequent experiments carried out with LCL samples sonicated for 35 cycles using the Bioruptor illustrated the poor reproducibility of the sonication (***Figure 3.5***.b). Indeed,

samples from experiments 1 and 3 were under-sonicated with a large proportion of the chromatin fragments remaining larger than 1kb. The sonication in experiment 2 failed. Experiments 4 and 5 resulted in an enrichment of chromatin fragments between 100 and 300 bp.

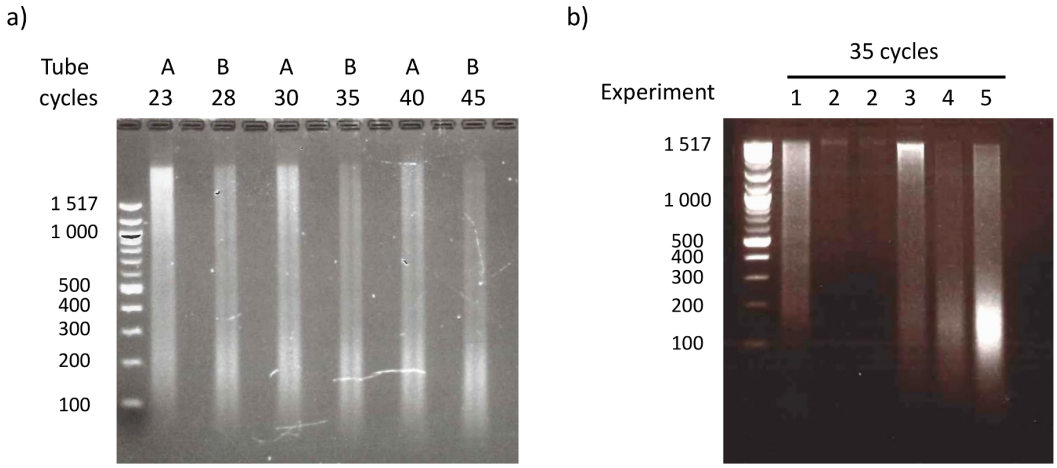


Figure 3.5: Sonication of double cross-linked LCL samples using the Bioruptor. a. Optimisation of the sonication parameters using a LCL sample, split in two 2 tubes, A and B. Aliquots were collected alternatively from tube A and B. 3. b. Assessment of the reproducibility of the sonication with optimised parameters across several ATRX ChIP experiments using LCL.

Due to the inconsistent results observed using the Bioruptor, I changed to the Covaris system. After optimisation, a duty of 12.5% and intensity of 5 for 25 min resulted in the majority of fragments being smaller than 300 bp (***Figure 3.6***.a). In subsequent experiments, however, a significant proportion of the double cross-linked chromatin became resistant to sonication (***Figure 3.6***.b) and increasing the stringency of the sonication settings did not solve the problem.

A comparison between the original optimisation experiment using the Covaris (***Figure 3.6***.a) with the subsequent experiments (***Figure 3.6***.b) led to the hypothesis that the time in storage at - 80°C (between the cross-linking step and the sonication step) is an

important factor affecting the sonication efficiency. Indeed, the samples used for the optimisation process were stored for a shorter period than the samples used in subsequent experiments. The impact of storage period on the sonication efficiency was further supported by the literature (Bortz and Wamhoff, 2011). Using the optimised settings previously defined on double cross-linked chromatin together with a shorter storage period of less than 72h at -80°C, resulted in a satisfactory and reproducible fragmentation (***Figure 3.6***.c). Yet, reducing the stringency of the sonication procedure, to limit eventual over-sonication, did not produce a similar sonication quality than with the previously optimised setting (***Figure 3.6***.c). Finally, I chose to use the Covaris with a duty of 12.5% and intensity 5 for 25 min on chromatin stored for less than 72h after double cross-linking as the new standard procedure for the ATRX ChIP experiments.

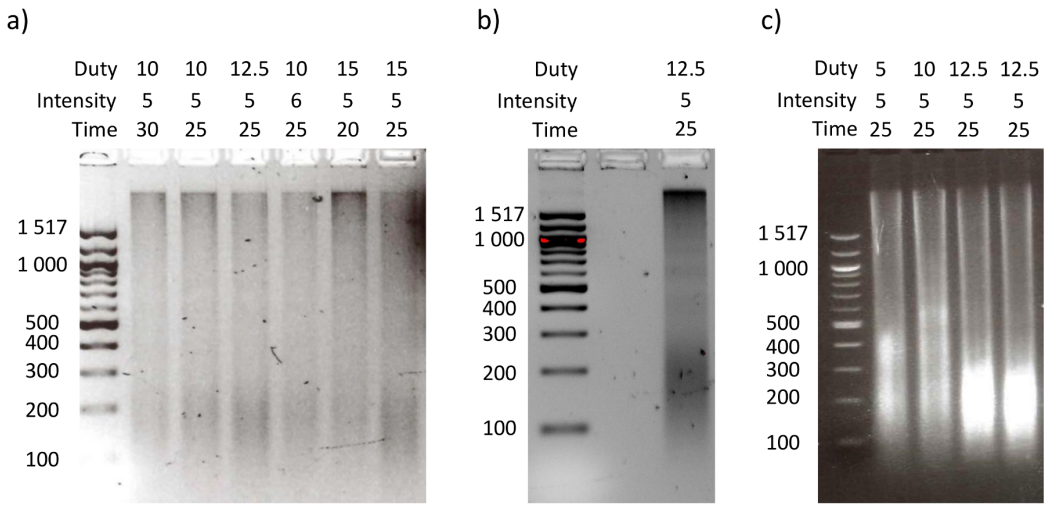


Figure 3.6: Sonication of double cross-linked LCL samples using the Covaris. a. Optimisation of the sonication parameters using double cross-linked LCL. b. Example of sonication artefact observed on double cross-linked LCL samples after more than a week of storage at -80°C. c. Optimisation and assessment of the reproducibility using double cross-linked LCL stored for less than 72h at -80°C before sonication

Further major modifications of the original ATRX ChIP-seq protocol included doubling the number of cells used as starting material per ChIP reaction, performing in-house library preparation and using a 150 cycles sequencing kit rather than a lower cycle kit to generate longer reads. As ATRX is known to associate with regions of repetitive DNA sequences, increasing the length of the sequenced fragments should increase the probability of obtaining reads containing unique sequences that flank repeats. These unique sequences improve the mapping accuracy. Using an even higher cycle sequencing kit, such as 300 cycles, did not appear cost efficient based on the average size of the sonicated chromatin (around 150 bp) (***Figure 3.6***.c).

3.2.4. Generation of robust ATRX ChIP-seq datasets in LCL

Once optimised, the ATRX ChIP-seq protocol was used to produce the core data for the present study (Truch et al., 2018). To this end, ATRX ChIP-seq experiments were performed in three independent LCL derived from healthy donors: FF-Ctr, TA-Ctr and CB-Ctr using two different anti-ATRX antibodies, H300 and ab97508, both targeting the C-terminus of the full-length protein. Each trio of ATRX ChIP-seq experiments performed with the same anti-ATRX antibody were analysed separately as biological replicates and also as a pool. Visualisation of the optimised ATRX ChIP-seq data using the UCSC genome browser (Kent et al., 2002) revealed reproducible, high quality data with low to very low background signal in the input (***Figure 3.7***); a significant improvement over previous datasets. Indeed, the ATRX ChIP-seq datasets generated with the newly optimised protocol did not display a "Furry" input, representing sites of putative preferential sonication, either at open or closed chromatin regions (***Figure 3.8***). Indeed, the high background signal observed in previous ATRX ChIP-seq datasets was not always associated with open chromatin regions suggesting more complex

underlying causes than simple over-sonication. One possible explanation could have been the presence of copy number variants (CNVs) within these regions. To investigate this, the same region was examined in the new ATRX ChIP-seq datasets derived from the same LCL (FF-Ctr) used in the original work. Using the new protocol, no "Furry" peaks were detected in the ATRX ChIP-seq or related input using either antibody, which showed that these peaks were not due to CNVs and confirmed the reliability of these new datasets (***Figure 3.8***.b).

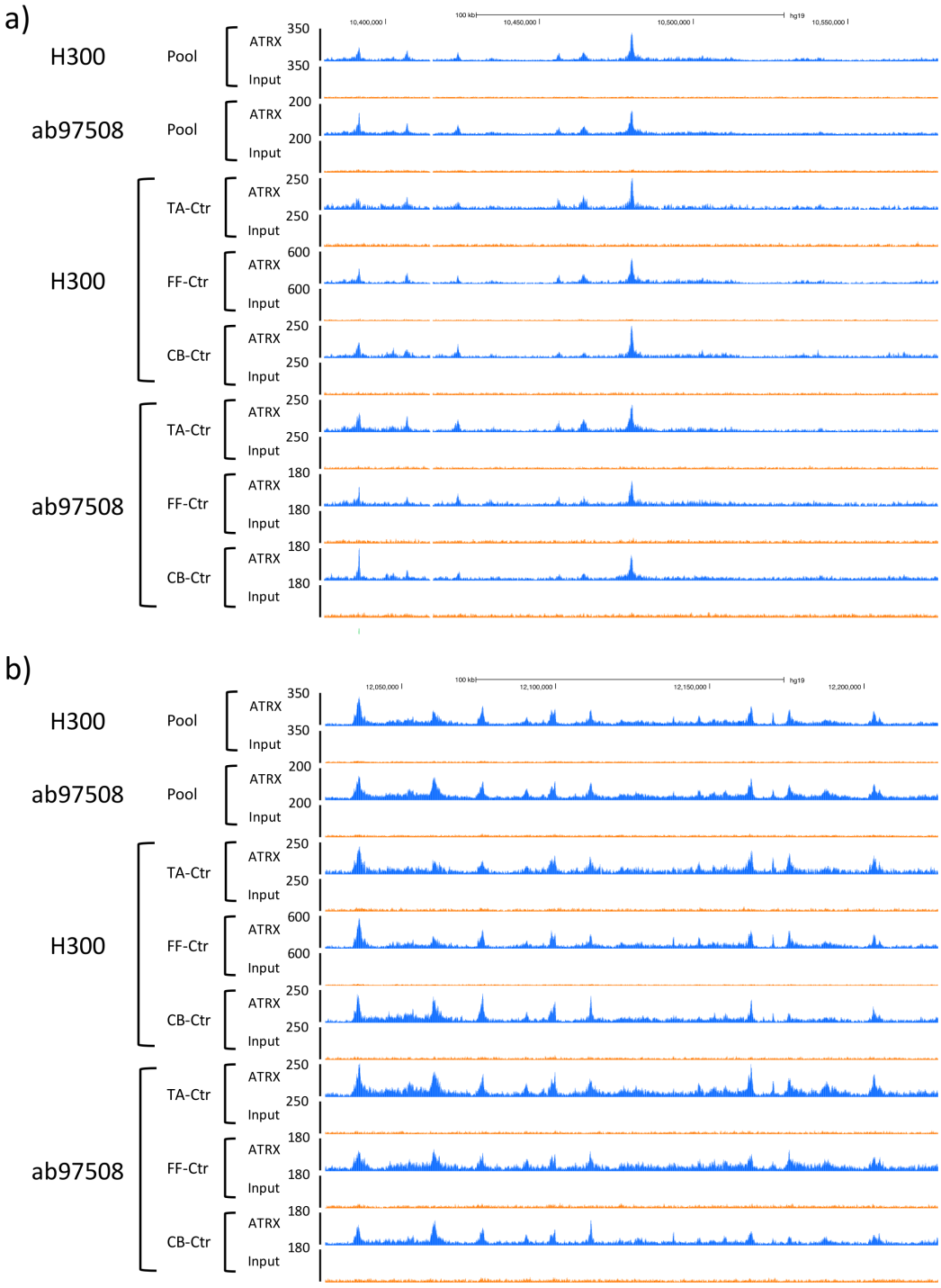


Figure 3.7: Screenshot of the hg19 UCSC genome browser showing the ATRX ChIP-seq experiments performed in LCL. a. 200kb region showing ATRX enrichment. b. 200kb region highly enriched in ATRX binding sites. Pool: pool of the three ATRX ChIP-seq experiments done on FF-Ctr, TA-Ctr and CB-Ctr LCL using the same anti-ATRX antibody. H300: anti-ATRX antibody from Santa Cruz. Ab97508: anti-ATRX antibody from Abcam.

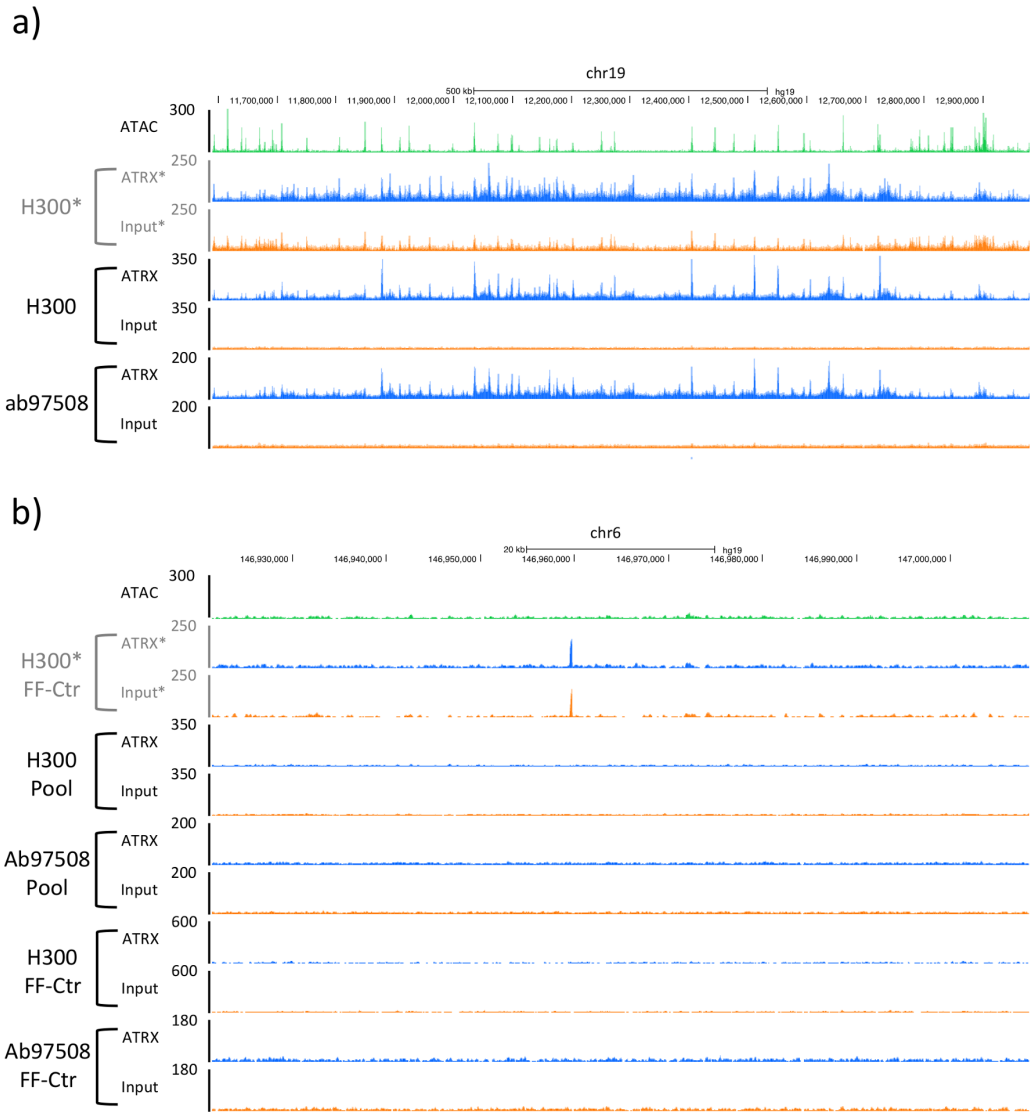


*Figure 3.8: Screenshot of the hg19 UCSC genome browser showing the improvements of the ATRX ChIP-seq experiments performed in LCL. a. Reduction of the 'Furry' background at open chromatin regions. b. Reduction of the background at closed chromatin region. Pool: pool of the three ATRX ChIP-seq experiments done on FF-Ctr, TA-Ctr and CB-Ctr LCL using the same anti-ATRX antibody. * Data generated by a former member of the group before the optimisation of protocol described here.*

PCA analysis showed good clustering of the ATRX ChIP-seq experiments. The sub-clusters showed a batch effect associated with the antibody used in both ATRX ChIP and input samples (***Figure 3.9***). The input samples, even though more tightly clustered, do not show a sub-clustering based on the source of the LCL and the two input samples derived from TA-Ctr LCL were at the opposite extremes of the input clusters. Conversely, the ATRX ChIP samples displayed some secondary sub-clustering based on the source of the LCL. By pooling the biological triplicates, the disparity is reduced between the H300 and ab97508 pools (***Figure 3.9***).

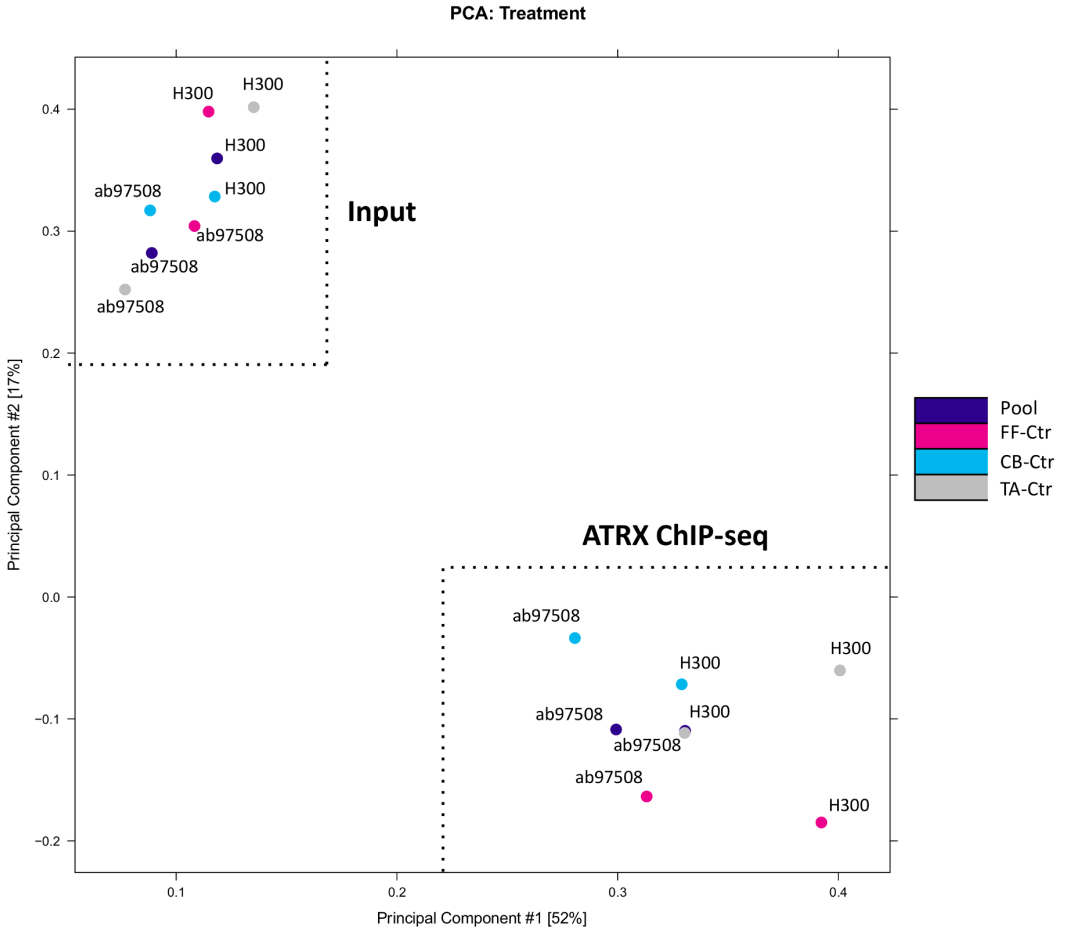


Figure 3.9: PCA analysis of the ATRX ChIP-seq experiments. Pool: pool of the three ATRX ChIP-seq experiments done on FF-Ctr, TA-Ctr and CB-Ctr LCL using the same anti-ATRX antibody. H300: anti-ATRX antibody from Santa Cruz. Ab97508: anti-ATRX antibody from Abcam.

ATRX ChIP-seq datasets were then further validated by ChIP-qPCR at known positive and negative controls, rDNA and DIST, respectively, as well as at newly identified targets (***Figure 3.10***).

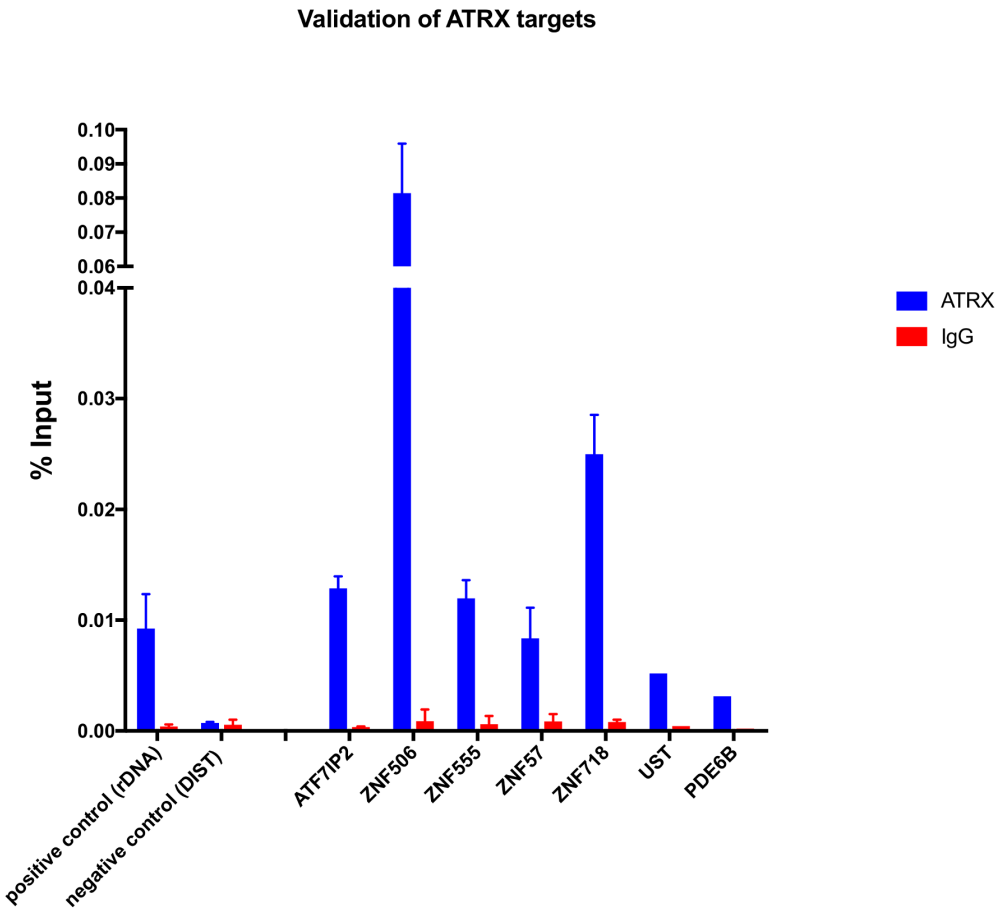


Figure 3.10: Validation of ATRX targets by ChIP-qPCR.

3.2.5. Genome-wide ATRX distribution; a challenge for peak calling

The generation of a robust collection of ATRX ChIP-seq experiments resulted in very large and complex datasets to interpret. Therefore, to analyse the genome-wide distribution of ATRX, I used a peak calling approach to identify the genomic regions enriched in ATRX. Preliminary inspection of the datasets highlighted the complexity of the ATRX enrichment with a diversity of peak shapes and densities across the

genome (***Figure 3.11***. a and b). ATRX binding patterns included sharp, well-defined peaks (***Figure 3.11***.b) as usually observed in transcription factors or ATAC-seq datasets, and broad peaks similar to those seen with histone marks such as H3K4me1 or H3K9me3 (***Figure 3.11***.a). In addition to this heterogeneity, ATRX ChIP-seq contained some extreme changes in background enrichment within the ChIP-seq data themselves dependent on the genomic region (***Figure 3.11***.a). These characteristics, peculiar to ATRX, created many challenges in generating accurate peak calls.

I tested two peak caller packages, MACS (Zhang et al., 2008) and HOMER (Heinz et al., 2010) and observed significant over-calling of peaks using the default parameters (***Figure 3.11***.c). Eventually, I opted for HOMER as the peak caller for this study as it offered a wide range of parameters for optimisation.

By increasing the stringency of the parameters used, the specificity of the peak calling was improved (***Figure 3.11***.c). Over-stringent parameters reduced the overall number of putative peaks identified but improved the resolution by unmerging peak regions into individual peaks. However, it did not completely solve the problems associated with over-called peaks as some of them were still retained as highlighted by red arrows in ***Figure 3.11***.c. When using a dataset with a slightly higher background signal, setting the parameters too stringently resulted in under-calling of some enriched regions (as shown by red arrows in ***Figure 3.11***. d). No matter which peak caller was used, or how the parameters were set, each peak call would retain some of these artefacts. After individual peak calling of the six ATRX ChIP-seq datasets (from FF-Ctr, CB-Ctr and TA-Ctr using either H300 or ab97508 anti-ATRX antibody), the peak calls were merged and peaks were discarded if they were present in fewer than three datasets. This approach reduced the presence of over-called peaks and took into account potential under-calling, individual background (presence or absence of a peak in a dataset

specific to the LCL used) as well as the background differences across the ATRX ChIP-seq experiments. In addition, the antibody-specific peaks i.e. those only present in datasets using one antibody but not the other, were removed.

I assessed each peak calling by direct visualisation using MIG (McGowan et al., 2013) and identified that few very large peaks (more than 10kb long) were present in the final peak call dataset. Nevertheless, these appeared to be real ATRX-enriched regions, associated with very low input background, as illustrated at the *ZNF506* gene locus (***Figure 3.11***.e), which was validated by ChIP-qPCR (***Figure 3.10***), and were therefore retained. These regions are associated with ZNF genes and will be further discussed in the next chapters. By contrast and despite the use of the input dataset during peak calling, several peaks with significant input background were identified (***Figure 3.11***. f and g). These suspicious enrichments, similar to signal artefact blacklisted regions (Kundaje, 2013), were excluded. Ultimately, in depth analysis identified 8,454 sites of significant ATRX enrichment genome-wide in LCL.

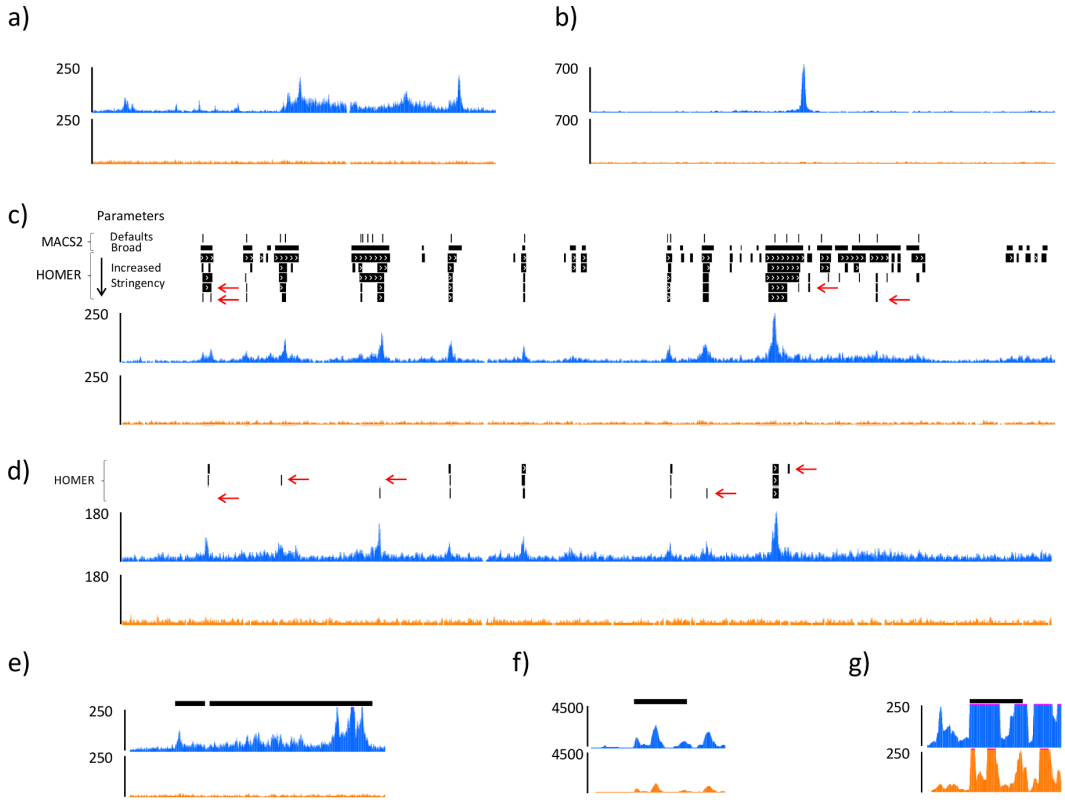


Figure 3.11: Examples of heterogeneity of the ATRX enrichment across the genome and challenges of peak calling in ATRX ChIP-seq datasets. Data were visualised on UCSC genome browser; in each case, blue: ATRX ChIP-seq, orange: input. a and b. Example of the heterogeneity of the ATRX enrichments across the genome with high variability in peak shapes at two regions. a. High variability in ATRX enrichment within a window larger than 100kb. b. Example of a sharp ATRX peak in a region (larger than 50kb). c and d. Examples of different peak calls using two peak callers, MACS2 and HOMER, and independent ATRX ChIP-seq datasets. c. ATRX ChIP-seq performed with H300 anti-ATRX antibody. Increasing stringency of the peak calling parameters of HOMER associated peak calling from top to bottom. d. ATRX ChIP-seq performed using the ab97508 anti-ATRX antibody. e) Example of a region highly enriched in ATRX resulting in a very large peak spanning the ZNF506 gene locus even with optimised parameters (over 40 kb). f and g. Example of potential artificial ATRX enrichment near to the centromere of chromosome 16, using two different scales, due to mapping artefacts not removed by the signal artefact blacklisted regions. Same ChIP-seq dataset used in a, b, e, f and g. Data normalised by read per 100 millions.

3.3. Discussion

In order to understand the normal role of ATRX and comprehend how mutations contribute to human disease, it has been essential to determine its genome-wide distribution. However, producing high quality ATRX ChIP-seq datasets turned out to be more challenging than expected despite the improvements of NGS technologies over the years. With the aim of generating a robust genome-wide ATRX ChIP-seq dataset suitable for the analysis of its binding sites and to assess the effects of pathogenic ATRX mutations on gene expression and the chromatin landscape, I first validated the LCL required for this study. Using Sanger sequencing, I observed the presence of specific base pair substitutions in the ATRX gene in LCL derived from three affected males as previously reported (Gibbons et al., 2008). I detected the full-length and truncated isoform of the protein, using an antibody targeting the N-terminus of ATRX. Consistent with previous observations, SB-R246C and RA-S576X showed a reduced signal for both the full-length (~280 kDa) and truncated (~200 kDa) isoforms compared to the samples derived from their unaffected fathers (Gibbons et al., 2008). MF-D2035V, displays a mutation in the ATPase domain, therefore did not lead to changes in the expression of ATRXt. Finally, an extra signal was observed in RA-S576X sample corresponding to the mutant associated truncated ATRX protein (Picketts et al., 1996). Interestingly, in RA-S576X case, alternative splicing occurs, by-passing the mutation and resulting in a mutant knock down rather than a knock out of the full-length protein (Gibbons et al., 2008). These results agree with published data and show that ATRX mutations destabilise the full-length protein to various extents (Argentaro et al., 2007, Mitson et al., 2011).

For the ChIP-seq experiments, I only used antibodies which recognise the full-length ATRX. There is no major phenotypic difference associated with mutations that affect

the full-length protein and those that affect both the full-length and truncated isoforms; this observation emphasises the essential role of full-length ATRX protein for a normal development. I therefore validated two anti-ATRX antibodies which target the C-terminal region of the protein for subsequent use in the ChIP-seq experiment.

I optimised the ATRX ChIP-seq protocol described by (Law et al., 2010) to increase the sensitivity of the assay and reduce the background. Increasing the amount of starting material, limiting the storage period of cross-linked samples and improving the sonication efficiency as well as in-house library prepraration and sequencing are key factors in this process. Using the optimised protocol (Truch et al., 2018), I produced a robust ATRX ChIP-seq dataset with high definition and low input signal in three independent LCL using two anti-ATRX antibodies. Using PCA analysis, I observed a clustering of the ATRX ChIP-seq data based on the antibody and sample used. Pooling the biological replicates based on the antibody used further reduced the variability.

The ATRX ChIP-seq data showed heterogeneity in peak shapes and distribution, thus making peak calling complex. Peak caller packages such as MACS (Zhang et al., 2008) and HOMER (Heinz et al., 2010) were designed to peak call datasets containing transcription factor-like (sharp peaks) or histone-like (broad peaks) distributions. Furthermore, improving the ATRX ChIP-seq protocol resulted in datasets with very low background leading to over-calling peaks when using default or more stringent parameters with two independent peak callers (MACS2 and HOMER). Even if most of the regions over-called represent real enrichment, they would impede the subsequent analyses by increasing the background noise. The fine-tuning of the peak caller parameters and the use of the input datasets were required to achieve a qualitative peak call.

Using stringent parameters, I identified 8,454 peaks across the LCL's genome, which would be the key dataset for the rest of this study. Despite the use of strict criteria for peak calling, the ATRX targets identified in LCL were more than eight times more abundant than in the original study performed in erythroblasts (Law et al., 2010). Careful assessment of the peak calling results using MIG revealed the presence of artefacts, even after filtering for blacklisted regions and using the input as background. These observations emphasised the need for assessing the peak calling results and not just using default parameters without further validation. Despite the care taken in this analysis, it must be acknowledged that a "perfect" qualitative ATRX peak call could not be achieved. In the final peak call, I noticed some very large peaks that might, in an ideal world, be split into sub-peaks giving a more qualitative result; the presence of some over- and under-called regions in the final peak call is expected. With the aim of improving the peak call quality of such datasets in the future, I am collaborating with the Hughes group to develop a new peak caller able to deal with very complex enrichment patterns.

Beyond the limitation of peak calling, the complexity of the ATRX distribution, including both transcription factor-like sharp peaks and broad histone enrichment-like patterns may indicate the presence of several subgroups among the ATRX targets and/or suggest various modes of interactions with DNA sequences. Indeed, a composite of direct and indirect interactions with DNA has to be considered as ATRX has been shown to interact with chromatin through binding with H3K9me3 for instance (Iwase et al., 2011, Eustermann et al., 2011) as well as with DNA secondary structures such as G4 (Law et al., 2010). Additionally, it is possible that ATRX, as a DNA translocase, forms very dynamic, transient interactions with naked DNA or other proteins.

Furthermore, the heterogeneous distribution of ATRX may reflect its different actions or modes of recruitment at various loci.

Despite the challenges to overcome, the generation of the current ChIP-seq datasets provided the opportunity to further investigate ATRX targets, as discussed in subsequent chapters, and to understand more deeply and to determine the potential role of ATRX at these sites. Furthermore, this optimised ATRX ChIP-seq protocol (Truch et al., 2018) provides a useful resource for the laboratory and beyond and may be adapted to successfully perform ChIP-seq experiment with other challenging proteins displaying very dynamic, transient and or indirect interactions with DNA.

Chapter 4. Genome-wide characterisation of the ATRX binding sites

4.1. Introduction

The cellular localisation of a protein may provide valuable hints to unravel its role and/or function. In the case of a chromatin interacting protein, further analyses of the characteristics of its genomic binding sites may help determine its method of recruitment, its direct targets and provide a subset of attractive sites to investigate in further details. Since its discovery, many studies have used such an approach to investigate the role of ATRX. In this context, microscopy has been a valuable tool to analyse ATRX distribution in both mouse and human. Indeed, several studies provided essential information regarding the spatio-temporal distribution of ATRX and showed that the protein mainly localised to heterochromatin including telomeres, pericentric regions and rDNA as well as at PML-NBs (Gibbons et al., 2000, McDowell et al., 1999, Xue et al., 2003, Wong et al., 2010). However, with the advent of the ChIP-seq approach it has been possible to map ATRX distribution at much higher resolution revealing novel genomic sites, paving the way to a deeper understanding of its function. Thus, the first ATRX ChIP-seq carried out in mouse and human were a watershed in the field and revealed new ATRX targets including GC-rich tandem repeats predicted to form G4 (Law et al., 2010). Furthermore, ATRX ChIP-seq performed in human erythroid cells revealed the presence of ATRX binding sites at the disease-relevant alpha globin locus.

These datasets have been reanalysed in several studies and a significant amount work has been undertaken, especially in mouse (Voon et al., 2015, Ratnakumar et al., 2012, Clynes et al., 2013). Surprisingly, by 2018, only two ATRX ChIP-seq experiments were

published in human and very few potential direct targets have been identified in this organism (Law et al., 2010, Valle-García et al., 2016). The production of a new robust and sensitive dataset for ATRX ChIP-seq would give us the opportunity to re-examine the characteristics of ATRX binding sites genome-wide and hopefully identify previously undetected targets and gain new insights into ATRX function. As ATRX is known to interact with highly repetitive sequences that are unfavourable for standard pipelines involving the genomic mapping of Illumina-based short read sequences, I divided the analysis in two steps. Firstly, I performed an enrichment analysis of the repetitive elements (repeatome) based on the method developed by Day and colleagues. This method is designed to analyse enrichment at repetitive sequences independently of their position in the final human genome assembly (Day et al., 2010). To bypass the limitations associated with repetitive sequences in the standard analysis, sequencing reads were directly mapped to the human repeatome without preliminary filtering processes such as removing of the potential PCR duplicates. As a control, the enrichment was compared to the input dataset, which had been similarly treated. Both the ChIP and input samples were normalised for the number of reads mapped to the repeatome. This approach was used for the original analysis of the first ATRX ChIP-seq datasets (Law et al., 2010). Secondly, I investigated the ChIP-seq dataset using the peak call described in the previous chapter. With this stringent peak call, my aim was to optimise the identification and characterisation of ATRX binding sites and their potential subgroups. To do so, I took advantage of publicly available datasets such RefSeq dataset (O'Leary et al., 2015) and supplemented those with further experiments to produce a comprehensive analysis.

4.2. Results

4.2.1. Analysis of ATRX enrichment at repetitive elements using direct mapping of ChIP-seq data

Previous study has shown that ATRX interacts with rRNA, G-rich telomere and subtelomeric repetitive regions (Law et al., 2010). With the help of Steven Taylor, reads from the ATRX ChIP-seq data and their related inputs were directly mapped to human repeatome (Day et al., 2010). Repeats associated with the large and small subunits of rRNA showed the highest maximum likelihood enrichment estimates (mle), mle = 3.021 and mle = 2.721, respectively (***Figure 4.1***.a). Other repetitive sequences enriched in ATRX included a high number of tRNA and simple repeats (***Figure 4.1***.b). ATRX was also enriched at subtelomeric TAR1 repeats (mle = 2.002) (***Figure 4.1***.c). However, no significant ATRX enrichment over input was observed at Telomeric repeats (CCCTAA)n (mle = 0.236) (***Figure 4.1***.c). Other repeats enriched in ATRX included the GC-rich LSAU (mle =2.588), the small nuclear RNA U2 (mle = 1.982), the Long terminal repeat (LTR) MER57E3 (mle = 1.885), the pericentromeric satellite HSATII (mle = 1.384) and the LTR LTR2B (mle = 1.232) repeats (***Figure 4.1***.d).

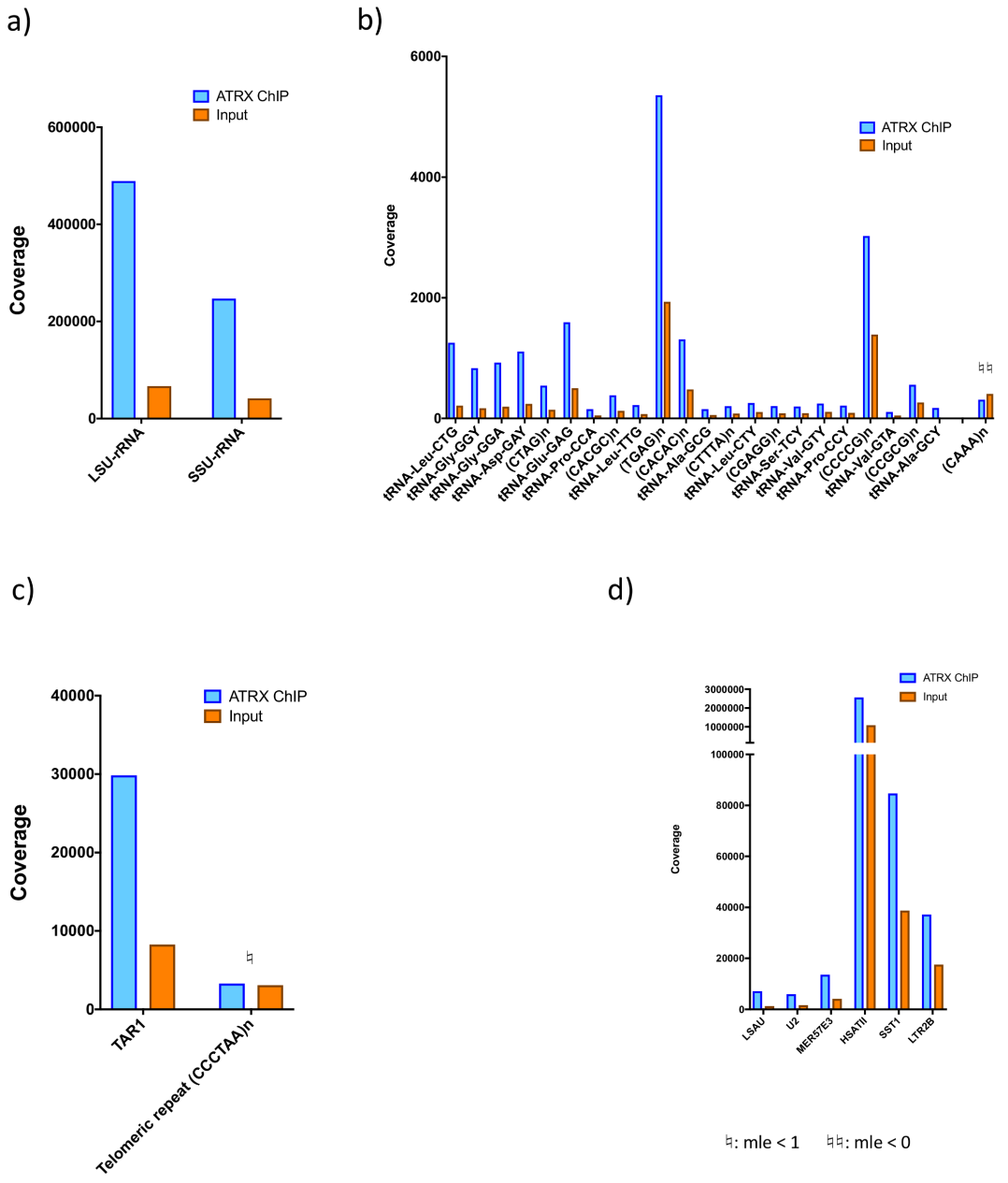


Figure 4.1: Analysis of ATRX binding at repetitive elements using direct mapping of the ATRX ChIP-seqs and related inputs. a. Repetitive regions with the highest mle value. b-d. Repetitive regions enriched in ATRX plots on different graphs due to scale constraints. With the exception of (CAAA)n and Telomeric repeat (CCCTAA)n, the mle > 1.15 and the ATRX fold enrichment over input > 2 for all the targets. b. ATRX Enrichment at tRNA and simple repeats. (CAAA)n example of simple repeats that were not enriched in ATRX (mle = -0.228). c. ATRX enrichment at subtelomeric TAR1 and telomeric (CCCTAA)n repetitive sequences. TAR1: mle = 2.002 and Telomeric repeat (CCCTAA)n: mle = 0.236. d. Additional repetitive sequences enriched in ATRX with mle > 1.15 and the ATRX fold enrichment over input > 2.

4.2.2. Genome Ontology analysis of ATRX binding sites

To characterise the ATRX binding sites identified in Chapter 3, I first performed a gene ontology analysis using HOMER (Heinz et al., 2010). The top terms were all associated with ZNF genes. Other enriched terms were associated with the regulation of various biological processes and molecular functions associated with nucleic acid binding. Additionally, enrichment at repetitive sequences such as GC-rich, low complexity, LTR2B, MER57E3 and a subset of simple repeats were also detected.

p-value	Term	Number of Genes in Term	Number of Target Genes in Term
1.34E-86	KRAB	350	253
1.80E-85	Krueppel-associated_box	366	258
3.18E-84	KRAB	350	254
2.40E-82	Znf_C2H2/integrase_DNA-bd	679	382
2.21E-79	Znf_C2H2-like	800	422
7.18E-76	Znf_C2H2	766	404

Table 4.1: Top terms of genome ontology highlighting a strong association between the ATRX binding sites and ZNF genes.

p-value	Term	Number of Genes in Term	Number of Target Genes in Term
1.05E-64	regulation of cellular metabolic process	5290	1672
6.31E-63	regulation of RNA metabolic process	3356	1162
7.13E-63	regulation of RNA biosynthetic process	3272	1139
2.64E-62	regulation of nucleobase-containing compound metabolic process	4114	1362
3.31E-62	regulation of transcription, DNA-templated	3232	1126

Table 4.2: Top terms of genome ontology associated with biological processes

p-value	Term	Number of Genes in Term	Number of Target Genes in Term
1.45E-53	DNA binding	2346	859
1.40E-31	nucleic acid binding	3789	1167
1.54E-22	binding	12394	3133

Table 4.3: Top terms of genome ontology associated with molecular function

Name	Overlap (number of peaks)	p-value
GC_rich|Low_complexity|Low_complexity	810	2,51E-28
LTR2B|LTR|ERV1	61	3,3E-21
MER57E3|LTR|ERV1	63	7,06E-17
TGAGn|Simple_repeat	17	3,26E-10
TGn|Simple_repeat	425	1,55E-08
CAn|Simple_repeat	395	1,7E-08
Simple_repeat	1978	3,05E-07
CGGn|Simple_repeat	147	5,88E-07
CCGn|Simple_repeat	140	7,62E-07

Table 4.4: Top terms of genome ontology associated with repetitive regions

4.2.3. ATRX is enriched at promoters and gene bodies

To analyse the location of ATRX relative to genes, I overlapped the 8,454 ATRX peaks identified in Chapter 3 with the dataset from RefSeq (O'Leary et al., 2015). 25% of the ATRX binding sites were found at gene promoters, almost 40% within gene bodies and more than 35% were intergenic (***Figure 4.2***.a). As almost two third of the ATRX binding sites appeared to be intragenic, I further analysed the average distribution of ATRX across genes and observed a strong signal enrichment at the transcription start site (TSS) (***Figure 4.2***.b). A subtle ATRX enrichment was also noticed near the transcription end site (TES). Regarding the high enrichment of ATRX at TSS, I analysed the enrichment of the H3K4me3, a histone modification associated with active promoters (Santos-Rosa et al., 2002), at ATRX binding sites, which were subdivided based on their RefSeq location. Using H3K4me3 ChIP-seq data performed in the same LCL background as the ATRX ChIP-seq (FF-Ctr, TA-Ctr and CB-Ctr), I detected a high enrichment in H3K4me3 at ATRX binding sites spanning the TSS (promoter) whereas the gene body and intergenic ATRX binding sites were depleted for this chromatin mark. This observation suggests that ATRX may have an affinity for a subset of putative active promoters (***Figure 4.2***.c).

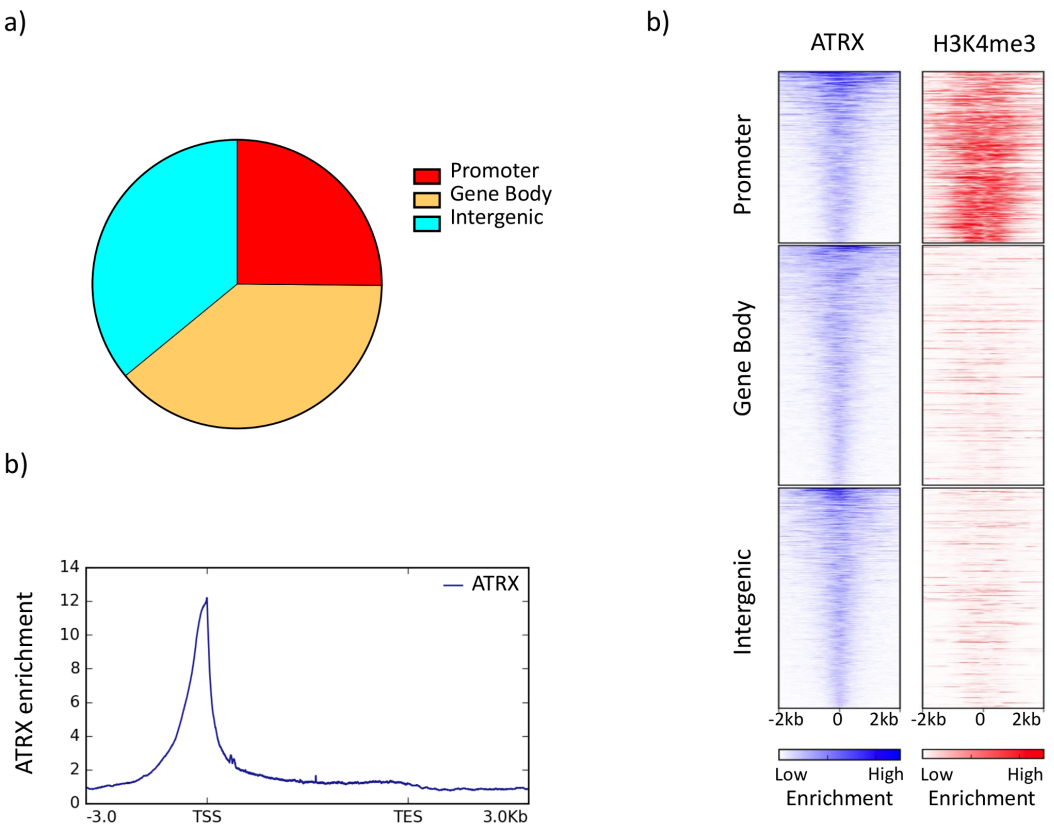


Figure 4.2:Distribution of the ATRX binding sites relative to genes. a. Analysis of ATRX binding sites relative to the RefSeq annotation. b. Average signal profile of ATRX across all RefSeq genes and their 3kb flanking regions. TSS: Transcription start site. TES: Transcription end site. c. Heatmap showing the enrichment of the H3K4me3 histone marks corresponding to each ATRX binding site.

4.2.4. ATRX interacts with CpG island containing regions

A common feature of most of the human promoter regions is the presence of CpG islands. Thus, I sought for the presence of CpG islands at ATRX binding sites using the CpG island annotation from the UCSC dataset. Approximately 30% of the ATRX binding sites contained annotated CpG islands with a high concentration at promoter regions (***Figure 4.3***.a). By contrast, only 9% of the annotated CpG islands presented an ATRX enrichment (***Figure 4.3***.b).

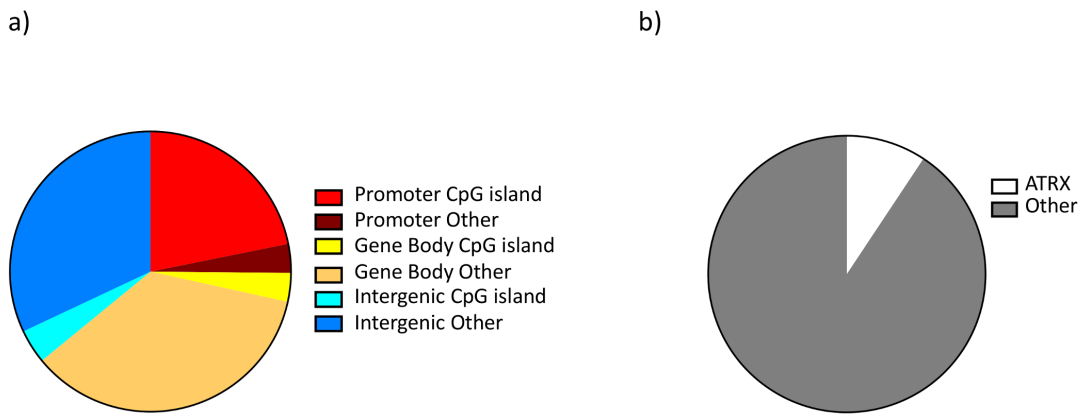


Figure 4.3: ATRX and annotated CpG islands. a. Distribution of annotated CpGs islands at ATRX binding sites. b. Proportion of ATRX binding sites at annotated CpG islands.

I performed Bio-CAP-seq experiments to confirm the presence of CpG islands at a subset of ATRX binding sites (Blackledge et al., 2011). This technique takes advantage of the properties of the zinc finger CxxC (ZF-CxxC) domain that specifically recognise unmethylated CpG clusters. By coupling ZF-CxxC domain with biotin, non-methylated clusters of CpG that are bound to the complex can then be purified from fragmented genomic DNA using Neutravidin beads and subsequently purified using sequential elution with increasing salt concentration (***Figure 4.4***.a). First, I tested the enrichment of CpG clusters by qPCR at three regions: (1) at the GAPDH promoter which displays non-methylated CpG islands and was used as positive control, (2) approximately 600 bp downstream of the GAPDH promoter region to test the robustness of the assay, and (3) at a G-depleted region, without annotated CpG islands, localised in chromosome 8, used as negative control (***Figure 4.4***.b). At low salt concentration (300 mM), the eluate showed a high background with abundant unspecific enrichment at the GAPDH gene body and the negative control. Increasing the salt concentration results in an enrichment of the signal at the positive control and a decrease of the background at the GAPDH gene body and negative control. 700 mM and 1,000 mM fractions displayed a

significant enrichment over the background, 36 and 24 times, respectively. As described by Blackledge and colleagues (Blackledge et al., 2011), most of the unmethylated CpG clusters are present in the 700 mM fraction (***Figure 4.4***.b). Thus, the 700 mM and 1,000 mM fractions were pooled and sequenced. I analysed the datasets and corrected for potential artefacts using the input associated with each Bio-CAP-seq experiment. The Bio-Cap-seq showed a signal enrichment at unmethylated annotated CpG islands and a background signal at fully methylated annotated CpG islands and CpG clusters identified by Infinium Human Methylation 450 Bead Array platform (Methyl450) (Consortium, 2012) or Reduced Representation Bisulfite Seq (RRBS) (Meissner et al., 2008) (***Figure 4.4***.c). Moreover, Bio-Cap-seq displayed enrichment over regions that were not annotated as CpG islands; including promoters and unmethylated CpG clusters tested by Methyl450 or RRBS. By overlapping the Bio-CAP-seq datasets with the annotated CpG islands and the ATRX peak call datasets, I observed that almost 40% (corresponding to 939 + 2,346 + 64 in ***Figure 4.4***.d) of the ATRX binding sites contained CpG clusters. The large majority of these clusters were partially to fully non-methylated (corresponding to 939 + 2,346 in ***Figure 4.4***.d). Surprisingly, a very restricted number of the ATRX binding sites (<1 %) contained fully methylated CpG islands (corresponding to 64 in ***Figure 4.4***.d). By contrast, approximately a third of the annotated CpG islands were characterised as fully methylated based on the Bio-CAP-seq data (corresponding to 9,045 in ***Figure 4.4***.d). Taken together, all these results suggest that ATRX has an affinity for CpG cluster-containing regions and preferentially interacts with CpG clusters that are partially to fully non-methylated rather than with full methylated CpG islands.

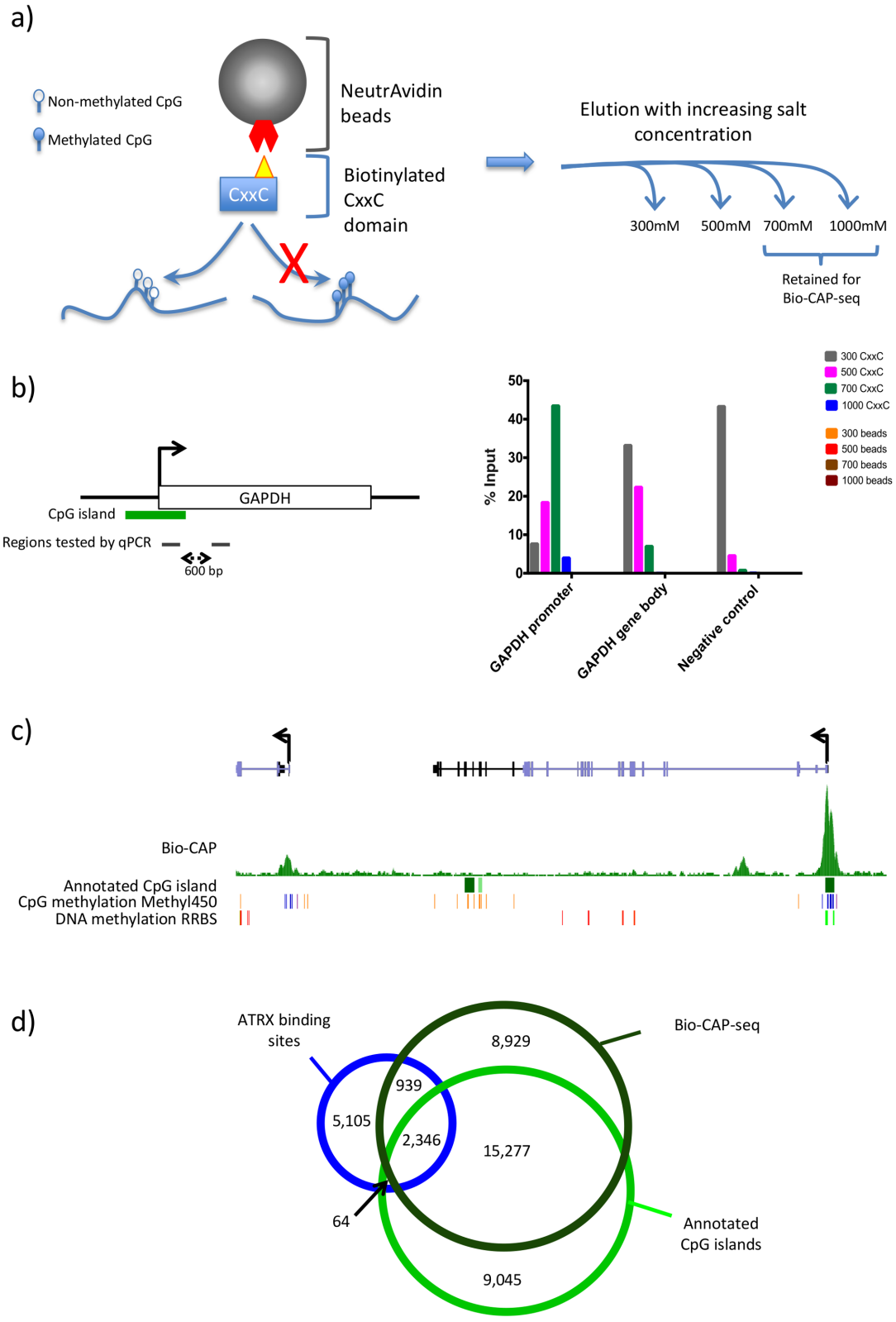


Figure 4.4: Analysis of the CpG clusters using Bio-CAP. a. Schematic representation of basic Bio-CAP principle. b. Analysis of the enrichment in CpG clusters during the sequential elution using Bio-CAP followed by qPCR. 300: elute obtained after elution using 300mM NaCl. 500: elute obtained after elution using 500mM NaCl. 700: elute obtained after elution using 700mM

NaCl. 1000: elute obtained after elution using 1000mM NaCl. c. Screen shot illustrating the ability of Bio-CAP-seq to detect unmethylated CpG clusters. Bio-CAP is a pool of three biological replicates using LCL derived from three independent unaffected controls. CpG methylation Methyl450; associated color code: Orange: methylated, Purple: partially methylated, Blue: Unmethylated. DNA methylation RRBS (Meissner et al., 2008); associated color-code: Red: 100% methylated, Yellow: 50% methylated and Green: 0% methylated. d. Overlap between ATRX binding sites, annotated CpG islands and Bio-CAP-seq datasets.

4.2.5. Almost 50% of the ATRX binding sites contain putative G4 forming sequences

Previous work has shown that CpG clusters associated with putative G4 forming sequences were less prone to high levels of DNA methylation (Halder et al., 2010). ATRX has been shown to bind G4 (Law et al., 2010). Moreover, I observed that a significant proportion of ATRX binding sites in LCL were associated with partially to fully nonmethylated CpG clusters. Therefore, I investigated the presence of putative G4 forming regions at ATRX binding sites. Based on putative quadruplex sequence (PQS) predicted by Quadparser used in (Sahakyan et al., 2017), I observed almost 50% of the ATRX binding sites contained PQS, representing a total of 11,046 PQS. PQS were present in 21% of the promoter peaks, 13% of the gene body peaks and 12.5% of the intergenic peaks (Figure 4.5.a). The subgroup of ATRX binding sites associated with promoters showed the strongest association with PQS with more than 80% (1777 out of 2126) of its targets containing at least one PQS (Figure 4.5.a). On average this subgroup also displayed more PQS per peaks (with an average of approximately 3) than the gene body and intergenic subgroups of ATRX binding sites, both with an average < 1 (Figure 4.5.b). A larger number of PQS containing ATRX binding sites were found in the promoter subgroup. Therefore, I corrected for this overrepresentation by considering only PQS-containing ATRX binding sites in each subgroup to determine

the average number of PQS per ATRX binding sites in each subgroup. The enrichment in PQS at promoter spanning ATRX binding sites remained even under these conditions. Thus, an average of 3.5 PQS per ATRX binding sites spanning promoters whereas the gene body and intergenic subgroups of ATR binding sites showed an average of 2.2 PQS (Figure 4.5.c). Even though almost 50% of the ATRX binding sites overlapped with putative G4, these QPS only represent a small percentage of the total predicted G4 forming sequences (~ 1.5%) (Figure 4.5.d). Altogether these results suggested that ATRX has an affinity for a restricted number of PQS containing regions at least in LCL.

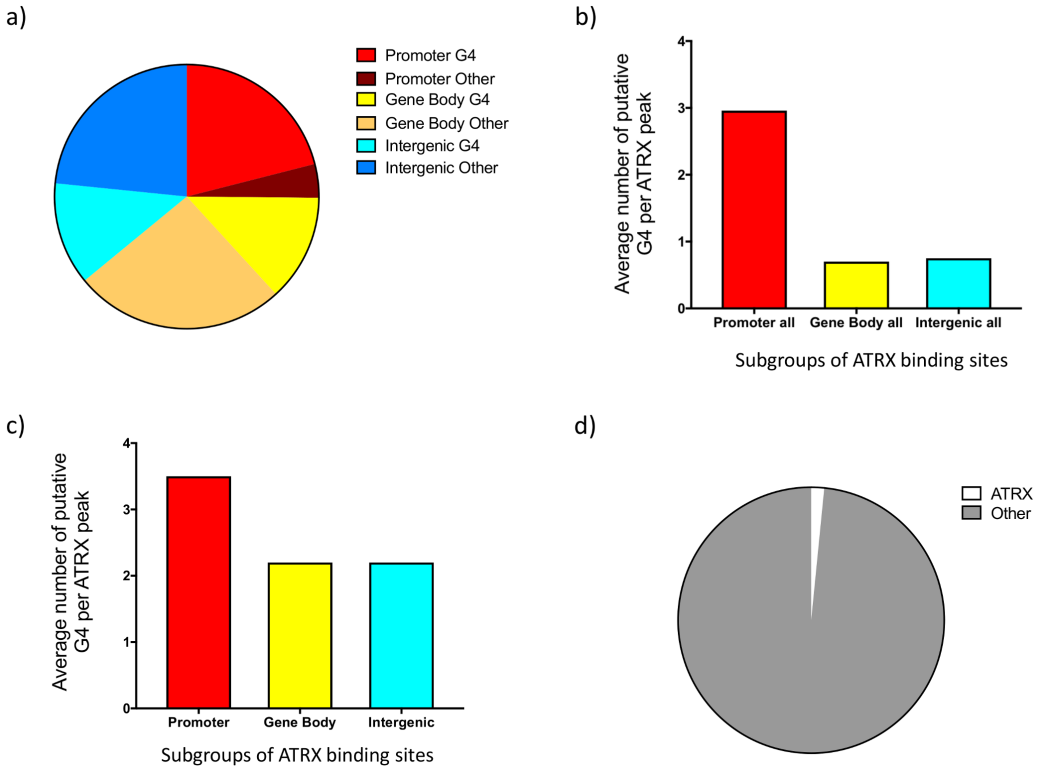


Figure 4.5: Analysis of putative G4 forming sequences at ATRX binding sites. a. Distribution of the PQS and non-PQS containing sequences at ATRX binding sites. b. Average number of PQS per ATRX binding site present at each subgroup: promoter, gene body and intergenic ATRX binding sites. c. Average number of PQS per ATRX binding site present at each subgroup only considering the ATRX binding sites spanning PQS. d. Proportion of ATRX binding sites at PQS genome-wide.

4.2.6. The majority of the ATRX binding sites contain nucleosome-depleted regions

Previous genome-wide analysis on PQS has associated high enrichment of PQS with gene promoter regions correlating with nuclease hypersensitive sites (Huppert and Balasubramanian, 2007). Based on the PQS enrichment at ATRX targets spanning TSS, I performed ATAC-seq experiments in four LCL derived from unaffected controls. As experimental controls, I also carried out ATAC-seq experiments on genomic DNA (gDNA) formerly extracted from two of these LCL. I hypothesised that if the Tn5 transposase, which is used to insert sequencing adapters into accessible regions of chromatin in the ATAC-seq protocol (Buenrostro et al., 2013), had no bias toward specific sequences, it should therefore evenly transpose into gDNA and the resulting sequenced track should not display any significant enrichment over background. Therefore, I used the ATAC-seq performed on gDNA as a control in a similar manner than an input in ChIP-seq experiment (discussed in Chapter 3). I considered any regions significantly enriched in these samples as artefacts and removed them for the subsequent analyses.

Using the ATAC-seq dataset, I investigated the chromatin accessibility and identified that almost 90% of ATRX binding sites contained regions of open chromatin (Figure 4.6.a). Regarding the very high proportion of ATRX binding sites displaying open chromatin regions, I performed the reverse analysis to assess for the presence of potential artefacts. Less than 15% of the open chromatin regions identified by ATAC-seq contained at least one ATRX binding site (Figure 4.6.b), thus supporting the notion that ATRX was associated with a subset of nucleosome-depleted regions.

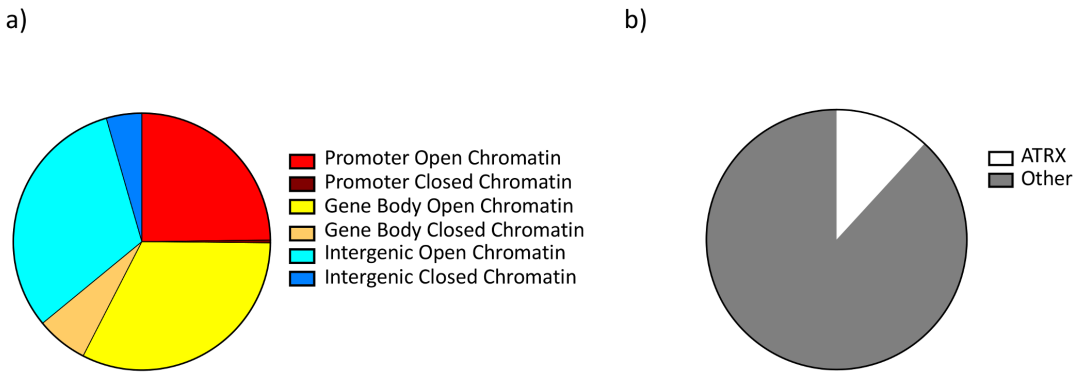


Figure 4.6: Analysis of chromatin accessibility at ATRX binding sites. a. Representation of open chromatin sites at ATRX binding sites. b. Proportion of ATRX binding sites at open chromatin sites identified by ATAC-seq.

4.2.7. Almost 95% of the ATRX binding sites contained at least one TFBS

Open chromatin has been associated with regulatory elements including promoters, CTCF and transcription factors binding sites (Song et al., 2011). Based on the high enrichment in open chromatin regions at ATRX binding sites, I assessed this dataset for the presence of transcription factors binding sites (TFBS). Using Uniform TFBS from ENCODE, I collected the available datasets containing the TFBS (including RNA polymerase subunits) identified by ChIP-seq performed in LCL (Consortium, 2012). Using these datasets, I discovered that almost 95% of the ATRX binding sites contained at least one TFBS. By analysing the TFBS for each factor independently, the RUNX3 TFBS were the most frequent at ATRX binding sites with 85% of them containing at least one RUNX3 binding site. Other TFBSs including PAX5, NFIC, EBF1, ATF2, EP300, FOXM1, Sp1 and POU2F2 were present at least in 60% of the ATRX binding sites. Interestingly, when focusing on the percentage of ATRX binding sites at each TFBS instead of the percentage of each TFBS at ATRX binding sites, the distribution differed. ZNF274 was the most enriched in ATRX binding sites, which were present at

more than 50% of them. By contrast only 1% of the ATRX binding sites displayed ZNF274 TFBS. Similarly, ATRX was present at more than 45% of the POLR3G TFBS but ATRX binding sites were strongly depleted in POLR3G TFBS. This discrepancy was explained by the low number of TFBS (233 and 211, respectively) for these two proteins. They represent the less enriched TFBSs in the datasets. By contrast, the other proteins included in the top 5 of the least frequent TFBSs, BRCA1, ZZZ3 and NFE2, representing 551, 713 and 772, respectively, did not showed such high ATRX enrichment (e.g. ZZZ3 was among the fifth TFBS with the lowest enriched in ATRX (11%)). On the other hand, RUNX3 contained the highest number of TFBS (67965) and even though highly enriched at ATRX binding sites, only 15% of its sites where enriched in ATRX. Similarly, CTCF and SMC3 had a large number of TFBS, 62580 and 30517, respectively. Yet, they were only in the second quarter regarding their enrichment at ATRX binding sites. Furthermore, CTCF and SMC3 were among the three TFBS least enriched in ATRX binding sites. These observations suggested that ATRX has a particular affinity for RUNX3 TFBS containing regions in LCL.

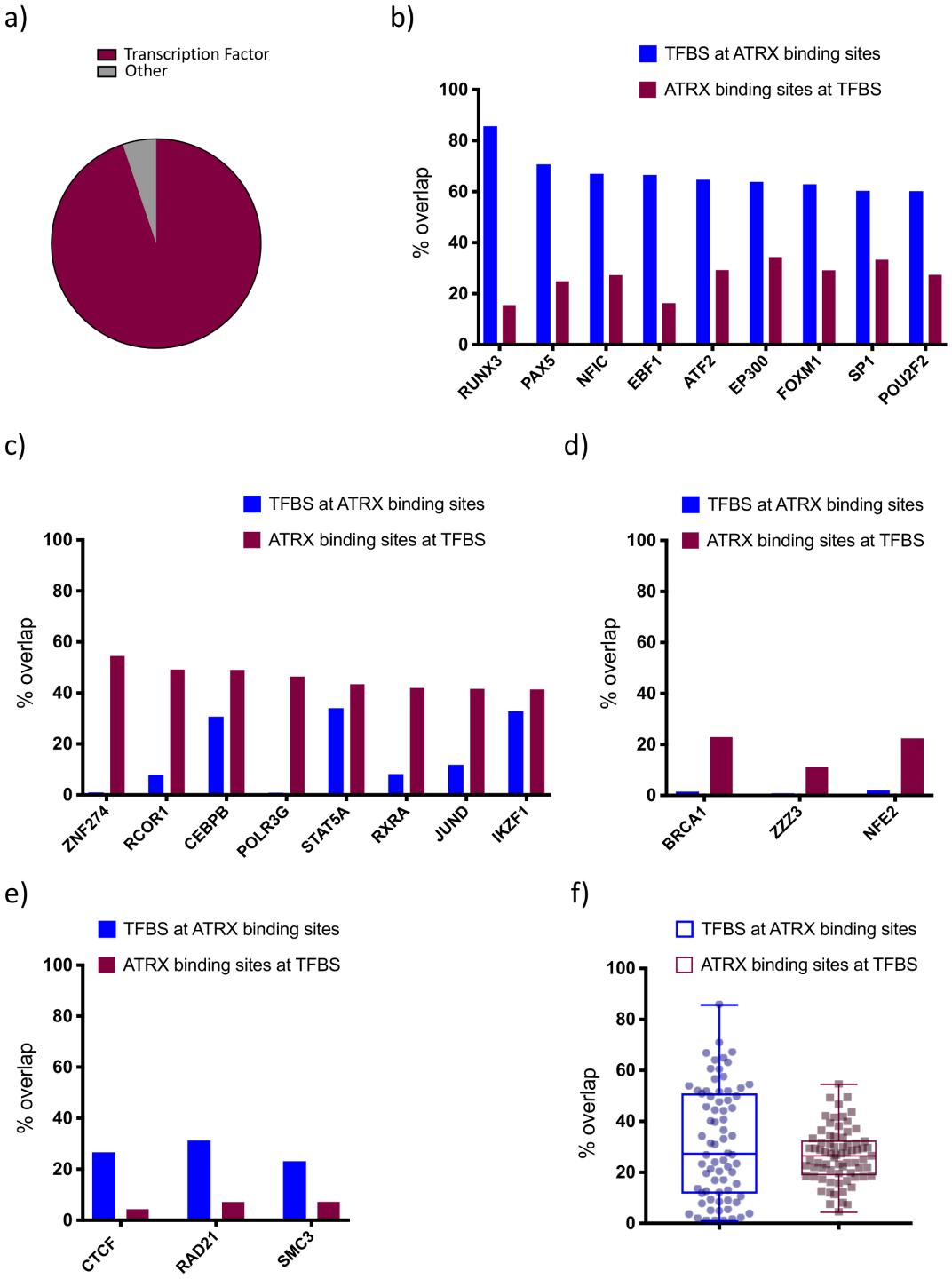


Figure 4.7: Analysis of TFBS relative to ATRX binding sites. a. Enrichment of TFBS at ATRX binding sites. b-f. Analysis of TFBS at ATRX binding sites (in blue) vs ATRX binding sites at TFBS (in burgundy) based on their % overlaps. b. TFBS most enriched in ATRX binding sites. c. TFBS most enriched in ATRX. d. TFBS less enriched at ATRX binding sites. e. TFBS less enriched in ATRX. f. Box and whiskers blots showing the distribution enrichment of all the TFBS tested at ATRX binding sites (blue) and of ATRX at each TFBS (burgundy) based on their percentage overlaps.

4.2.8. Motif analysis conducted on ATRX binding sites identified consensus binding sequences associated with transcription factors

Subsequently, I performed motif analyses to identify potential enrichment motifs at ATRX binding sites. In a first step, I focused on all the ATRX binding sites and used the human genome as background. The top hits for the analysis of the known motifs were associated with BATF and the RUNX family (***Figure 4.8***). BATF and RUNX3 were present in more than 50% of the ATRX binding sites (***Figure 4.7***).

Rank	Motif	Name	P-value	% of Targets
1		BATF	1.00E-146	11.48%
2		RUNX(Runt)	1.00E-142	13.26%
3		RUNX1(Runt)	1.00E-141	17.14%
42		Egr2(Zf)	1.00E-17	4.11%
45		Egr1(Zf)	1.00E-14	11.34%

Figure 4.8: Motif analysis carried out on all ATRX binding sites to identified enrichment in know motifs.

Then, I attempted to untangle the ATRX affinity for a subset of open chromatin regions and not to another (***Figure 4.6***). So, I restricted the motif analysis to ATRX binding sites containing open chromatin regions. Instead of using random sequences from the human genome as background, I selected the open chromatin regions identified by

ATAC-seq that did not overlap with any of the ATRX ChIP-seq peaks. In this analysis the top candidate motifs were also associated to RUNX family (***Figure 4.9***).

Rank	Motif	Name	P-value	% of Targets
1		RUNX (Runt)	1.00E-39	15.71%
2		RUNX1 (Runt)	1.00E-31	19.56%
3		RUNX2 (Runt)	1.00E-27	16.25%
11		Egr2(Zf)	1.00E-08	4.38%
12		Egr1(Zf)	1.00E-08	12.14%

Figure 4.9: Motif analysis carried out on ATRX binding sites containing at least one open chromatin region and using the ATAC-seq peaks not associated with ATRX as background.

Then, I performed the motif analysis using only the ATRX binding which did not contain open chromatin regions and I used the human genome as background. In this case, G-rich candidate motifs associated with EGR family supplanted the motif associated with the RUNX family as top ranked (***Figure 4.10***). However, the p-values were relatively high. Indeed, in the analysis considering all the ATRX binding sites, the first motif associated with the EGR family was only ranked 42, nevertheless its p-value

was lower than in the latter analysis in which it was ranked 1 (***Figure 4.8*** and ***Figure 4.10***).

SP1 binding sites were present in more than 50 % of the ATRX binding site (***Figure 4.7***.b) and its binding site is a putative G4 forming sequences. Therefore one could expect that motif associated with SP1 would be present among the best ranked motifs in all three analyses. Unexpectedly, motif associated with SP1 was only detected within the top 15 highest ranked motifs in the third analysis (***Figure 4.10***). Furthermore, the more the dataset tested was restrictive, the less significant the p-values were (***Figure 4.8 - Figure 4.10***).

Rank	Motif	Name	p-value	% of Targets
1		Egr2(Zf)	1.00E-11	5.54%
2		Egr1(Zf)	1.00E-05	11.19%
3		Tcf4	1.00E-05	7.22%
4		RUNX(Runt)	1.00E-05	12.03%
5		RUNX1(Runt)	1.00E-04	15.27%
11		Sp1(Zf)	1.00E-03	3.66%

Figure 4.10: Motif analysis carried out on ATRX binding sites, which did not contain open chromatin regions, and using random sequences from the human genome as background.

4.2.9. ATRX binding sites are enriched in active regulatory elements

A large proportion of the ATRX binding sites contained open chromatin regions and TFBS. Therefore, I further investigated the presence of regulatory elements in this dataset using GenoSTAN (Genomic STate ANnotation) (Zacher et al., 2017). This model was designed to identify chromatin states such as promoters and enhancers within a dataset using chromatin marks and accessibly. To perform this analysis, I selected all the open chromatin sites identified by ATAC-seq complemented with all the ATRX binding sites that do not contain open chromatin regions. Then, I used chromatin marks, RNA polymerase and CTCF ChIP-seq datasets performed in LCL for the GenoSTAN analysis at these selected sites. By refining the analysis using TFBS and RefSeq dataset, I identified 7 chromatin states including active promoters, active promoters associated with CTCF, active enhancers, CTCF sites, poised enhancers (containing active and repressed chromatin marks, or high H3K4me1 but low H3K27ac), repressed regions and background regions (Figure 4.11.a. Left pie chart). Within this dataset containing more than 63, 000 sites, the most represented states were (1) Poised enhancers (~ 32%), (2) CTCF sites (~ 25%) and (3) the active promoters and enhancers (~ 16% each) (Figure 4.11.a. Left pie chart). I then extracted the ATRX binding sites from the dataset analysed by GenoSTAN and evaluated their distribution relative to the identified states (Figure 4.11.a Right pie chart). Interestingly, the ATRX binding sites showed a strong enrichment (~ 50%) toward the active enhancers (versus ~ 16% when considering all the analysed sites) (Figure 4.11.a. Right pie chart). ATRX binding sites also displayed enrichment in active promoters (~28%) compared to the dataset with all sites (~16%). By contrast, ATRX binding sites appeared depleted in CTCF (~0.5%) sites and poised enhancers (~1.7%) relative to all analysed sites (~ 25%

of CTCF and ~ 32% poised enhancers). Overall, ATRX binding sites appeared highly enriched in active regulatory elements.

To reinforced this observation, I analysed the ATRX binding sites in the context of the chromatin state segmentation done by hidden Markov model (ChromHMM) from ENCODE dataset available for the LCL (Ernst et al., 2011). Similarly to the GenoSTAN analysis, I first assessed the distribution of the chromatin states at all the sites contained in the ChromHMM dataset (Figure 4.11.b. Left pie chart). Similar to the data obtained by GenoSTAN, the most represented state, among approximately 570,000 identified sites, was associated with weak/poised enhancers (~ 31%). Subsequently, I assessed the distribution of the chromatin states identified by ChromHMM at ATRX binding sites only. Similarly to the GenoSTAN analysis, I observed a strong enrichment of active enhancers and promoters, composing approximately 43% and 33% of all the ATRX binding sites, respectively (Figure 4.11.b. Right pie chart). Furthermore, I also observed a decrease in the percentage of insulators (associated with CTCF sites) and weak/poised enhancers compared to the dataset containing all sites identified by ChromHMM. Taken together, these observations strongly support that ATRX binding sites were enriched in active regulatory elements.

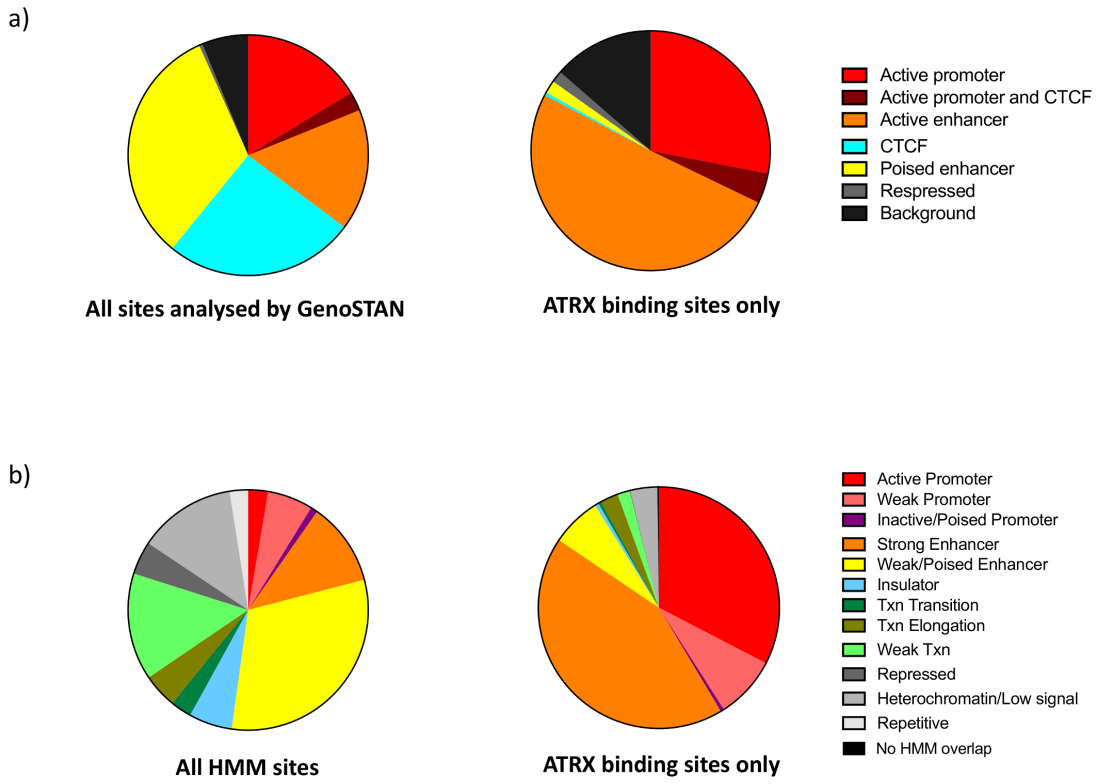


Figure 4.11: Identification of regulatory elements at ATRX binding sites. a. Analysis performed using GenoSTan. Pie chart representing the distribution of the identified chromatin states at all the sites included in the dataset (left), at the ATRX binding sites only (right). b. Analysis performed using publicly available ChromHMM dataset done in LCL. Pie chart representing the distribution of the identified chromatin states at all the sites identified by ChromHMM (left), at the ATRX binding sites only (right). Txn: transcription

4.2.10. ATRX binds functional enhancer clusters of the *MYC* gene

To further assess the presence of ATRX at active enhancers, I tried to identify functionally tested enhancers in LCL. Work from Liang and colleagues investigated the effects of RNA expression of enhancers, located −428 and −525 kb upstream of the *MYC* promoter, on *MYC* expression (Liang et al., 2016). After observing the presence of bi-directional expression at these enhancers, they used shRNA knockdown strategy against these targets and analysed the *MYC* expression. The shRNA experiment resulted in more than 80% knock down of the −428 and −525 kb enhancer expression and was associated with a strong reduction (> 70%) of *MYC* expression (Liang et al., 2016).

Therefore, I analysed the ATRX enrichment at and around the *MYC* locus. ATRX displayed a complex enrichment pattern in this region and more than ten ATRX binding sites were observed including at the −428 and −525 kb enhancers (Figure 4.12). ATRX enrichment overlapped with active chromatin marks such as open chromatin regions identified by ATAC-seq and histone modification such as H3K27ac, H3K4me1 or H3K4me3. ATRX binding sites were also enriched in H3.3. ChromHMM performed in LCL (Ernst et al., 2011) further confirmed that most of the ATRX binding sites co-localised with active enhancers (orange). I performed Capture-C using the *MYC* promoter as viewpoint to assess the interactions between *MYC* promoter and the ATRX binding sites within this 3 Mb region. The Capture-C pattern was also very complex and highlighted interactions between the *MYC* promoter and the enhancers bound by ATRX including the −428 and −525 kb enhancers supporting the ATRX association with a subset of active regulatory elements.

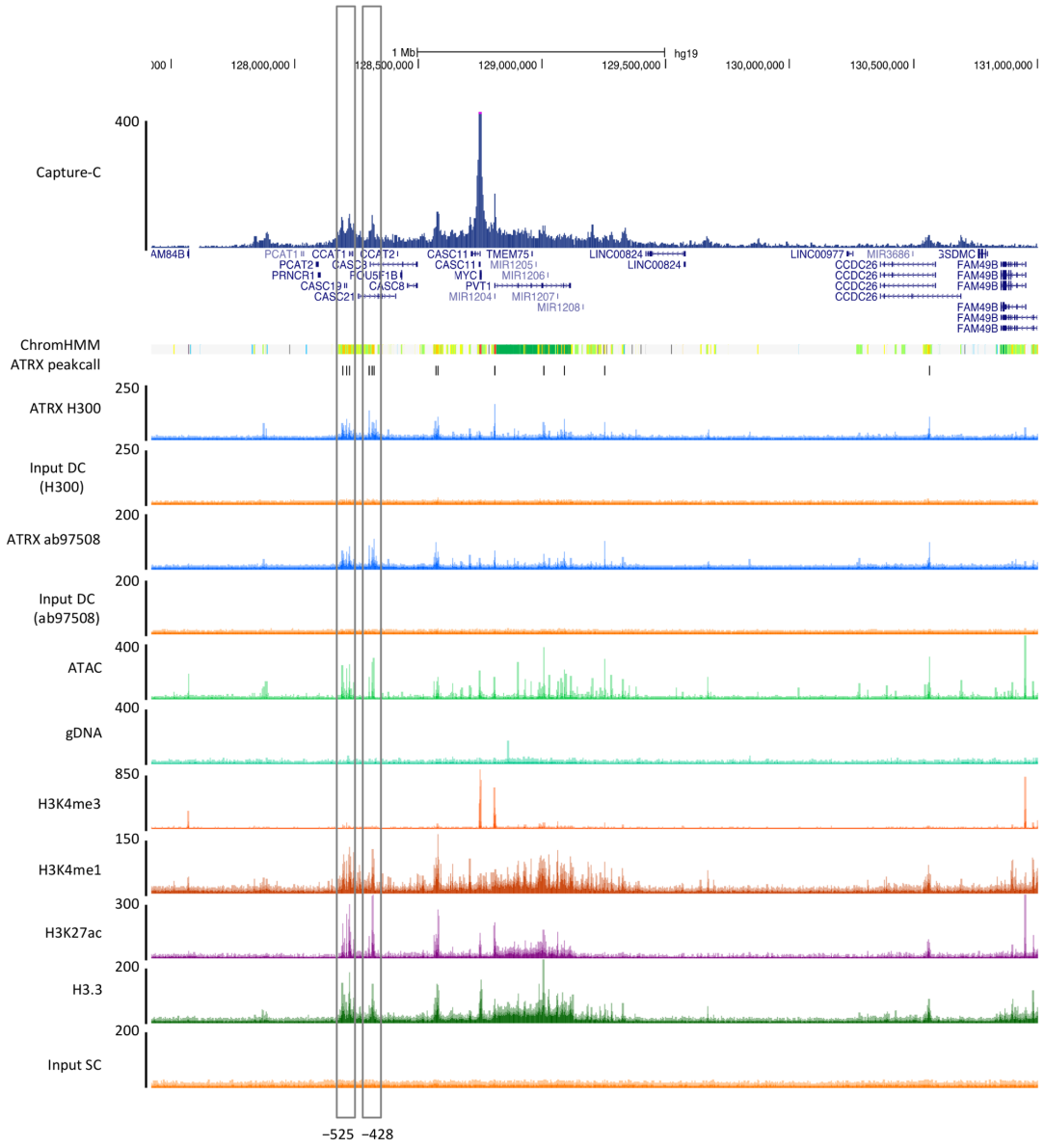


Figure 4.12: ATRX enrichment around the MYC locus. The grey border rectangles highlight cluster of enhancers containing the −428 and −525 tested by Liang and colleagues (Liang et al., 2016). ChromHMM color code corresponds to the legend in Figure 4.11.b. The main features including strong enhancer (orange), active promoter (red), weak enhancer (yellow) and transcription (green)

4.3. Discussion

Using a robust dataset of ATRX ChIP-seqs described in Chapter 3, I have investigated the ATRX enrichment. I confirmed that ATRX was enriched at previously known targets including the genomic loci of the large and small subunits of rDNA. This characteristic appeared to be shared, at least to some extent, across cell lines and species and it is suggested that ATRX is important to maintain the integrity of these sites (Gibbons et al., 2000, Law et al., 2010, Udugama et al., 2018). In ATR-X patients, rDNA showed an abnormal DNA pattern (Gibbons et al., 2000) and in mouse ES cells, loss of ATRX results in more than 50% of copy number losses at both SSU-rDNA and LSU-rDNA (Udugama et al., 2018). These results were further supported by the lower rDNA copy observed in ATRX-mutated human ALT samples than in non-ALT tumours (Udugama et al., 2018). In accordance with ATRX occupancy at PCH, ATRX was also enriched at the pericentromeric satellite HSATII. Overexpression of the non-coding HSATII RNA has been highly correlated with human cancers including pancreatic ductal adenocarcinomas, kidney, ovarian, lung and prostate (Ting et al., 2011). As observed in the original ATRX ChIP-seq performed in human erythroid cells, ATRX was also enriched at the subtelomeric TAR1 repeats in LCL. However, no significant enrichment was observed at telomeric repeats (Law et al., 2010). One possible explanation could be the differences between the cell types used for the analysis. Indeed, ATRX occupancy at telomeres has been observed to differ between cell types and suggested to decrease during differentiation of mouse ES cells (Wong et al., 2010). Therefore, it is not surprising if ATRX enrichment over input is not detected in terminally differentiated cells. Furthermore, ATRX enrichment at telomeric repeats in human erythroid cells (approximately 16 times over input) was less marked than in mouse ES cells (approximately 40 times over input) (Law et al., 2010). Besides, ATRX

was also enriched at a subset of simple repeats, LTR associated with retrotransposons and tRNA supporting previous data (Law et al., 2010, Voon et al., 2015, Sadic et al., 2015). ATRX enrichment over input was also present at the small nuclear RNA U2. Its associated non-coding RNA is a component of the spliceosome, a complex involved in the splicing of the mRNA precursors (Black et al., 1985). Altogether, the enrichment analysis of repetitive elements confirms most of the data previously described and unveils novel potential repetitive targets for ATRX.

Further enrichment analysis based on ATRX peak calling supports the ATRX enrichment at GC-rich, low complexity sequences, and emphasised the affinity of ATRX for ZNF genes. These results are in accordance with the most recent published ATRX ChIP-seq performed in human cell lines (Valle-García et al., 2016). Biological processes and molecular function enrichment analyses showed enrichment in regulatory processes and nucleic acid binding functions, respectively. These observations are in line with the observed enrichment in ZNF genes as they are putative DNA binding transcription factors.

The analysis of ATRX binding sites based on RefSeq annotation showed that almost 65% of them are either at the promoter region or at the gene body. These data are similar to those observed in human erythroblast but differed from mouse ES cells in which the majority of the ATRX binding sites were intergenic (Law et al., 2010). The analysis of the H3K4me3 enrichment suggested that ATRX may preferentially interacts with active promoters in LCL. Subsequent experiments using annotated CpG island datasets and Bio-CAP-seq assays confirmed that, as previously observed, ATRX interacts with a subset of CpG clusters. Interestingly, the quasi-totality of them is either partially or fully unmethylated. These results are in accordance with the enrichment of ATRX in active chromatin promoters and support the observations of previous work done in

mouse forebrain showing that ATRX preferentially binds to the unmethylated ICR of a subset of imprinted loci (Kernohan et al., 2014). Additionally, these results are not contradictory with the works showing that in mouse ES cells, ATRX preferentially interacted with the methylated alleles of DMR in the imprinted loci (Voon et al., 2015). Indeed, these loci would also be detected by Bio-CAP-seq because they contained unmethylated CpGs. However, regarding the difference observed in ATRX binding and functions at imprinted loci between different tissues/cell types in mouse, any comparisons with results obtained in terminally differentiated human cells would be uncertain.

ATRX is known to interact with a subset of PQS including at telomeres and ψζ VNTR (Law et al., 2010). Among others, PQS have been associated with CpG-cluster-containing sequences and promoter regions (Huppert and Balasubramanian, 2007). Moreover, PQS CpG-cluster-containing sequences appeared to be less prone to high DNA methylation than non PQS ones (Halder et al., 2010). I investigated the presence of PQS at ATRX binding sites and I observed that approximately 50% of them contained at least one PQS, supporting previous published data. More than 80% of the ATRX bound promoter regions contained at least one PQS and on average these regions displayed a higher number of PQS than seen at ATRX binding sites in gene bodies or intergenic regions.

Active promoters, and active chromatin in general tend to be associated with open chromatin regions (Santos-Rosa et al., 2002, Schübeler et al., 2004). Surprisingly for a chromatin remodeler associated with heterochromatin regions, most of the ATRX binding sites contain nucleosome-depleted regions and these observations are unlikely to happen by chance. Indeed, based on 7 cell lines, only 9% of the human genome was identified as nucleosome-depleted (Song et al., 2011). Moreover, ATRX enrichment is

only detected at a subset of the open chromatin sites identified by ATAC-seq in LCL. These results suggest that (1) the overrepresentation of nucleosome-depleted regions within the ATRX binding sites is unlikely to be artificial and (2) the presence of open chromatin is not a sufficient feature by itself for ATRX recruitment and/or stabilization at that sites and consequently, additional factors may be involved.

Nucleosome-depleted sites often coincide with TFBS (Song et al., 2011). Interestingly, the majority of the ATRX binding sites were also overlapping such regions. A deeper analysis showed that amongst the TFBS tested, based on the availability of data in ENCODE, RUNX3 was the most represented in LCL. This co-occurence could be due to the large number of RUNX3 binding sites in LCL. Thus, RUNX3 binding sites were present in 85% of the ATRX binding sites, whereas ATRX occupancy was detected at only 15% of the RUNX3 binding sites. These results suggest that ATRX may be recruited to a subset of RUNX3 sites whereas the opposite situation appears to be very unlikely. Alternatively, their co-localisation may be the result of random distribution, as there is numerous RUNX3 binding sites, ATRX would be more likely to overlap with one of them. One observation against the hypothesis is the correlation between ATRX and CTCF sites. The number of CTCF sites and RUNX3 binding sites were similar and ATRX have been shown to play a role at CTCF sites at imprinted loci in mouse forebrain. Despite these circumstances, CTCF sites were not among the sites the more enriched at ATRX in LCL. Further motif analysis identified RUNX3 and the RUNX family in general as enriched motifs at ATRX binding sites. The RUNX family is evolutionarily conserved across metazoans. RUNX3 has been associated with various biological processes, including neurogenesis (Levanon et al., 2002) and the development of the gastrointestinal tract (Fukamachi and Ito, 2004). In human B cells and LCL, *RUNX3* expression is required to repress *RUNX1* genes illustrating the

interplay involved between RUNX genes (Spender et al., 2005). Moreover, *RUNX3* is considered as a tumour suppressor and its inactivation has been reported in several cancers (Spender et al., 2005).

In contrast to *RUNX3* binding sites, ATRX occupancy was detected at more than 50% of the ZNF274 sites. This KRAB-containing ZNF has been associated with recruitment of SETDB1 and TRIM28 including at the last 3' exon of ZNF genes (Valle-García et al., 2016). This study also supported the ability of ATRX to co-immunoprecipitate SETDB1, ZNF274 and TRIM28 (Valle-García et al., 2016). However, a more recent study using Flag-IP and anion exchange chromatography suggested that ATRX and SETDB1 associate with DAXX in two distinct complexes and that SETDB1 did not co-immunoprecipitate ATRX (Hoelper et al., 2017). This disparity raises questions: what is the basis of ATRX recruitment – does ATRX enrichment at more than 50% of the ZNF274 binding sites result from the presence of H3K9me3 or is there a more direct association between them? Similarly, almost half of the POLR3G binding sites overlap with ATRX. This subunit of the RNA polymerase III is notably involved in the transcription of structural RNA including rRNA and tRNA (Lund et al., 2017). These observations are in line with the results of the current study showing that ATRX is enriched at rDNA and a subset of tDNAs.

ATRX enrichment in regulatory elements was further supported by unsupervised analyses using both GenoSTAN and ChromHMM. The results highlighted that the majority of the ATRX binding sites overlapped with active promoters or enhancers. The portion of CTCF sites however was not high. ATRX binding sites were present at functionally tested enhancer clusters associated with *MYC* in LCL. I further confirmed that these *MYC* enhancers were interacting with the *MYC* promoter using Capture-C. Until now, the HIRA complex has been considered as the chromatin remodeling

complex responsible for H3.3 deposition at active chromatin (Goldberg et al., 2010). The results presented above now suggest that ATRX plays a role in this process and offer an explanation regarding the H3.3 deposition at TFBS which was, in previous experiments, unaffected by HIRA or ATRX KO (Goldberg et al., 2010). Indeed, it is possible that both complexes display some pathway redundancy for a subset of targets including these TFBS. Regarding these two complexes, it would not be the first insight that supports pathway redundancy, shared targets and association with nucleosome-depleted regions. In drosophila, Xnp (ATRX homolog) and HIRA factors, both co-localised with nucleosome-depleted chromatin through independent binding (Schneiderman et al., 2012). Both are also involved in H3.3 deposition and displaced after the assembly of new nucleosomes. Whereas Xnp or HIRA single mutants are viable, xnp/HIRA double mutants die during larval development (Schneiderman et al., 2012). Finally, supporting the role of ATRX in H3.3 deposition at regulatory elements, DAXX binding sites have been recently identified at active regulatory elements in human (Puto et al., 2015). The ATRX association with active regulatory elements may explain its different distribution between cell types and its link with transcribed regions. Increasing numbers of studies suggest that G4 may be involved in gene regulation and DNA replication, two processes that are also associated with ATRX. Thus, PQS and G4 have been found at regulatory elements including promoters and putative enhancers *in vitro* and *in vivo* (Hänsel-Hertsch et al., 2016, Chambers et al., 2015). One hypothesis suggests they may be involved in the stabilisation of the promoter-enhancer interactions (Hegyi, 2015). DNA methylation has also been suggested to affect G4 stability (Stevens et al., 2014) including at imprinted loci and consequently may affect the binding of transcription factors. Considering that the majority of ATRX binding sites in LCL appeared to be associated with active regulatory elements, these observations ask new

questions regarding the role of ATRX that should be further investigated. Does the chromatin environment vary at these regulatory elements in ATRX deficient cells? Could it shed light on previously unknown mechanisms or on the contrary, does it reinforce existing hypotheses that explain the phenotypes observed in diseases associated with ATRX mutations? For example, is there some pathway redundancy with HIRA? Might redundancy explain how H3.3 is deposited at TFBS when either HIRA or ATRX is lost? Further works are undoubtedly required and identifying novel genes affected by ATRX mutations may be a valuable contribution to these investigations. Ultimately, these data are the first study that strongly associates a significant proportion of ATRX binding sites with putative active euchromatin regions in human.

Chapter 5. Identification of novel genes affected by ATRX mutations

5.1. Introduction

ATRX has previously been shown to play a significant role in the transcription of a subset of genes in different organisms. In human, the downregulation of alpha globin gene expression in ATR-X syndrome is probably the most extensively characterised (Law et al., 2010). ATRX is enriched at the ψζVNTR within the alpha globin cluster. The length of this VNTR is highly variable, and strikingly, the longer it is, the more severe is the downregulation of adjacent genes (even between patients with the same ATRX mutation) (Law et al., 2010). The same phenomenon was also observed at the *NME4* locus where another VNTR, enriched for ATRX, is situated in the first intron of the gene. The severity of the downregultation of *NME4* and its downstream gene, *DECR2*, are also correlated with the length of the VNTR. These GC-rich VNTRs are predicted to form G4 structures (Law et al., 2010). Beyond the alpha globin cluster and *NME4*, depletion of ATRX has been shown to affect retrotransposon silencing (Sadic et al., 2015, Elsässer et al., 2015) although a recent investigation suggests that DAXX, but not ATRX, is required for silencing of endogenous retroviruses (Hoelper et al., 2017). In mouse, ATRX depletion has been shown to affect the expression of imprinted genes (Kernohan et al., 2010, Voon et al., 2015). Indeed, in mouse embryonic stem cells, ATRX KO results in the upregulation of alleles that are normally silenced. In mouse forebrain, ATRX KO affects the expression of the maternal allele of *H19* (Kernohan et al., 2010, Kernohan et al., 2014). However, despite the increasing evidence of a relationship between ATRX and the regulation of gene expression, our

knowledge of human targets affected by pathogenic ATRX mutations remains surprisingly limited.

Therefore, I investigated the effects of pathogenic ATRX mutations on the regulation of gene expression in LCL. I identified new gene targets dysregulated in ATR-X patients using a microarray approach and further validated these targets by qPCR and Fuidigm Biomark systems. Then, I selected potential direct targets based on the localisation of ATRX binding relative to the differentially expressed genes (DEG). These data provided the foundation for further attempts to understand the underlying mechanisms by which ATRX may influence gene expression.

5.2. Results

5.2.1. Gene expression analysis of patient-derived LCL using microarrays

To identify novel genes affected by ATRX mutations, I performed a microarray-based expression analysis using RNA from LCL derived from 28 patients and 20 healthy controls (including FF-Ctr, CB-Ctr and TA-Ctr controls used in the ATRX ChIP-seq experiments) (***Table 2.1*** Chapter 2). Previous work had shown that gene expression could be more variable in ATR-X patients that normal controls (Law et al., 2010); consequently, a larger number of cases than controls were selected. The cases were subdivided in two subgroups. The first contains 19 cases with a broad range of pathogenic ATRX mutations whereas the second is composed of 9 cases with the most common mutation (p.R246, ***Figure 5.1***) seen in almost 20% of ATR-X patients. This experimental design not only allows a case-control study but also a comparison between (1) the effects of a specific mutation versus a range of other pathogenic mutations and (2) the effects of mutations present in the ADD domain versus mutations present in the

ATPase domain. For practical reasons, the samples were split into five batches (***Table 2.1*** Chapter 2). Samples from each of the four subgroups (controls, cases with p.R246C, cases with mutations in the ADD domain other than p.R246C and cases with mutation within the ATPase domain) were distributed in each experimental batch to ensure proper correction of batch effects,.

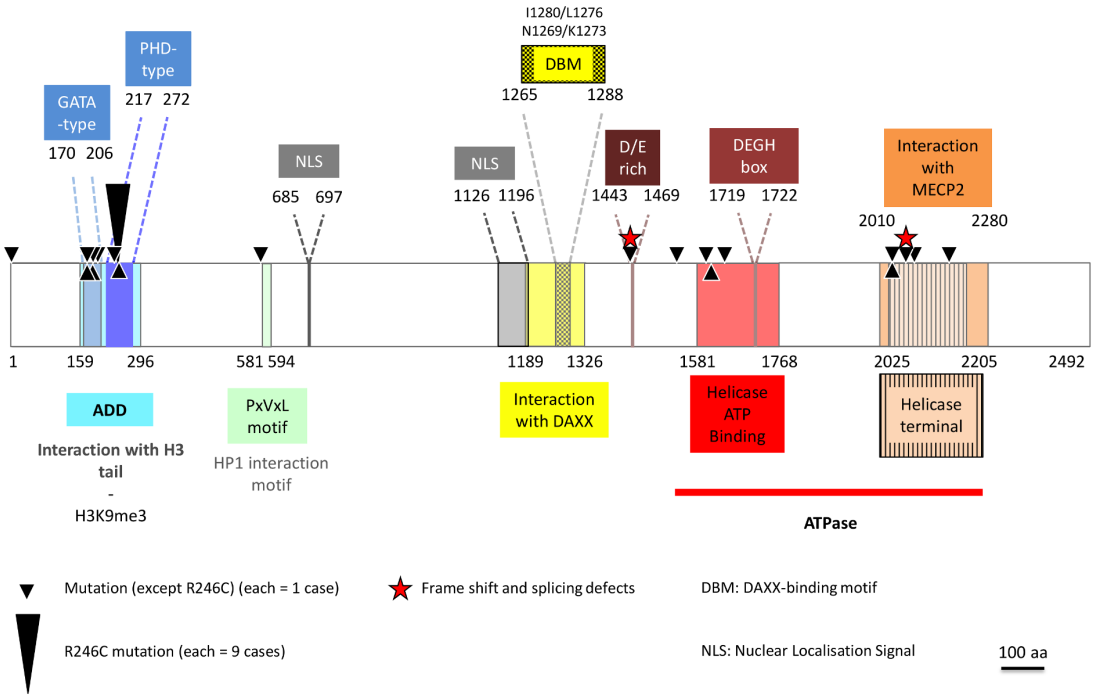


Figure 5.1: Schematic representation of the full-length ATRX protein depicting the position of the ATRX mutations present in the selected LCL derived from ATR-X patients.

The robustness of this experiment design has been assessed by a power calculation, performed by Emmanouela Repapi, using the microarray data from the 2nd and 3rd batches. This analysis supported the experimental design as a sample size of 40 (20 controls and 20 cases) was predicted to detect a 1.5-fold change at 99% of the probes present on the Affymetrix ChIP (corresponding approximately to logFC = 0.18) and therefore give a power of 99% under these conditions (***Figure 5.2***).

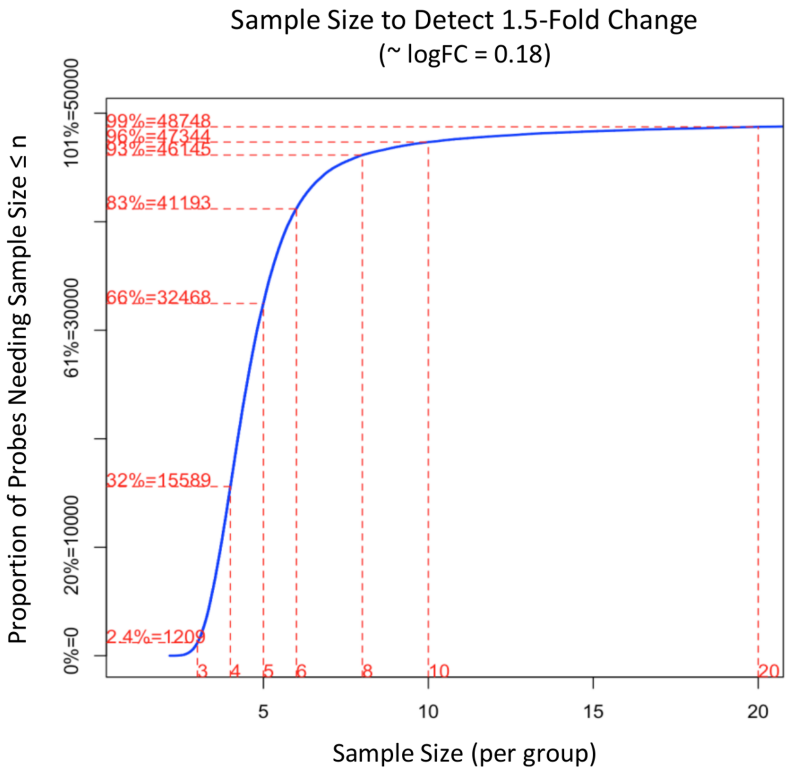


Figure 5.2: Power calculation analysis to detect 1.5-fold change. The analysis is based on detection of the probes present in the array depending of the sample size present in each group (Control and Cases). Analysis performed by Emmanouela Repapi using the R package size functions.

The raw data from the microarray experiment were normalised and corrected for the batch effect (with the help of Emmanouela Repapi). The candidate list was first determined based on an adjusted p-value ≤ 0.05. In total, 388 probes displayed a p-value < 0.05, representing 234 unique DEG (***Figure 5.3***. a and b). The DEG were evenly shared between upregulated (118 DEG) and downregulated (116 DEG) genes. However, only two genes display a |logFC|> 2 (RASGEF1A and LPAR6) (***Figure 5.3***.a) and less than 50% (103) of the DEG have a |logFC| > 0.5 (***Figure 5.3***.a and c). Considering only the DEG with a |logFC| > 0.5, a change in the distribution of the DEG is noticed with an over representation of upregulated DEG (~64%) (***Figure 5.3***.c).

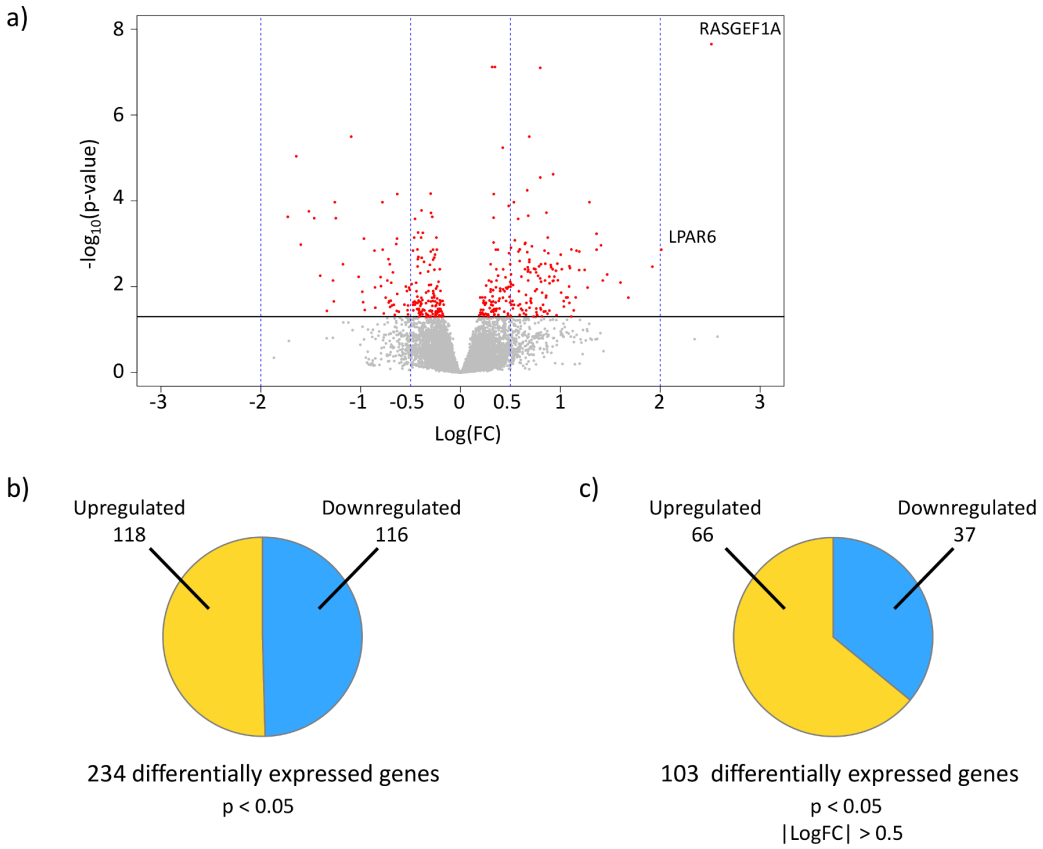


Figure 5.3: Analysis of the DEG obtained by microarray-based experiment. a) Volcano plot highlighting the statistically significant probes (in red) with an adjusted p-value ≤ 0.05 (horizontal black line). The vertical blue lines highlight the regions |logFC|=2 and | logFC|= 0.5. b) Pie plot representing the unique DEG with an adjusted p-value ≤ 0.05. c) Pie plot representing the unique DEG with an adjusted p-value ≤ 0.05 and a |logFC| > 0.5.

Subsequently, gene ontology analyses (Da Wei Huang et al., 2008) revealed a significant enrichment in biological processes linked to transcription in both datasets comprising the 234 significant DEG and the 103 significant DEG showing |logFC| > 0.5 (***Table 5.1***). A noticeable number of ZNF genes were found, which is consistent with these gene ontology analyses (***Table 5.1***).

Gene List	Term	Count	p-value	Genes	%
DEGs p < 0.05	transcription, DNA-templated	32	0.042	ZNF555, ZNF57, ZNF347, CNOT6, ZNF32, ZBTB38, ZNF30, HOXA1, LPXN, STAT4, LBH, TEAD4, VPS36, ZNF595, TXNIP, MBD3L3, L3MBTL4, TRNP1, ZNF506, ARID5A, ZNF649, MAPK11, MBD2, ZNF626, MXD1, ATF7IP2, PURA, MSL3, ZNF718, ADNP2, MAFIP, ZNF418	13.6
	regulation of transcription, DNA-templated	27	0.025	ZNF555, ZNF57, ZNF347, CNOT6, ZNF32, ZNF30, HOXA1, LPXN, STAT4, MEIS2, LBH, MYB, VPS36, ZNF595, L3MBTL4, TRNP1, ZNF506, ZNF649, ZNF626, ATF7IP2, MSL3, RPS6KA5, ZNF718, ADNP2, TCFL5, MAFIP, ZNF418	11.5
	angiogenesis	7	0.045	KLF5, FMNL3, NUS1, LEPR, SOX18, ENPEP, PNPLA6	3.0
	T cell costimulation	6	0.002	BTLA, TRAC, CD80, KLRK1, KLRC4-KLRK1, CD28	2.6
	peptidyl-tyrosine dephosphorylation	6	0.006	PTPRC, PTPRG, DNAJC6, PTPDC1, PTPRO, RNGTT	2.6
	negative regulation of gene expression	6	0.021	SLC35C2, BAK1, AIF1, TRIM6, CCR1, CD28	2.6
DEGs p < 0.05 |logFC| > 0.5	regulation of transcription, DNA-templated	15	0.020	ZNF595, ZNF555, L3MBTL4, ZNF506, ZNF57, ZNF626, ATF7IP2, RPS6KA5, ZNF30, HOXA1, ZNF718, STAT4, MEIS2, LBH, TCFL5	14.0
	lipid transport	4	0.007	NME4, APOL2, APOL1, ABCA6	3.7
	central nervous system development	4	0.024	NDN, AHI1, VCAN, PDX1	3.7
	protein targeting to Golgi	3	0.005	RGPD2, RGPD1, RGPD4	2.8
DEGs p < 0.05 |logFC| > 0.5	regulation of transcription, DNA-templated	13	0.001	ZNF555, ZNF30, ZNF595, ZNF718, L3MBTL4, STAT4, MEIS2, LBH, TCFL5, ZNF506, ZNF57, ZNF626, ATF7IP2	24.5
ATRX binding site within 100kb	transcription, DNA-templated	13	0.009	ZNF555, TXNIP, ZNF30, ZNF595, ZNF718, L3MBTL4, STAT4, LBH, ZNF506, ZNF57, ZNF626, ATF7IP2, ZBTB38	24.5
DEGs p < 0.05 |logFC| > 0.5	regulation of transcription, DNA-templated	11	4.48E-05	ZNF555, ZNF30, ZNF595, ZNF718, L3MBTL4, LBH, MEIS2, ZNF506, ZNF57, ZNF626, ATF7IP2	40.7
Intragenic ATRX binding site	transcription, DNA-templated	11	4.07E-04	ZNF555, ZNF30, ZNF595, ZNF718, L3MBTL4, LBH, ZNF506, ZNF57, ZNF626, ATF7IP2, ZBTB38	40.7
DEGs p < 0.05 logFC < 0.5	regulation of transcription, DNA-templated	11	1.40E-06	ZNF555, ZNF30, ZNF595, ZNF718, STAT4, L3MBTL4, MEIS2, ZNF506, ZNF57, ZNF626, ATF7IP2	52.4
ATRX binding site within 100kb	transcription, DNA-templated	11	1.53E-05	ZNF555, ZNF30, ZNF595, ZNF718, STAT4, L3MBTL4, ZNF506, ZNF57, ZNF626, ATF7IP2, ZBTB38	52.4
DEGs p < 0.05 logFC > 0.5	negative regulation of gene expression	3	0.023	AIF1, CCR1, CD28	9.38
ATRX binding site within 100kb	positive regulation of monocyte chemotaxis	2	0.027	AIF1, CCR1	6.25

Table 5.1: DAVID gene ontology analysis of biological processes. Count: number of DEG associated with GOTERM. %: percentage of DEG associated with GOTERM

5.2.2. Assessing the microarray data in the context of published work

The microarray data were compared with the relevant literature. Interestingly, the expression of *ATRX* did not show a significant change at the RNA level between cases and controls (***Figure 5.4***). These results corroborated the previous work showing that ATRX mutations mainly reduce the level of ATRX protein (Argentaro et al., 2007). In contrast, the RNA levels of *NME4* and *DECR2* were significantly downregulated; logFC values were -1.34 and -0.23 compared to control, respectively (***Figure 5.4***). *NME4* expression is therefore more affected than *DECR2* by ATRX mutations and

NME4 expression data in ATR-X patients were more widely spread out compared to controls (***Figure 5.4***). These observations are also consistent with previous works (Argentaro et al., 2007, Law et al., 2010).

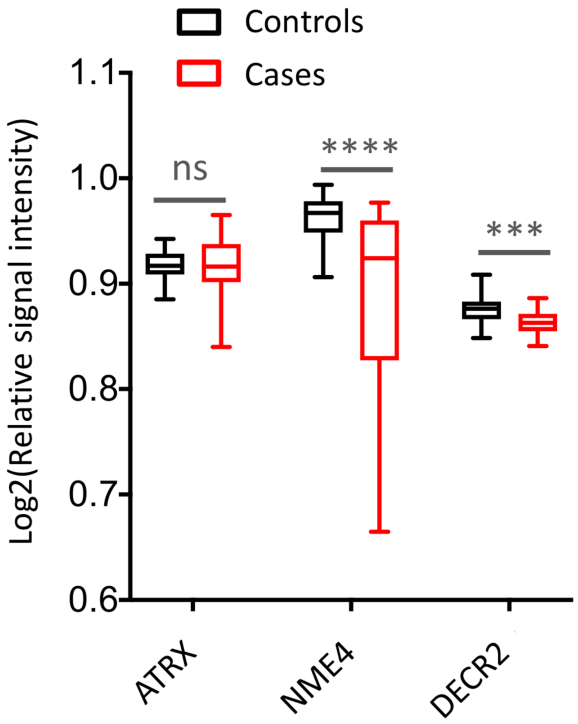


*Figure 5.4: Analysis of the microarray data associated with ATRX, NME4 and DECR2 genes. Box and whiskers plot showing the minimum, 25th percentile, median, 75th percentile and maximum of the relative mRNA expression based on the microarray data. *** p-value < 0.001 and **** p-value < 0.0001. ns: not significant. p-value determined by Mann-Whitney test.*

5.2.3. Assessing the effects of ATRX mutations based on their location

Taking advantage of the experimental design, the effects of various ATRX mutations on gene expression were compared based on their occurrence in patients or their location in the protein. First, the samples displaying the common p.R246C mutation were compared to all the other cases, independently of the location of the mutation, to assess if this specific mutation has specific effects on gene expression. No statistically significant difference (p-value < 0.05) between the two groups was observed (***Figure 5.5***.a). Next, the analysis was further restricted to compare the common p.R246C mutation with the other mutations located in the same domain of the ATRX protein, the ADD domain. Here again no significant difference could be detected (***Figure 5.5***.b).

As, the ADD domain is composed of GATA-like and PHD-like subdomains, this result could reflect a specificity of each sub-domain instead of a specific effect of the common p.R246C mutation *per se*. If it was the case, the effects would be diluted in an analysis only taking into account the ADD domain. To assess this hypothesis, the mutations found in the PHD-like subdomain were compared with those present in the GATA-like subdomain. Even with these criteria, no significant differences were identified (***Figure 5.5***.c). However, by increasing the specificity of the analysis, the number of samples corresponding to the criteria also decreases and so does the power. Therefore, using the available data, no differential effects could be observed between the common p.R246C mutation and the other mutations in the ADD domain or its sub-domains.

The majority of pathological ATRX mutations are located in the ADD and ATPase domains; I next compared the effects of missense ATRX mutations in these two domains. One mutation in the ATPase domain, which results in a frame shift and splicing defects, was excluded to avoid possible bias (***Figure 5.1***). Indeed, this mutation leads to the production of approximately 10% of normally spliced product and two

abnormal spliceoforms that are likely to result in nonsense mediated decay (Picketts et al., 1996). Interestingly, by comparing mutations lying in the ADD domain with those found in the ATPase domain, a differential effect could be detected on a unique gene, *NME4* (p-value < 0.001) (***Figure 5.5***.d). To further analyse the effects of mutations in the ADD versus ATPase domain on *NME4* relative to controls, the relative signal intensities corresponding to *NME4* were plotted for the controls, all the cases, the cases with mutations in the ADD domain and the cases with the mutations in the ATPase domain (***Figure 5.5***.e). The data suggested that mutations in the ATPase domain appeared to have a more dramatic effect on *NME4* expression compared to mutations in the ADD domain and showed an overall more significant downregulation of *NME4* compared to controls (***Figure 5.5***.e).

With the exception of *NME4* expression, no other significant differences have been found by comparing different types of mutations used in the microarray assay (***Figure 5.5***. a-d), therefore all the cases were used as a single group for the rest of the study.

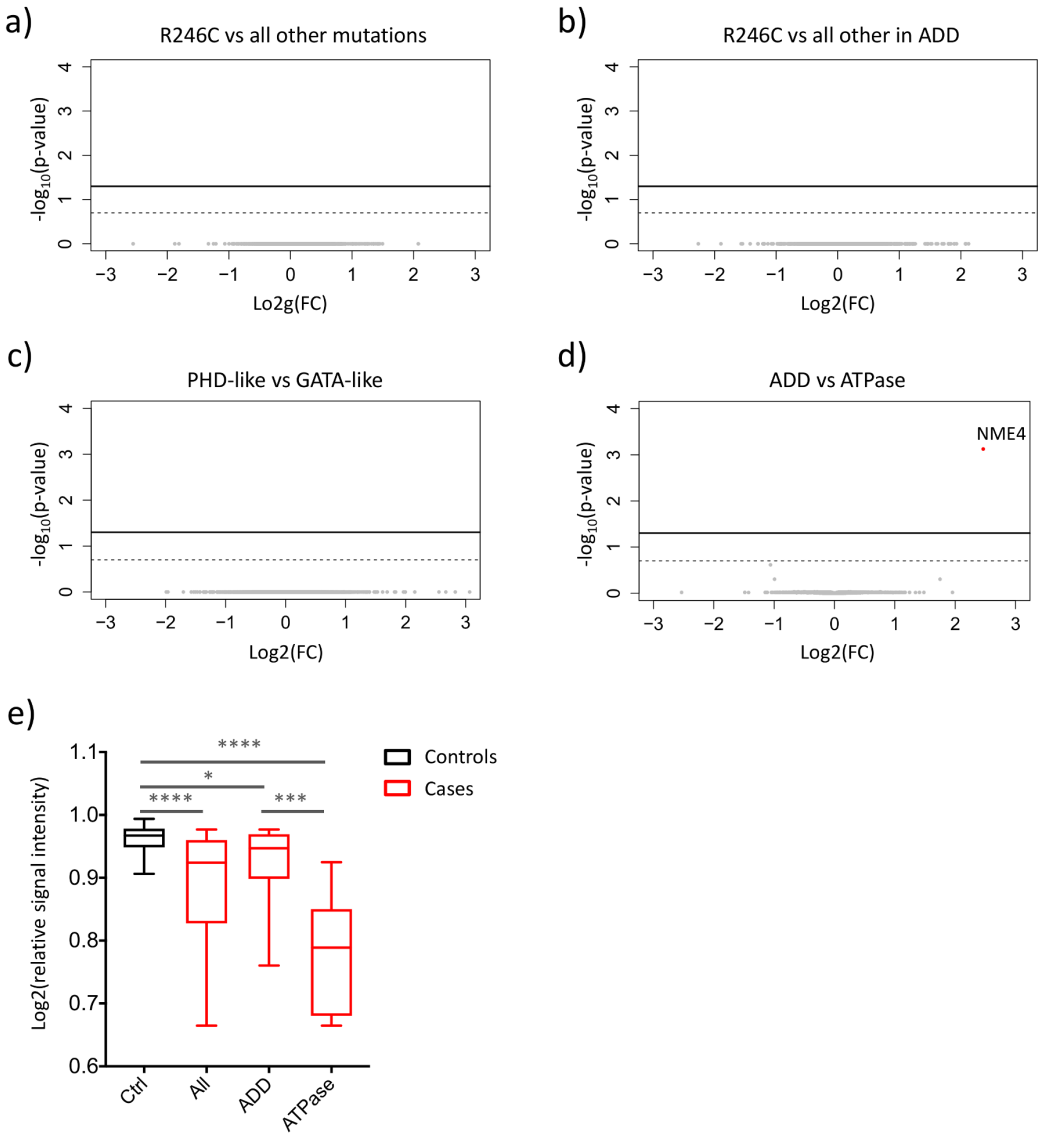


*Figure 5.5: Comparison of ATRX mutations based on their occurrence in patients and their location. a) Differential analysis of the gene expression between the common p.R246C mutation vs all other mutations. b) Differential analysis of the gene expression between the common p.R246C mutation vs all other mutations lying in the ADD domain. c) Differential analysis of the gene expression between the ATRX mutations within the PHD-like vs GATA-like subdomains. d) Differential analysis of the gene expression between the ATRX mutations lying in the ADD domain vs those lying in the ATPase domain. a-d) Black horizontal line represents a p-value = 0.05 and the grey dash horizontal line represents a p-value = 0.2. Grey dots represent no significant probes. Red dot represents a significant probe with a p-value < 0.05. e) Box and whiskers plot (minimum, 25th percentile, median, 75th percentile and maximum) of the relative signal intensity related to NME4 at different subgroups. * p-value < 0.05, *** p-value < 0.001 and **** p-value < 0.0001. p-value determined by Mann-Whitney test.*

5.2.4. Identification of candidate genes potentially directly affected by ATRX mutations

The expression data containing the DEG with a p-value < 0.05 and a |logFC| > 0.5 were cross-referenced with the ATRX ChIP-seq data to identify the genes that were potentially directly affected by ATRX mutations. Using perl and shell scripts, the DEG containing at least one intragenic ATRX binding site were characterised and subdivided based on their change in expression compared to controls (down- or up- regulated) (***Figure 5.6***.a and b). The analysis was further extended by looking at ATRX binding sites in proximity to the DEG using different windows up to 100kb away from each DEG (***Figure 5.6***. a and b). Therefore, DEG were classified based on a priority order as: (1) displaying at least one intragenic ATRX binding site, (2) displaying at least one ATRX binding site in proximity up to 10 kb away from the DEG and (3) displaying at least one ATRX binding site in proximity between 10kb and 100 kb away from the DEG (***Figure 5.6***.b). The DEG that did not display at least one ATRX binding site under these conditions were considered as potential indirect ATRX targets. A manual verification of the selected DEG revealed some inaccuracies regarding the annotation of the gene location in the Affymetrix dataset associated with some probes. As an example, the probe 11738632_s_at, associated with the *GPAT2 gene*, was incorrectly associated with a region of more than 1,000 kb including many genes (***Figure 5.7***). After correction for the gene location, 58 DEG out of 103 were identified as potentially directly affected by ATRX mutations (***Figure 5.6***. a and b) with approximately 59% upregulated (35 DEG) and 41% downregulated (24 DEG). Even though the upregulated DEG remain in the majority, their share is slightly reduced compared to their ~ 64% share in the original 103 DEG (***Figure 5.3***.c). Strikingly, the distribution of the ATRX binding sites corresponding to the selected DEG differed between the down- and up-

regulated DEG. Indeed, the distribution of the ATRX targets seemed to be skewed toward intragenic regions for the 24 downregulated DEG. By contrast, the 34 upregulated DEG showed similar enrichment at intragenic and distal regions (between 10 to 100 kb form the DEG) (***Figure 5.6***. b). To test the robustness of these observations, the analysis was repeated using ATRX peak calling of increasing stringency. Even with more stringent ATRX peak calls, the distribution of the ATRX targets remained skewed toward intragenic regions for the downregulated DEG (***Figure 5.6***. d and f). By contrast, for the upregulated DEG, ATRX enrichment shifted from the intragenic regions to distal regions (10 - 100kb) (***Figure 5.6***. d and f). Moreover, the more stringent the ATRX peak call was, the higher the proportion of downregulated DEG was (representing ~ 41% to 52%) (***Figure 5.6***). These data provided the first insight into a potential relationship between the nature of the change in gene expression and the position of the ATRX binding sites relative to the DEG.

The 58 DEG displaying at least one ATRX binding site within the 100kb window were submitted to a gene ontology analysis (Da Wei Huang et al., 2008) and showed proportionally stronger enrichments for transcription processes (24.5%) than the full list of candidates (14%) (***Table 5.1***). This enrichment increased to 40% when focusing on the 31 DEG displaying at least one intragenic ATRX binding site. Furthermore, as the distribution of the ATRX targets differed between down- and up- regulated DEG, each group was submitted to the gene ontology analysis independently. Limiting the analysis to the downregulated DEG further selected for transcription processes with more than 50% enrichment (***Table 5.1***). Conversely, gene ontology analysis based on upregulated DEG showed less than 10% enrichment for the top biological process, negative regulation of gene expression.

Taken together, these data suggest different processes acting on the regulation of the DEG that are down- and up- regulated in ATR-X patients. It has to be considered that by affecting the expression of genes involved in transcription, ATRX mutations may lead to significant secondary effects.

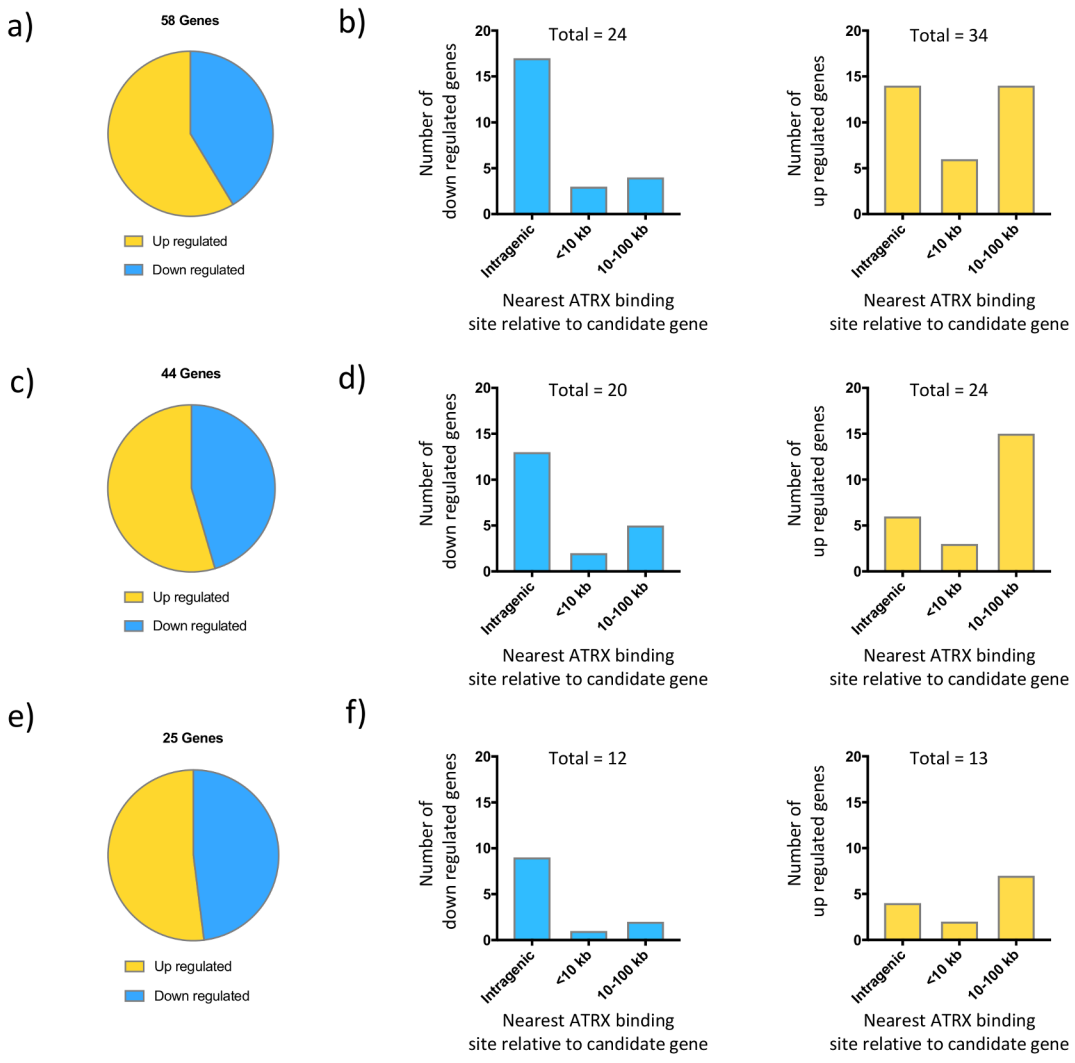


Figure 5.6: Identification of DEG potentially directly affected by ATRX mutations in LCL. a) and b) using the general ATRX peak call dataset described in Chapter 3 containing ATRX binding sites independent of the antibody used, individual background and present in at least 3 replicates. c) and d) using the more stringent ATRX peak call dataset for which ATRX binding sites were independent of the antibody used, individual background and present in at least 4 replicates. e) and f) using the highly stringent ATRX peak call dataset for which ATRX binding sites were independent of the antibody used, individual background and present in at least 5 replicates.

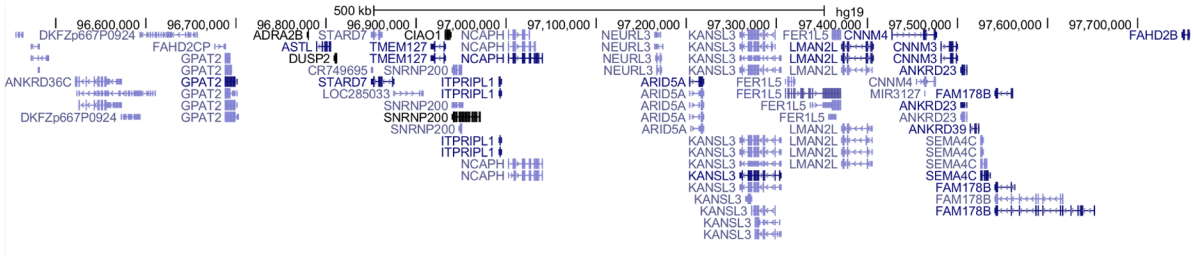


Figure 5.7: Example of inaccurate annotation of the gene position in the Affymetrix probe 11738632_s_at.

5.2.5. Validation of the microarray

After carrying out preliminary tests by standard qPCR to validate the cDNA, test most of the Taqman probes, and perform some pre-validation analysis, I subsequently validated the microarray data using the Fluidigm BioMark HD high throughput qRT-PCR system. Using this system, it was possible to analyse forty (out of forty-eight) of the samples used in the microarray (batches 1, 2, 3 and 5, see ***Table 2.1***) in two biological and two technical replicates and a relevant number of no reverse transcription controls using two house-keeping genes (*GAPDH* and *RPL13A*). In addition to the DEG tested, non-DEG (*MYC*, *TAB2* and *ZNF554*) were included in the gene-test panel as negative controls (***Figure 5.8***). The non-DEG (based on the microarray data) were also unchanged in expression using the Fluidigm platform (***Figure 5.8***). Furthermore, all but one (*MYOM2)* of the microarray-based DEG tested, were significantly differentially expressed between controls and cases (***Figure 5.8***). The microarray data showed that *ZNF595*, *MEIS2* and *ATF7IP2* were amongst the most downregulated genes whereas *UST, GDA* and *CSTA* showed significant over-expression in patients-derived LCL (***Figure 5.8***). Whereas *ZNF555*, *ZNF718*, *ZNF595*, *ZNF57* and *ZNF506* were all statistically significantly downregulated, *ZNF554*, a non-DEG located upstream of

ZNF555, was not significantly affected by ATRX mutations (***Figure 5.8***). Overall, the data obtained with the Biomark supported the microarray data.

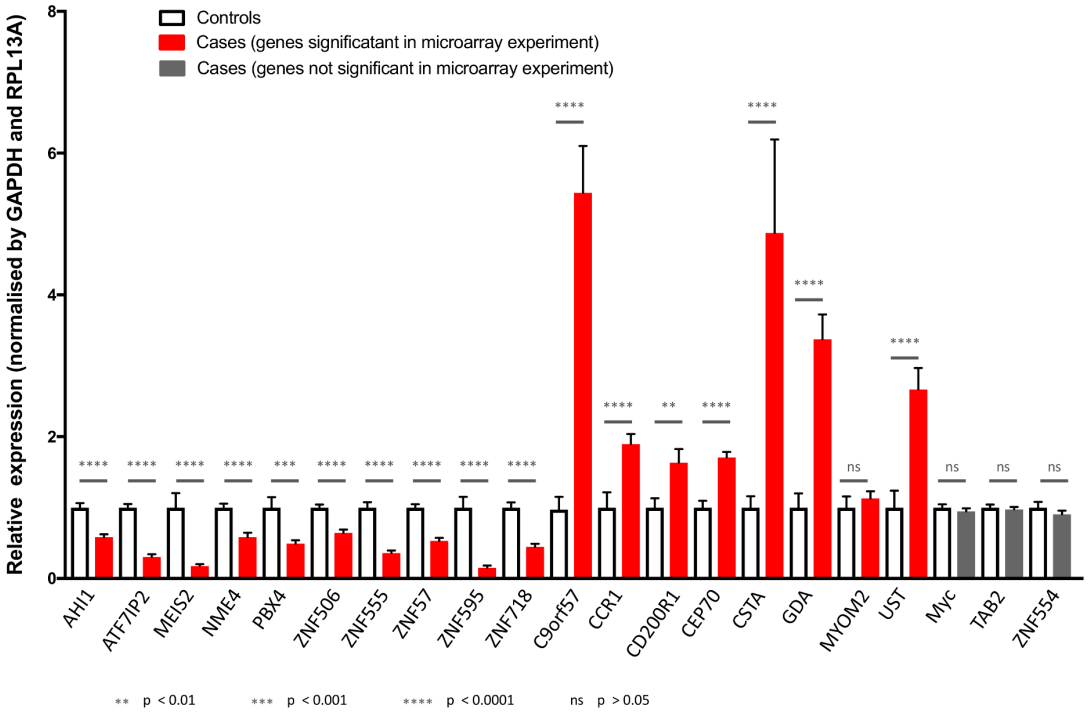


Figure 5.8: Validation of the microarray experiment using 40 LCL samples derived from different individuals (17 controls and 23 cases) using two biological and technical replicates. For display purposes, the normalised samples were divided by the mean of the normalised controls. Mean representation; error bars represent SEM. p-value determined by Mann-Whitney test.

5.2.6. Downregulated ZNF genes with intragenic ATRX binding sites are enriched in members of the KRAB-containing C2H2-ZNF subfamily

Many of the DEG were ZNF genes but not all ZNF genes were DEG. To further characterise the differentially expressed ZNF genes that were potentially directly affected by ATRX mutations, I analysed their genomic context, amino acid sequence and the ATRX distribution. Interestingly, when ATRX was mutated, only one ZNF with |logFC| > 0.5, *RNFT2*, was upregulated and all the ZNF genes that were

downregulated displayed intragenic ATRX binding sites (***Table 5.2*** and ***Figure 5.6***). Moreover, with the exception of *RNFT2 and L3MBTL4*, all the ZNF genes belonged to the Zinc finger C2H2 (C2H2-ZNF) (IPR013087) superfamily (***Table 5.2***, ***Figure 5.9*** and ***Figure 5.10***). Each C2H2 domain is composed of various numbers of 28 aa-repeats containing conserved cysteine and histidine residues. Within these DEG, the number of repeats varied from 10 (*ZNF506*) to 18 (*ZNF595* and *ZNF30*) (***Table 5.2***). No correlation between the numbers of C2H2 repeats and the severity of the dysregulation was observed. Indeed, *ZNF595* and *ZNF30* both contained the same number of repeats, however, *ZNF595* had the largest logFC (< -1.5) and *ZNF30* was barely above the |logFC| > 0.5 threshold (***Table 5.2***).

With the exception of ZBTB38, all the C2H2-ZNF proteins possessed a Kruppel-associated box or KRAB (IPR001909) domain (***Figure 5.9***). Each KRAB domain of these DEG was composed of approximately 40 aa corresponding to the KRAB-A box followed by a more polymorphic subdomain sometimes termed the KRAB-B, KRAB-b or KRAB-C box (***Figure 5.9***. a). Strikingly, ATRX enrichment differed greatly between the KRAB-containing C2H2-ZNF and the others. Whereas, L3MBTL4 and ZBTB38 displayed few and distinct ATRX binding sites within their gene bodies (***Figure 5.10***), the KRAB-containing C2H2-ZNF had a strong ATRX enrichment within their promoter regions and high levels of ATRX across their gene bodies (***Figure 5.11***). Yet, ZNF554, used as negative control in the microarray validation (***Figure 5.8***), also belonged to the KRAB-containing C2H2-ZNF superfamily with eight C2H2 repeats. Even though not differentially expressed, it also displayed intragenic ATRX enrichment.

Gene Symbol	LogFC	p-value	KRAB domain	Number of ZF-C2H2 repeats	Location of nearest ATRX binding site relative to the DEG
ZNF595	-1.730	0.00024	Yes	18	Intragenic
ZBTB38	-0.853	0.03096	No	10	Intragenic
ZNF718	-0.782	0.00011	Yes	11	Intragenic
ZNF57	-0.677	0.00465	Yes	14	Intragenic
ZNF506	-0.659	0.04555	Yes	10	Intragenic
ZNF555	-0.634	0.00007	Yes	15	Intragenic
ZNF30	-0.531	0.03541	Yes	18	Intragenic
ZNF626	-0.516	0.01056	Yes	12	Intragenic

Table 5.2: Characteristics of the differentially expressed ZNF genes belonging to the C2H2-ZNF superfamily

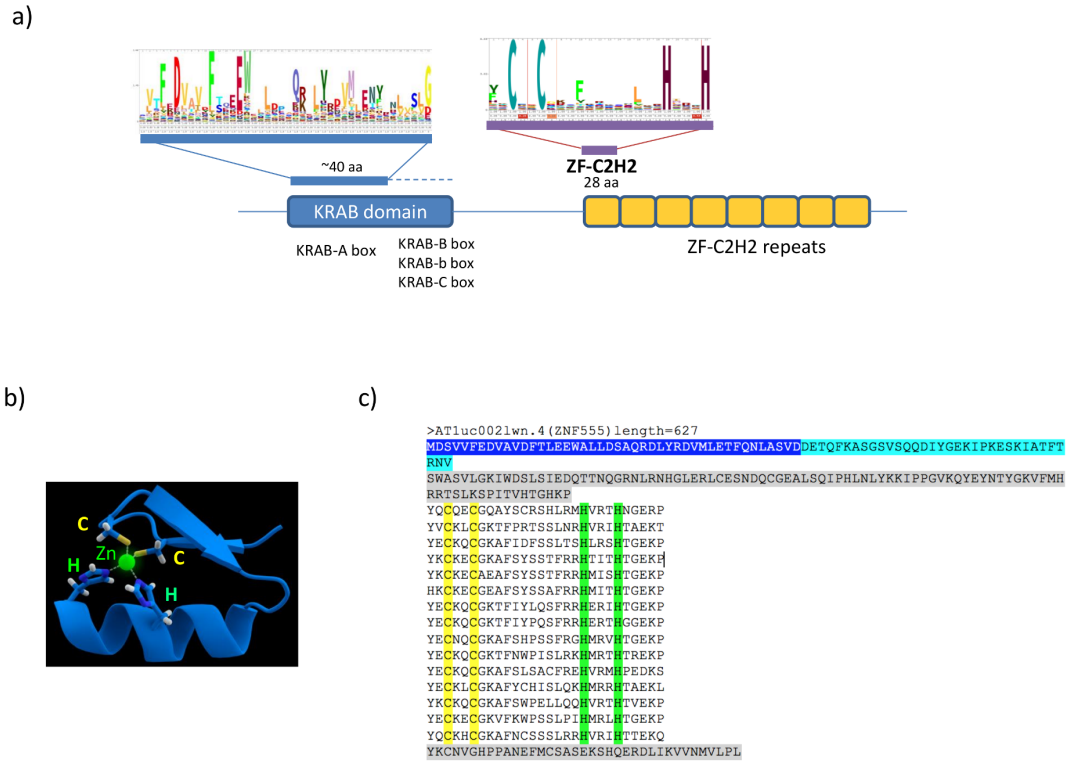


Figure 5.9: Characterisation of the differentially expressed ZNF genes. a. Cartoon showing the basic structure common to the tested ZNF proteins including a KRAB domain and Zn-C2H2 repeats associated with their Pfam HMM profiles (KRAB domain: PF01352 and ZF-C2H2: PF00096). b. Cartoon representing the ZF-C2H2 motif based on the X-ray structure of the protein Zif268 (PDB:1A1L) showing the presence of a Zinc ion interacting with the two highly conserved Cysteines and Histidines. c. Amino acid sequence of the ZNF555. KRAB domain is highlighted in blue, linker sequences were highlighted in grey, the two Cysteines and Histidines of each ZF-C2H2 were highlighted in yellow and in green, respectively.

Gene Symbol	logFC	p-val	Zinc finger type	Other domains	Location of nearest ATRX binding site relative to the DEG
L3MBTL4	-0.856	0.01043	C2HC-type	MBT repeats and SAM	Intragenic
RNFT2	0.798	7.93E-08	C3HC4	Transmembrane	within 10-100kb

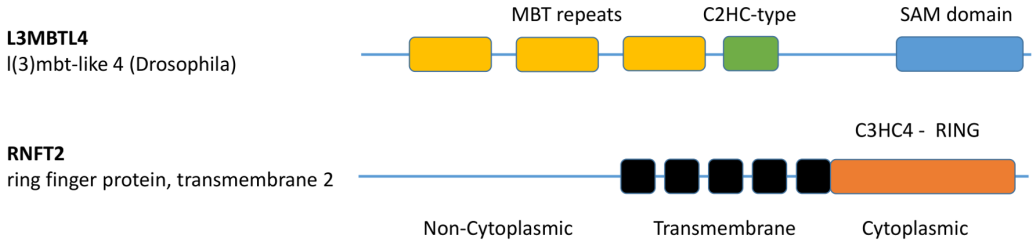


Figure 5.10: Differentially expressed ZNF genes not belonging to the C2H2-ZNF superfamily

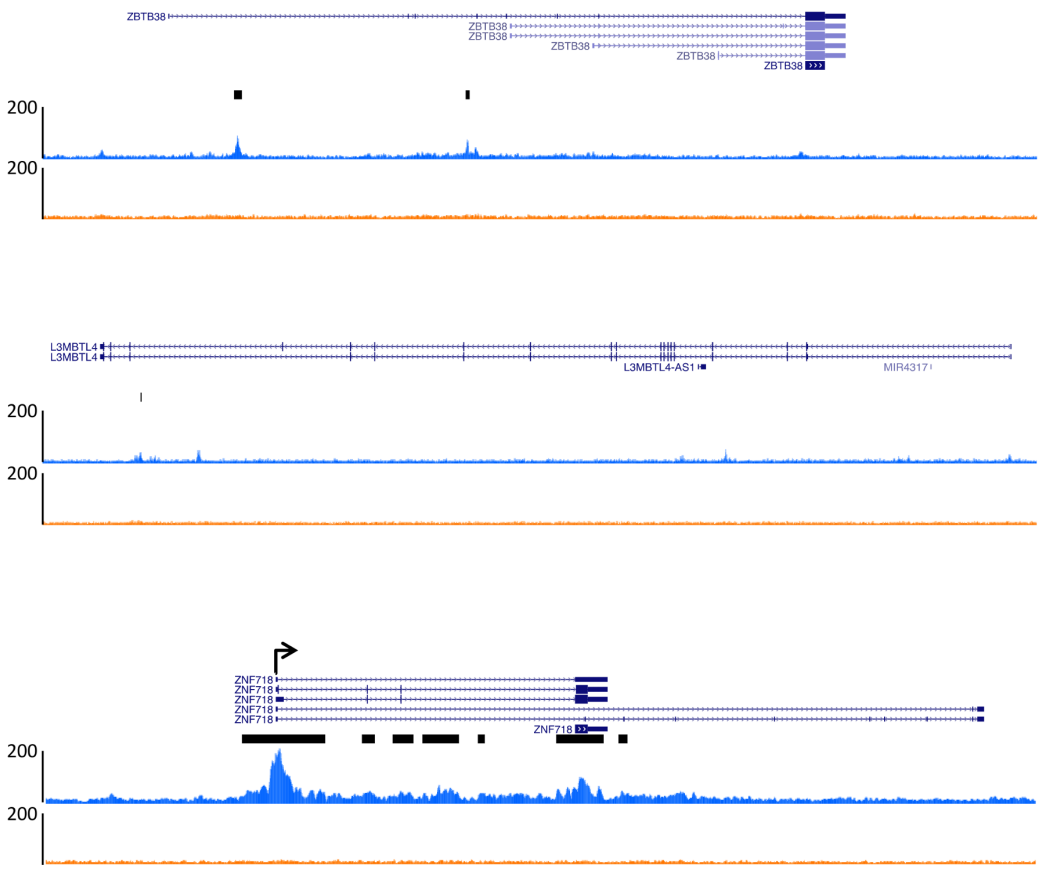


Figure 5.11: Variation in ATRX enrichment between differentially expressed ZNF containing, or not, KRAB and C2H2 domains. Top: ZBTB38, middle: L3MBTL4 and bottom: ZNF718. ATRX ChIP (blue), input sample (orange).

5.3. Discussion

Using a microarray approach, I have identified 234 DEG affected by ATRX mutations among which 103 showed a |logFC| > 0.5. These data showed that, at least in LCL, pathogenic ATRX mutations affect a limited number of genes and mostly lead to subtle changes in gene expression with only two DEG (*RASGEF1A* and *LPAR6*) having a |logFC| > 2. These results are in line with the low number of genes affected by ATRX mutations that have been identified and characterised over the last two decades. This restricted list of candidate genes is potentially linked with the essential role of ATRX. Indeed, ATRX inactivation in mouse results in defective formation of the extraembryonic trophoblast and embryonic death by day 9.5 post coitus (Garrick et al., 2006). Besides, analyses in ATR-X patients revealed that, even with a nonsense mutation such as p.S576X, patients were hypomorphic for ATRX instead of having a complete KO. Therefore, using LCL derived from ATR-X patients, it is reasonable to assume that the effects of disease-associated mutations are not as drastic as a complete ATRX KO. However, patient-derived cell lines may more accurately reflect the subtle perturbations leading to the disease *in vivo*. Additionally, there may be pathway redundancy associated with other chromatin remodeler complexes, such as HIRA, which may mask the full effects of ATRX dysfunction (Schneiderman et al., 2012).

I assessed the presence of any correlation between the type or location of the ATRX mutation and the change in gene expression. I did not observe any statistically significant differences between patients with the common p.R246C mutations and those either located throughout the ATRX protein or restricted to the ADD domain. Similarly, no significant differences were detected between the mutations located in the GATA-like and the PHD-like subdomains of the ADD domain. These observations are consistent with the hypothesis that mutations lead to loss of ATRX function as opposed

to a gain of function (haploinsufficiency is not considered, as *ATRX* is located on the X chromosome: the patients are males and therefore hemizygotes). However, increasing the specificity of comparisons involved smaller patient groups and led to a decrease of power. Thus, it could not be definitely ruled out, even unlikely based on the very high p-values obtained, that a larger panel would not reveal some differences. No statistically significant differences had been noticed between mutations affecting the ADD and the ATPase domains with one exception (NME4). It is therefore open to question whether the mechanism by which ATRX regulates NME4 expression is an exception rather than a rule or if *NME4* has unique features that result in this exception in LCL.

To identify the DEG directly affected by ATRX mutations, I analysed the location of ATRX binding sites relative to the DEG with a |logFC| > 0.5 and I characterised four subgroups: (1) DEG having at least one intragenic ATRX target, (2) DEG displaying at least one ATRX binding site in close proximity (less than 10 kb), (3) DEG having at least one ATRX target between 10kb and 100 kb away and (4) DEG without any ATRX enrichment within 100 kb. By doing so, I hypothesised that the DEG from subgroup 1 with at least one intragenic ATRX target have the greatest chance to be directly regulated by ATRX whereas the DEG in subgroup 4 are more likely to be indirectly influenced by ATRX. Any DEG from subgroups 2 and 3 could be regulated by ATRX through a proximal/distal mechanism as observed within the alpha globin cluster (Law et al., 2010). Based on the data described in Chapter 4, one must also consider whether the regulation of the DEG could occur via the action of ATRX at enhancers.

Additionally, many DEG were associated with transcriptional processes and some were indeed known or putative transcription factors. Thus, primary effects of ATRX mutations on the expression of genes involved in transcription, may lead to significant

secondary effects and it may prove challenging to untangle the direct and indirect roles of ATRX on the expression of DEG. Nevertheless, the pattern of ATRX enrichment differs between the up- and down- regulated DEG. Indeed, the downregulated DEG are predominantly enriched in intragenic ATRX binding sites, independently of the stringency of the ATRX peak call used. Furthermore, this dataset encompassed most of the DEG associated with the transcription processes including *ZNF* genes containing a KRAB domain (Lupo et al., 2013, Urrutia, 2003). This is the case for *ZNF555*, *ZNF57*, *ZNF595*, *ZNF718* and *ZNF506* that were further validated by the Fluidigm Biomark system. Identified in tetrapods and some archaic fishes but not in Zebrafish, Drosophila and yeasts, the presence of the KRAB domain in ZNF proteins has often been associated with repressor activity (Urrutia, 2003, Imbeault et al., 2017). Thus, it is conceivable that some DEG identified in subgroups 2 and 3 are indirectly affected by ATRX mutations via dysregulation of transcription factors and co-factors.

The list of downregulated DEG with intragenic ATRX binding sites also included the transcription factor interacting protein ATF7IP2 which interacts with Methyl-CpG Binding Domain Protein 1 (MBD1), the histone H3K9 methyltransferase SETDB1 and the transcription factor Sp1 (Ichimura et al., 2005). Interestingly, ATF7IP2 paralog, ATF7IP, along with MBD1 and SETDB1 have been implicated in the maintenance of X chromosome inactivation (Minkovsky et al., 2014) whereas PSM2 (protein similar to MCAF2), a protein structurally similar to ATF7IP2, has been associated with mouse gonad development (Cai et al., 2006); two processes affected by ATRX loss of function (Garrick et al., 2006, Lossi et al., 1999, Huyhn et al., 2011, Tang et al., 2004).
Additional downregulated DEG with intragenic ATRX binding sites were involved in transcription processes as the transcription factor PBX4 and the transcription regulator MEIS2. Strikingly, members of the Pbx and Meis families have been shown to associate

and may act as Hox cofactors. Thereby, in Zebrafish, PBX4 interacts with Meis3 and Hoxb1b forming a complex involved in hindbrain differentiation (Vlachakis et al., 2001) whereas Meis2 has been associated with midbrain development in chick (Agoston et al., 2012). Taken together, these data emphasise that beyond the enrichment in processes linked to transcription, selected DEG were also associated with developmental processes including brain development. This observation is further supported by the presence of *AHI1*. Mutations of this gene can result in midbrain-hindbrain malformation and cognitive impairment known as Joubert syndrome (Tuz et al., 2013). The role and misregulation of these DEG could shed light on the underlying causes leading to ATR-X syndrome and may deserve deeper analysis in future studies.

Despite the limitation of LCL as a model, they do display pathogenic ATRX mutations. Furthermore, the intrinsic role of ATRX on gene expression may not be drastically different across different cell types and, as long as appropriate target candidates are found, it may reveal information on the mechanisms by which ATRX mutations alter gene expression. In future experiments, it will be worth considering repeating the validation experiments on additional candidate genes if they appeared to be highly relevant later in the study (see Chapter 6) or on additional KRAB-containing C2H2-ZNF genes similar to the DEG but not significantly differentially expressed in the microarray experiments. Furthermore, I would like to use the Fluidigm BioMark platform to test some of the DEG identified in this study on RNA extracted from an ATR-X patient's blood as well as in sorted CD34 negative, CD19 positive B cells to assess the relevance of this analysis in primary cells.

Chapter 6. Investigation of the effects of ATRX mutations on chromatin landscape

6.1. Introduction

Studies of human disease have often led to the discovery of novel genes and the understanding of protein function. ATRX is no exception. The *ATRX* gene was discovered through studies focussing on the ATR-X syndrome (Gibbons et al., 1995). Over the years, animal models, cell lines and patients' samples have proved invaluable in understanding the role of ATRX (Clynes et al., 2013, Berube et al., 2005, Law et al., 2010, Voon et al., 2015). A very common approach has been to correlate the localisation of the proteins with DEG in absence of WT ATRX to further focus on the potential causes of these changes. Therefore, following the acquisition of reliable ATRX ChIP-seq data helping to identify the genome-wide binding of ATRX (Chapter 3), I analysed the chromatin environment at that sites (Chapter 4) and I selected potential direct DEG targets (Chapter 5). Now I propose to investigate the changes in the chromatin landscape associated with ATRX mutations. My results showing that ATRX is significantly associated with active regulatory elements, its affinity for KRAB-containing C2H2-ZNF genes and the observation that a subset of these genes (plus others) are differentially expressed provide new information for determining the precise role of ATRX.

Several mechanisms leading to the dysregulation of DEG may be considered. Indeed, work focusing on ATR-X syndrome suggested that ATRX might prevent and/or resolve the formation of G4 at VNTR sequences (Law et al., 2010). Nevertheless, many ATRX binding sites are not GC-rich and should not form G4. Therefore it is likely that our list of genes affected by ATRX mutations will contain subgroups where the mechanism of

perturbed gene regulation differs depending on their genomic and epigenetic contexts. I propose to study specifically these environments.

Although H3.3 deposition at actively transcribed genes is supposed to be HIRA-dependent in mouse ES cells (Goldberg et al., 2010), the present analysis of ATRX distribution in human LCL revealed that ATRX is enriched at a subset of such sites. Therefore, it is of interest to investigate if some of these sites showed changes in the H3.3 occupancy in ATRX deficient cells. It would not be the first example of an association between ATRX and intragenic deposition of H3.3. Thus, in mouse, it was suggested that ATRX may facilitate the transcription of a subset of aPAR genes especially at PQS regions (Levy et al., 2014). However, these studies did not determine if these regions could be associated with regulatory elements. Similarly, in human cell lines, ATRX enrichment was been associated with the 3' exon of ZNF genes but no correlation was found with gene expression (Valle-García et al., 2016). However, the authors did not investigate in depth the DEG between ATRX KO and WT. Furthermore, their ATRX ChIP-seq data did not allow them to detect ATRX enrichment at active regulatory elements. Other work in mouse, showed an association between ATRX binding and changes in gene expression and chromatin environment at imprinted loci (Kernohan et al., 2014, Voon et al., 2015). It is therefore of particular interest to (1) investigate the role of ATRX at imprinted loci and aPAR genes in human, bearing in mind that LCL are differentiated cell lines and may not be the optimum model, (2) investigate the effects of ATRX deficiency at regulatory elements and (3) identify potential subgroups of ATRX targets showing similar patterns of expression. All together, these new results should help us to develop theories regarding ATRX function.

6.2. Results

6.2.1. ATRX at imprinted genes

In mouse, several lines of evidence have associated ATRX with the regulation of a subset of imprinted genes and differentially methylated regions (Kernohan et al., 2014, Voon et al., 2015). Therefore, I analysed the chromatin landscape of imprinted genes identified during the microarray experiment. Two established (*NDN* and *MKRN3*) and two predicted (*LILRB4 and FUCA1)* imprinted genes (Anisimov, 2012) were identified among the DEG with a |logFC| > 0.5 (***Table 6.1***). *NDN*, *LILRB4* and *MKRN3* were upregulated whereas *FUCA1* was downregulated in patient-derived LCL. No relationship was observed between the parental origin of the expressed alleles and the change in gene expression. Indeed, among the three upregulated DEG, two are expressed from the paternal allele (*NDN* and *MKRN3*) whereas the third one was expressed from the maternal allele (*LILRB4*). None of the imprinted DEG displayed an intragenic ATRX binding site. Indeed, the established imprinted genes *NDN* and *MKRN3* did not show any ATRX binding sites within a 100kb region flanking the genes (***Figure 6.1***). Both genes are located within the Prader-Willi syndrome/Angelman syndrome (PWS/AS) imprinted region (Abreu et al., 2015, Rabinovitz et al., 2012). Despite the absence of ATRX enrichment, some changes in active chromatin marks between controls and cases were noticed; I observed an increase of H3K4me3 at these gene promoters and an increase in H3K27ac signal at MKRN3 (***Figure 6.1***).

By contrast with the established imprinted genes *NDN* and *MKRN3*, ATRX binding sites were identified within the 100kb regions flanking the *LILRB4 and FUCA1* genes. However, no changes in chromatin landscape could be observed at these genes. I performed a Capture-C experiment using the *LILRB4* promoter as viewpoint to test for interactions between the ATRX binding site observed upstream *LILRB4* and *LILRB4*

promoter (***Figure 6.2***). No significant interaction was detected suggesting that ATRX may not have a direct effect on *LILRB4* in LCL.

Gene	Status	Expressed Allele	LogFC	p-value
FUCA1	Predicted	P	-0.699	0.0212
MKRN3	Established	P	0.579	0.0275
LILRB4	Predicted	M	0.669	0.00157
NDN	Established	P	0.962	0.0463

Table 6.1: List of imprinted genes identified as DEG with |logFC| > 0.5. P: paternal, M: maternal. Status based on (Luedi et al., 2007, Anisimov, 2012).

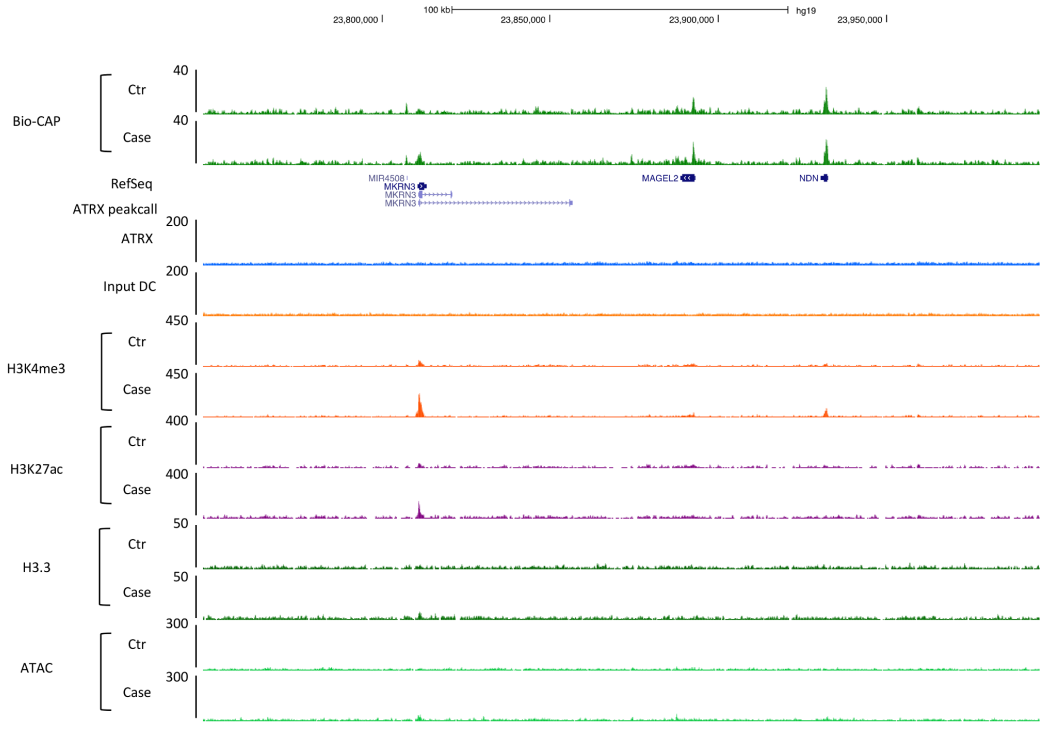


Figure 6.1: Example of imprinted genes (NDM and MKRN3) that are dysregulated in ATR-X patient derived LCL showing an absence of significant ATRX enrichment. All tracks represent the average of experiments performed in three biological replicates except for ATAC-seq that represents an average of four replicates and ATRX ChIP-seq that is an average of six replicates.

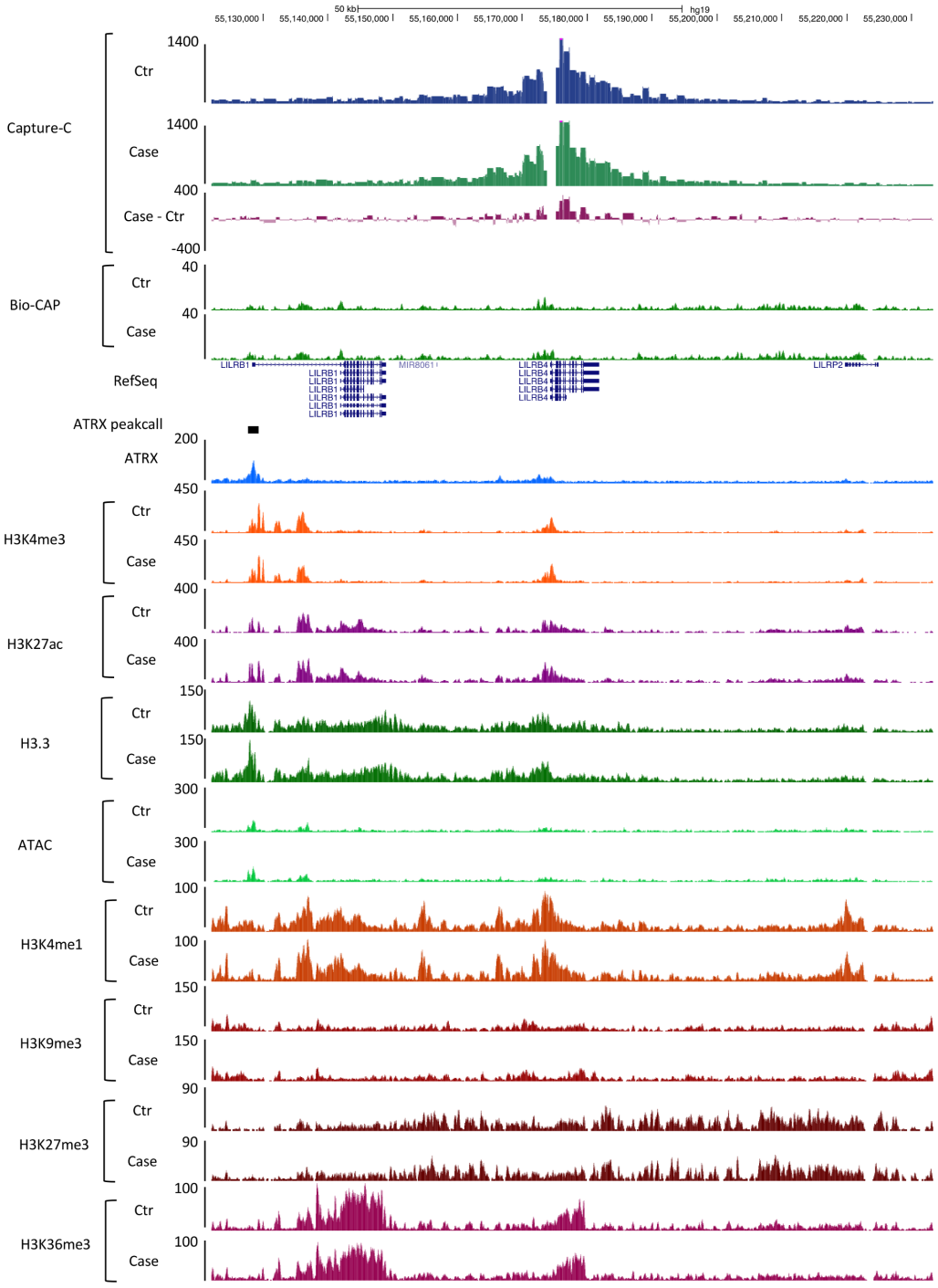


Figure 6.2: Analysis of ATRX localisation and chromatin interaction associated with the LILRB4 locus. All tracks represent the average of experiments performed in three biological replicates except for ATAC that represents an average of four replicates and ATRX that is an average of six replicates.

6.2.2. ATRX and aPAR gene expression

Previous studies in mouse have identified a subset of aPAR genes that were significantly downregulated in absence of ATRX (Levy et al., 2008, Levy et al., 2014). However, none of the aPAR homologs tested in LCL were found to be significantly differentially expressed in either microarray datasets or by RT-qPCR (***Figure 6.3***). Taken together, these data suggested that ATRX may not have a significant role in aPAR gene transcription in human LCL.

These differences between the results obtain in mouse and in human LCL suggest that the studies in mouse may not be relevant in a human context or at least in LCL. These discrepancies emphasise the importance to investigate human-derived cells and tissues to better understand the abnormalities associated with ATRX mutations leading to disease.

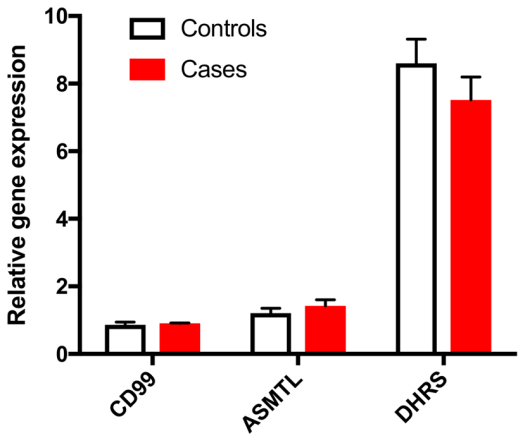


Figure 6.3: Expression analysis at a subset of aPAR genes in LCL using RT-qPCR. Error bar show the SEM. Experiment performed with three biological replicates.

6.2.3. ATRX at the NME4 Locus

The gene expression analysis described in the Chapter 5 confirmed the downregulation of NME4 in ATRX deficient cells as previously identified (Law et al., 2010). However, no significant ATRX peak was called at the NME4 VNTR in LCL. Interestingly, mapping issues were observed at this locus and could probably restrict the analysis regarding ATRX enrichment (especially at the VNTR region). Despite these limitations, changes in chromatin environment were detected at the locus and a clear decrease in active chromatin marks was observed in ATRX deficient cells (***Figure 6.4***).

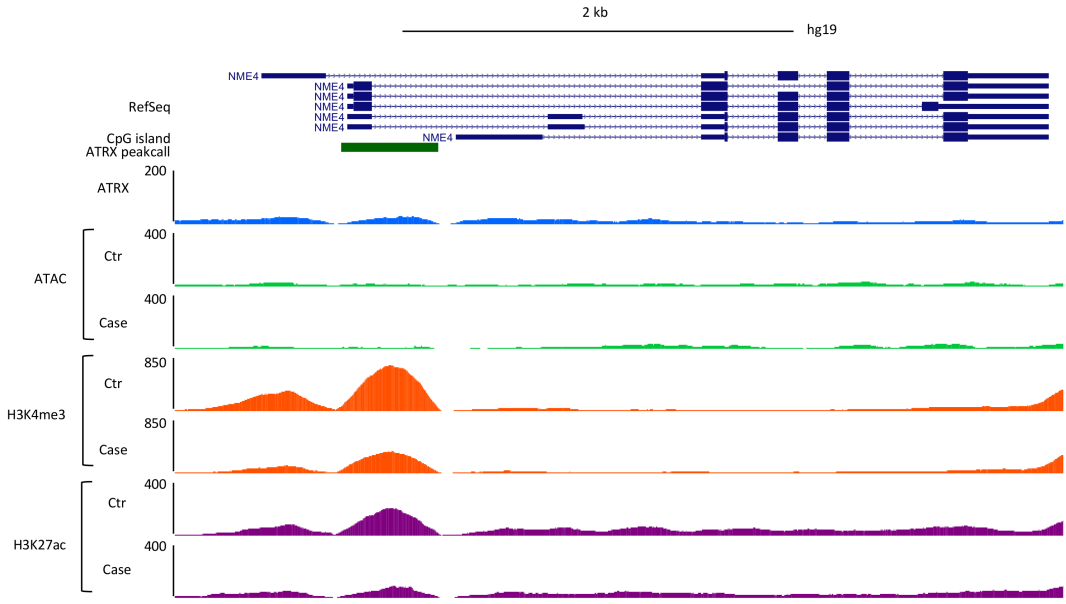


Figure 6.4: Changes in chromatin environment at the NME4 locus in LCL.

6.2.4. ATRX mutations correlate with changes in chromatin environment at regulatory elements associated with a subset of DEG

A large proportion of ATRX binding sites contained active enhancers and promoters. To assess the relevance of these observations in the context of ATR-X syndrome, I investigated whether the chromatin environment within the 58 DEG, short-listed in Chapter 5, differed between cases and controls. A subset of DEG showed changes in

chromatin environment at putative regulatory elements enriched in ATRX. It was, for instance, the case at the *CCR1* locus. *CCR1* was upregulated in patient-derived LCL. Several sites of ATRX binding were observed in the region flanking the *CCR1* gene (***Figure 6.5***). These sites displayed marks of active enhancers including nucleosome-depleted regions with H3K27ac and H3K4me1 enrichment in control LCL (***Figure 6.5***). Interestingly, ATR-X deficient cells showed some changes in chromatin environment at a subset of these sites including increases in chromatin accessibility and H3K27ac enrichment (***Figure 6.5***). I performed Capture-C experiments in the three Case-Control pairs (MF-D2015V/FF-Ctr, SB-R246C/CB-Ctr and RA-S576X/TA-Ctr) to determine if the *CCR1* promoter was interacting with these putative enhancers, which would suggest a potential regulatory association between these regions. The results highlighted a strong interaction between the *CCR1* promoter region and the downstream cluster of putative enhancers; both enriched in ATRX (highlighted in yellow in ***Figure 6.5***). An increase in chromatin accessibility was observed at these two regions in ATR-X cases. Similarly, an increase in H3K27ac signal at the cluster of putative enhancers in cases was noticed. No changes in repressive marks (H3K9me3 and H3K27me3) were identified between cases and controls at the ATRX binding sites within these regions. Nevertheless, a decrease in H3K27me3 in cases was detected downstream the putative enhancer cluster enriched in ATRX (highlighted by a grey box in ***Figure 6.5***). Interestingly, despite the presence of ATRX mutations, the H3.3 signal did not appear affected. No major changes in the 3 dimension chromatin interactions were detected by Capture-C between cases and controls. Nevertheless, a slight increase in interactions was observed between the enhancer cluster and the *CCR1* promoter in cases (***Figure 6.5***).

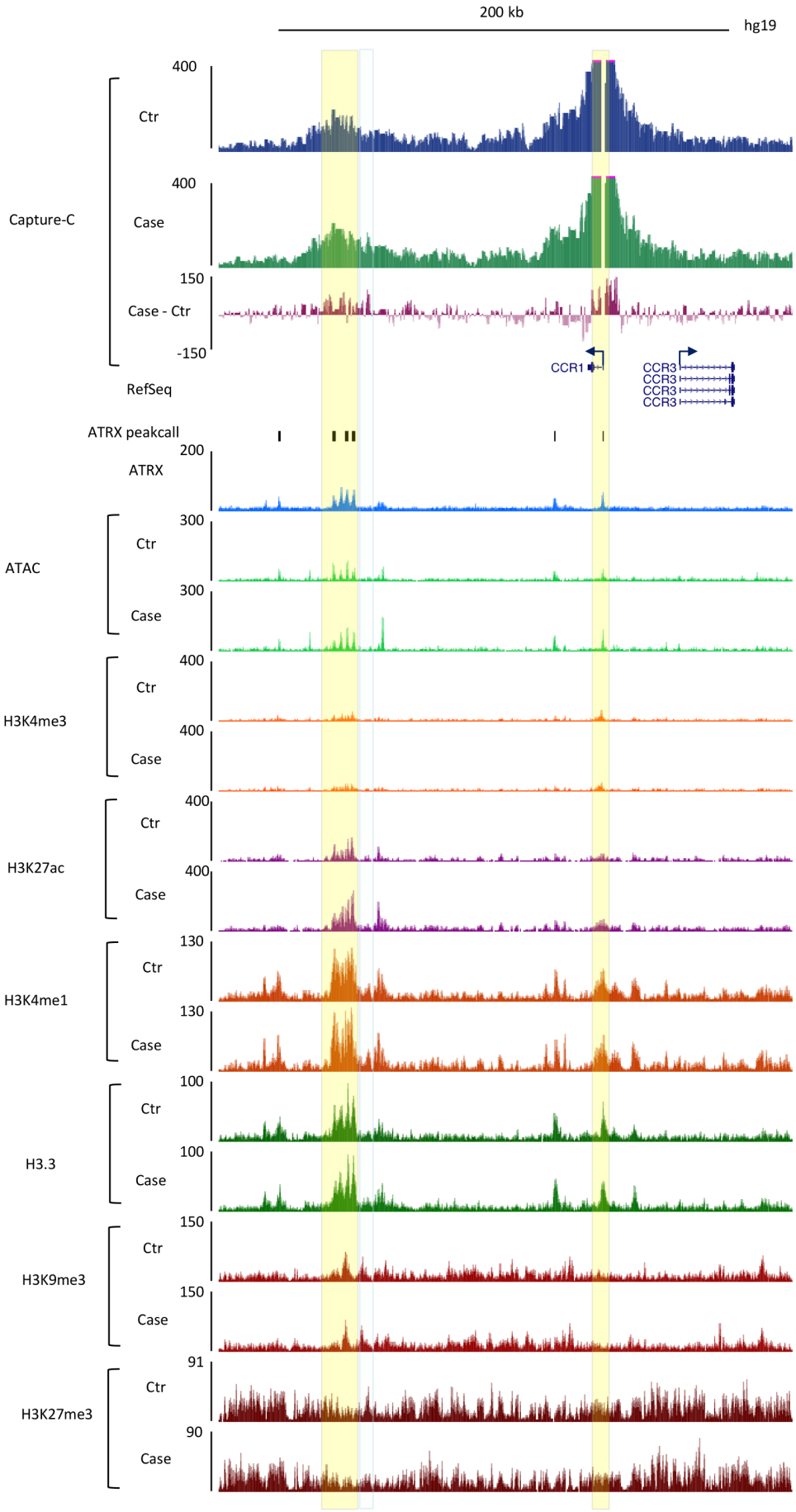

200 kb
hg19
Capture-C
Ctr
Case
Case - Ctr
400
150
-150
RefSeq
CCR1
CCR3
ATRX peakcall
ATRX
200
ATAC
300
H3K4me3
H3K27ac
H3K4me1
130
H3.3
100
H3K9me3
150
H3K27me3
91
90

Figure 6.5: Comparison of the chromatin environment at the CCR1 locus between LCL derived from unaffected controls and ATR-X patients. Due to the orientation of the CCR1 gene, the region located downstream this gene is displayed on the left side whereas the region upstream this gene is located on the right side of the picture. The regions enriched in ATRX that showed some changes in chromatin environment between controls and cases were highlighted in yellow. The additional light blue box showed the region associated with changes in H3K27me3.

The ATRX enrichments detected at the *CCR1* promoter and at the putative enhancer cluster were unlikely to be solely due to a "leak through" artefact associated with the strong cross-linking conditions used in ATRX ChIP-seq experiments. Indeed, as exemplified in ***Figure 6.6***, the strong interaction between the gene promoter and the CTCF binding sites near its 3' UTR did not result in apparent ATRX enrichment at the CTCF site but rather ATRX binding was restricted to the promoter region. This means that the strong cross-linking conditions of the optimised ATRX ChIP-seq are unlikely to produce artefactual ATRX enrichment at enhancers.

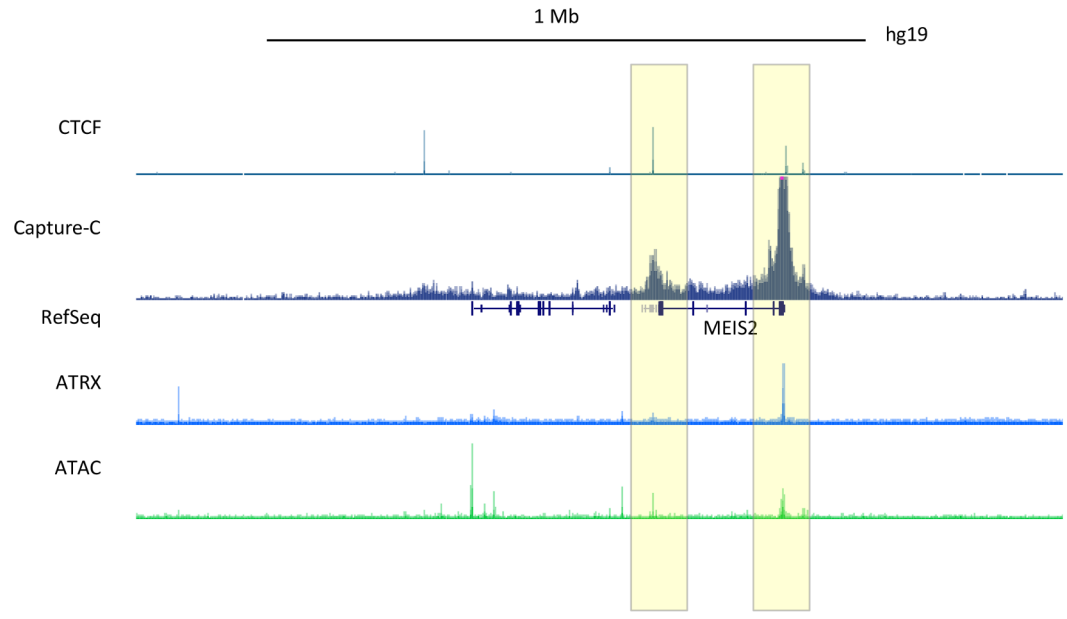


Figure 6.6: Capture-C experiment at the MEIS2 locus highlighting that the ATRX enrichment detected at the promoter region did not necessarily associate with ATRX enrichment at the promoter interacting regions

Interestingly, capture-C experiments showed that the *CCR1* promoter region was also strongly interacting with the *CCR5* locus (***Figure 6.7***). The reanalysis of the expression dataset obtained in Chapter 5 revealed that none of the CCR3, CCR2, CCR5 and CCRL2 genes reached the |logFC| >0.5 threshold. However, *CCR2* and *CCRL2* genes were observed to be differentially expressed (p-value < 0.05) in the microarray experiment and upregulated in ATR-X patient-derived LCL (*CCR2*: logFC = 0.46 and *CCRL2*: logFC = 0.23). These observations may suggest that ATRX deficiency directly affects the *CCR1*, *CCR2* and *CCRL2* genes. The regulatory elements enriched in ATRX may be involved not only in the regulation of *CCR1* but also of *CCR2* and *CCRL2* genes. It would be interesting to perform Capture-C experiments using the putative enhancer cluster enriched in ATRX as the viewpoint to assess its potential interaction with the *CCR2* and *CCRL2* genes.

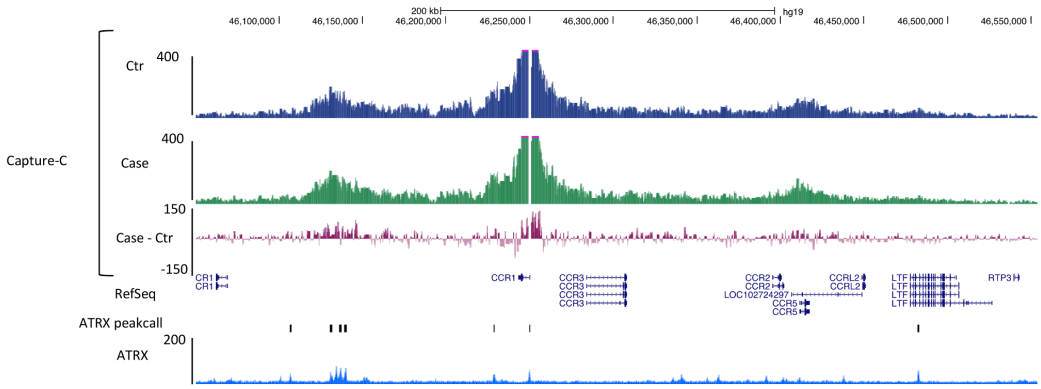


Figure 6.7: Capture-C experiments showing the interactions associated with the CCR1 promoter

The sulphotransferase *UST* (Uronyl 2-Sulfotransferase) is another upregulated DEG in ATR-X patient-derived LCL. Several ATRX binding sites were observed within *UST* gene body but not at its TSS (***Figure 6.8***). Two additional ATRX binding sites were

detected less than 100 kb upstream from the *UST* promoter. Both of the upstream intergenic ATRX binding sites showed slight enrichment in H3K9me3 but only one of them displayed an open chromatin region (***Figure 6.8***). In controls, this region was also enriched in H3K4me1 and H3.3 but not in H3K27ac. Interestingly, in the presence of mutant ATRX, this site showed H3K27ac enrichment, a slight increase in ATAC-seq and H3.3 enrichment (***Figure 6.8***). Similarly, the intragenic ATRX binding site proximal to the *UST* TSS displayed some marks of active enhancers (open chromatin region, enrichments in H3K27ac and H3K4me1). In cases, this site showed an increase in H3K27ac and subtle increase in open chromatin (***Figure 6.8***). These chromatin modifications are associated with enhancer activity. Three additional intragenic ATRX binding sites were observed at the 3' end of *UST* gene. Among those three, the one highlighted in yellow in ***Figure 6.8***, displayed chromatin marks characteristics of an active promoter with a high H3K4me3 signal at an open chromatin region. Furthermore, it overlapped with the TSS of the uncharacterised LOC105378047, which is affiliated to the non-coding RNA class based on the RefSeq annotation (highlighted in yellow in ***Figure 6.8***). By contrast with the two ATRX binding sites mentioned earlier, a decrease in active chromatin marks was observed in cases with reduction of the ATAC and H3K4me3 signals (***Figure 6.8***).

Similarly to CCR1, active sites enriched in H3.3 and ATRX did not showed a strong H3.3 loss in cases. Conversely, some sites may even display an increase in H3.3 signals under these conditions. Unexpectedly, the strongest change in chromatin accessibility was observed at a position without ATRX binding site or enrichment in H3K27ac and H3K4me3 (highlighted by a black box in ***Figure 6.8***).

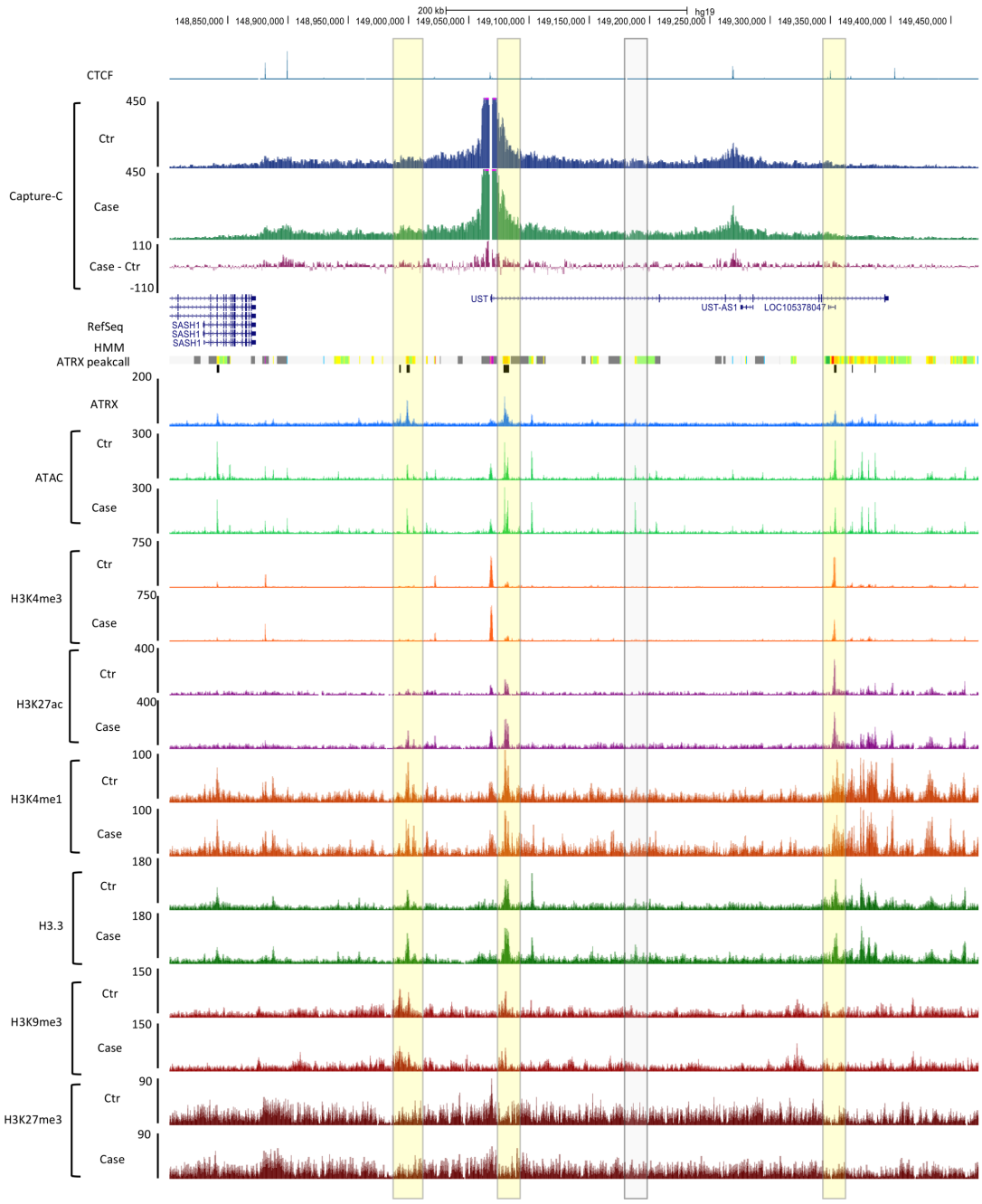


Figure 6.8: Characterisation of the chromatin environment at the UST locus in controls and cases. ChromHMM color code corresponds to strong enhancer (orange), active promoter (red), weak promoter (pink), weak/poised enhancer (yellow), inactive/poised promoter (purple), transcription related (green), insulator (blue) and repressed/heterochromatin/low signal (grey).

Adding to the complexity observed at the *UST* locus, an ATRX binding site was noticed at the non DEG *TAB2* promoter located downstream of the *UST* gene. I explored the possibility that some of the *UST* intragenic regulatory elements bound by ATRX might have subtle effects on *TAB2*. To ensure that the expression of *TAB2* was not affected by ATRX mutation, I validated microarray data related to *TAB2* during the Biomark experiment described in Chapter 5 and confirmed that *TAB2* expression was not significantly affected by ATRX deficiency. Then, I performed Capture-C using *UST* and *TAB2* promoters as viewpoints to investigate which regulatory elements interacted with these two promoters.

The *UST* promoter displayed strong interactions with its proximal intragenic and intergenic ATRX binding sites as well as with CTCF sites that are not associated with ATRX. These interactions were maintained in patient-derived cells (***Figure 6.9***). The three ATRX binding sites located near the 3'end of the *UST* gene had weaker interactions with the *UST* promoter. However, these three ATRX binding sites showed some interactions with *TAB2* promoter (***Figure 6.9***).

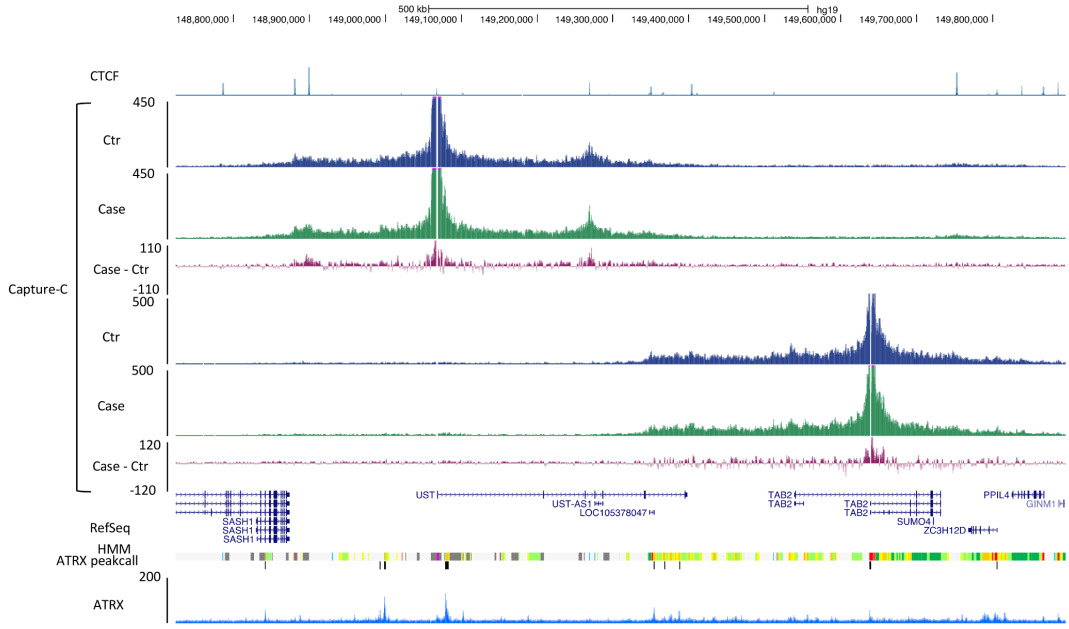


Figure 6.9: Complex chromatin interactions occurred across the UST locus.

Variations in chromatin landscape at putative regulatory elements associated with ATRX binding sites were not restricted to upregulated candidates. Indeed, the putative transcription factor PBX4 was downregulated in ATR-X patient-derived LCL. An ATRX binding site was identified within its gene body. In controls, this region displayed marks of active enhancers (H3K4me1, H3K27ac and open chromatin) and enrichment in H3.3 (***Figure 6.10***). In the presence of mutant *ATRX*, a decrease in active chromatin marks was observed and included a reduction in chromatin accessibility and in H3K27ac and H3K4me1 signals (***Figure 6.10***). Moreover, a decrease in H3.3 signal was also detected. Interactions between this putative enhancer and PBX4 promoter were observed by Capture-C and maintained in patient-derived cells (***Figure 6.10***).

All together these results suggest that ATRX plays a role in the maintenance of the chromatin integrity at putative enhancers associated with DEG.

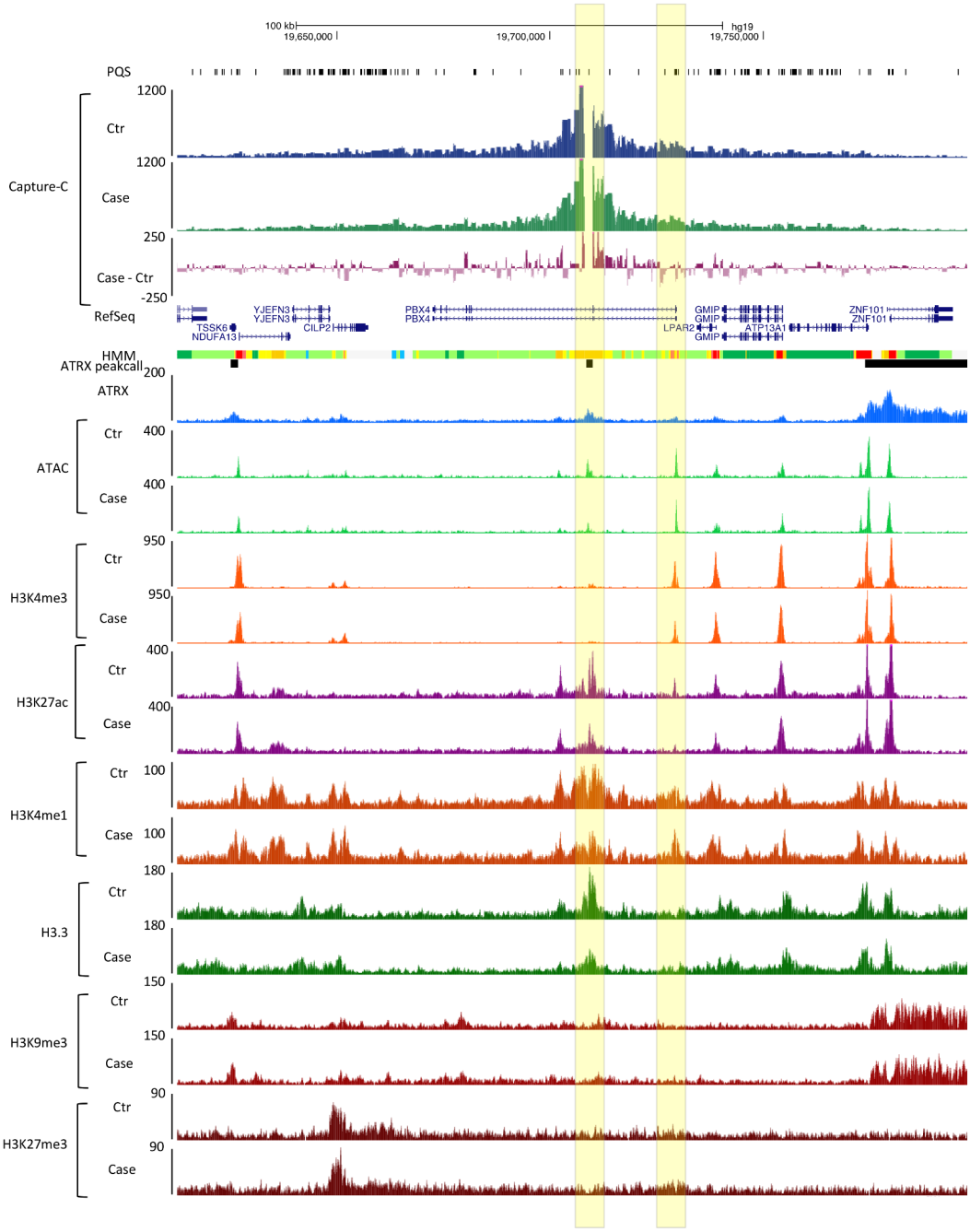


Figure 6.10: Chromatin environment at the PBX4 locus. ChromHMM color code corresponds to strong enhancer (orange), active promoter (red), weak enhancer (yellow), transcription related (green), insulator (blue) and repressed/heterochromatin/low signal (grey).

6.2.5. Chromatin interactions help discriminate between DEG that are affected directly or indirectly by ATRX

In the previous Chapter, I used an arbitrary threshold of 100kb regions surrounding the DEG to identify the presence of ATRX binding sites and so to select DEG that were potentially directly affected by ATRX. However, the more distal the ATRX binding site was, the less likely it was to play a direct role on the DEG expression. Using Capture-C experiments, I was able to confirm that some of these proximal ATRX binding sites could potentially affect the DEG. However, this analysis also revealed potential indirect ATRX effects. Indeed, one of the most downregulated DEG, the *BAIAP2L1* gene, displayed an ATRX binding site at less than 100 kb away from its 3' end. This ATRX binding site displayed the marks of a putative active enhancer (***Figure 6.11***). The *BAIAP2L1* promoter showed a decrease in active chromatin marks. The Capture-C analysis showed that in patient-derived cells, *BAIAP2L1* promoter (viewpoint) interactions with its gene body and around its 3' UTR were reduced. Despite these changes, no obvious interaction between *BAIAP2L1* promoter and surrounded ATRX binding sites were noticed, suggesting that *BAIAP2L1* was probably indirectly affected by ATRX mutations (***Figure 6.11***).

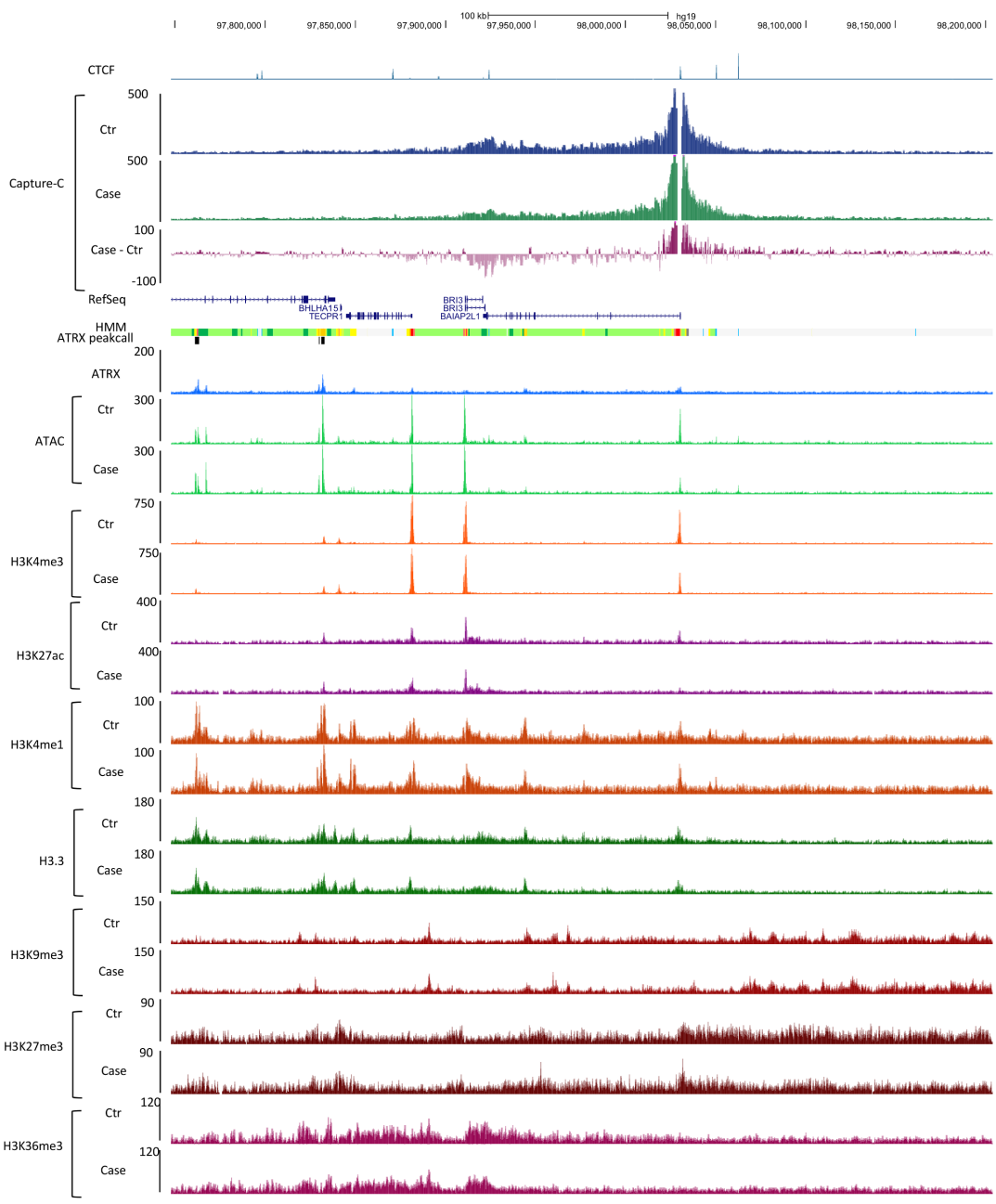


Figure 6.11: Chromatin environment at the BAIAP2L1 locus. ChromHMM color code corresponds to strong enhancer (orange), active promoter (red), weak promoter (pink), weak/poised enhancer (yellow), inactive/poised promoter (purple), transcription related (green), insulator (blue) and repressed/heterochromatin/low signal (grey).

6.2.6. ATRX mutations are associated with changes in chromatin accessibility

I further investigated the effects of ATRX deficiency on the genome-wide chromatin accessibility. Using Diffbind, I identified 907 sites that showed significant changes in chromatin accessibility between controls and cases (***Figure 6.12***). However, very few of them overlapped with ATRX binding sites or direct candidate genes (less than 60). This approach did not have a very high power due to the variability observed between the datasets. Indeed, the use of several case-control couples leads to greater variability than when using the same cellular background, which is preferred for DiffBind analyses. Therefore, a higher number of replicates would allow a higher power for the analysis. Nevertheless, this approach emphasised the presence of ATRX binding sites at promoters of DEG that showed decreases in expression and chromatin accessibility in patients.

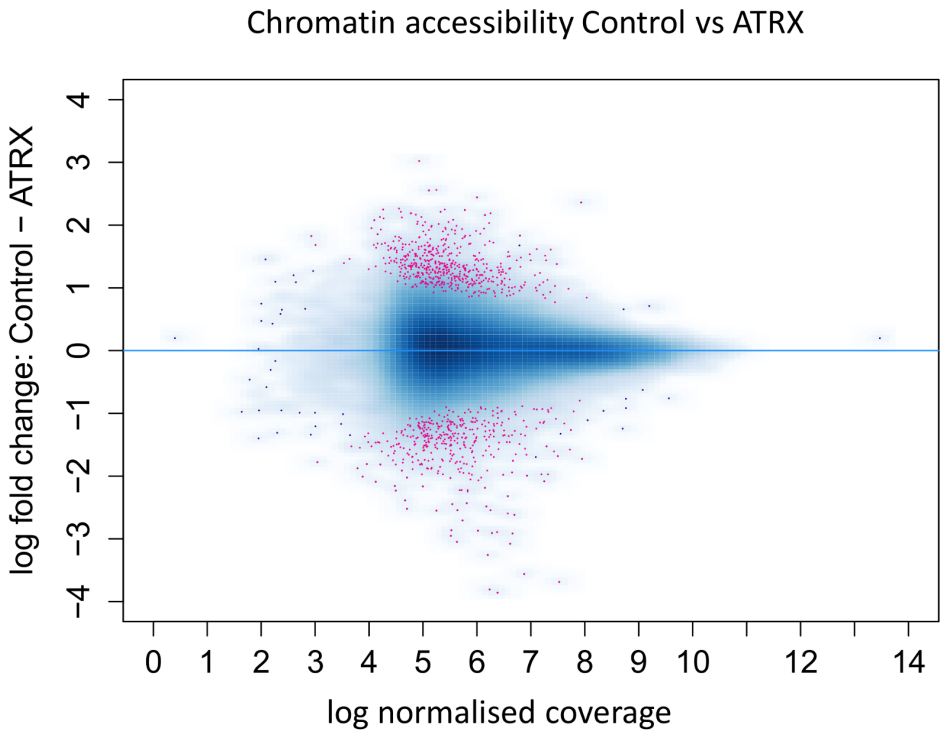


Figure 6.12: Diffbind analysis assessing for the difference in chromatin accessibility genome-wide between cases and controls. Significant sites (FDR < 0.1) were highlighted in pink.

ATF7IP2, a downregulated DEG, had an ATRX binding site at its promoter and showed a statistically significant decrease (FDR < 0.1) in chromatin accessibility in ATR-X patient-derived LCL (***Figure 6.13***). Interestingly, chromatin marks associated with active regions (H3K4me3 and H3K27ac) were also reduced. Even the H3.3 occupancy appeared to be depleted and the H3K36me3 enrichment across the gene was reduced, which are consistent with a decrease in transcriptional activity (***Figure 6.13***). Although *ATF7IP2* showed some H3K9me3 occupancy, no obvious changes were observed in this chromatin mark between cases and controls. No enrichment in H3K27me3 was found regardless of the ATRX status. The ATRX occupancy at *ATF7IP2* promoter region spans a predicted CpG island and PQS. Variability observed between the three case-control pairs, however, increases the complexity of the analysis. Although the trend of change was similar between the pairs (e.g.: enrichment reduced in cases compared to controls), the fold change could vary greatly as illustrated by H3K4me3 for instance (last six tracks in ***Figure 6.13***). This variability tended to correlate with the severity of the dysregulation, which was background-dependent too. Therefore, additional confirmation by ChIP-qPCR would be highly valuable to strengthen the data interpretation and obtain p-values. Increasing the number of case-control pairs or the number of replicates from the same pairs might also be considered. However, using LCL derived from different individuals offered an informative panel of changes reflecting the variability observed in patients.

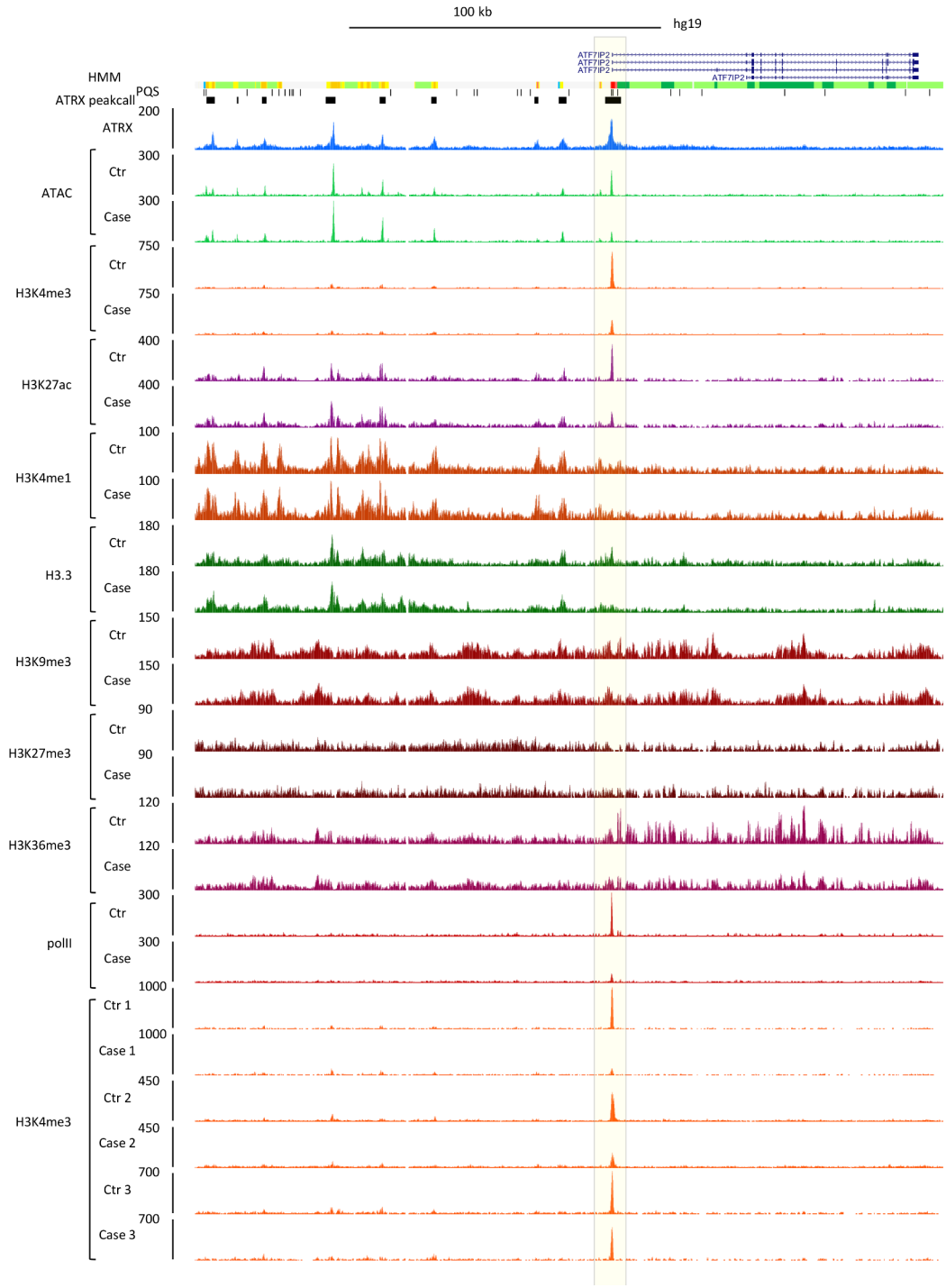


Figure 6.13:ATF7IP2 chromatin environment in LCL. ChromHMM color code corresponds to strong enhancer (orange), active promoter (red), weak promoter (pink), weak/poised enhancer (yellow), inactive/poised promoter (purple), transcription related (green), insulator (blue) and repressed/heterochromatin/low signal (grey).

Some of these results were further confirmed by ChIP-qPCR, demonstrating that enrichments in H3.3, RNApolII and H3K4me3 were statistically significantly reduced in cases at the *ATF7IP2* promoter whereas, H3K4me1 and H3K9me3 did not showed statistically significant changes (***Figure 6.14***). Importantly, an increased in H3 was noticed, supporting the reduced chromatin accessibility observed by ATAC-seq (***Figure 6.14***).

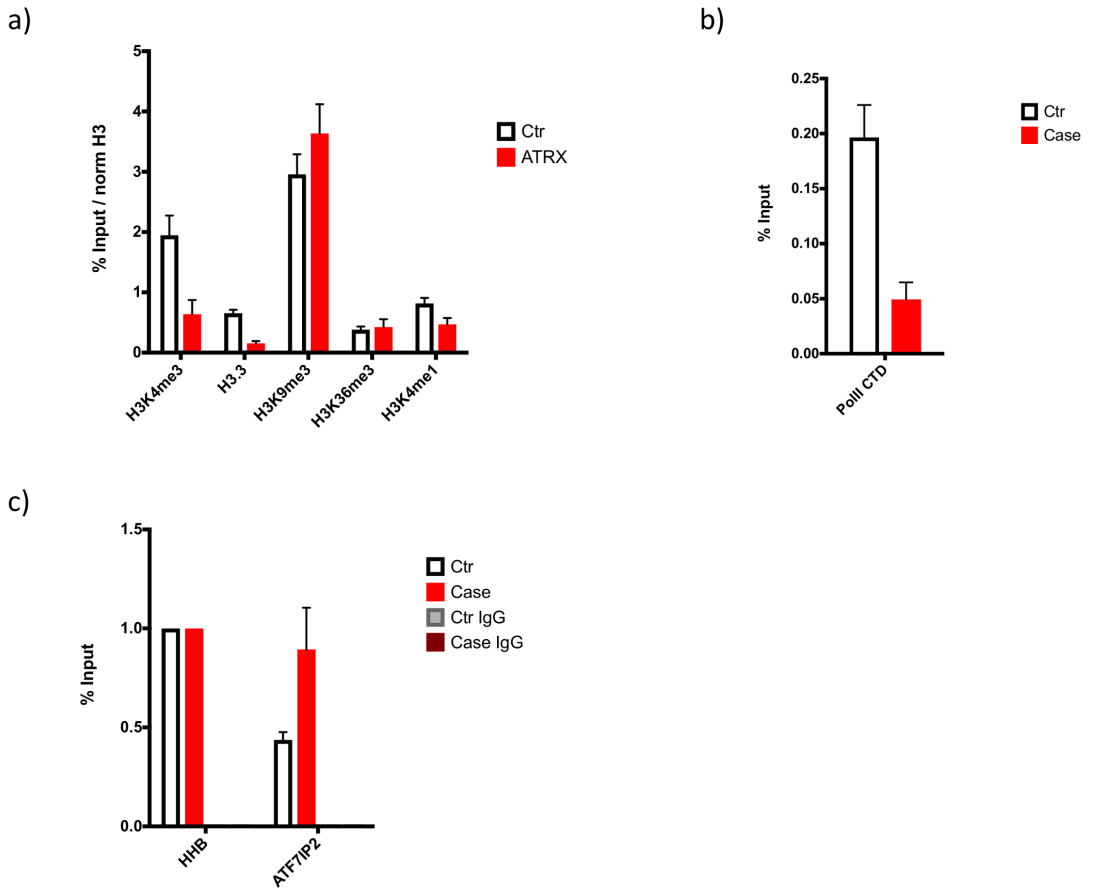


Figure 6.14: ChIP-qPCR analysing the changes in chromatin environment observed by ChIP-seq at the ATF7IP2 promoter. a. Chromatin marks normalised by H3. b. RNA polymerase. c. Histone H3. All the experiments were performed in three biological replicates except for H3.3 that was done in two biological replicates.

Another point of interest at the ATF7IP2 locus was the presence of upstream ATRX binding sites that co-localised with enhancer signatures containing open chromatin and H3K4me1 enrichment (***Figure 6.13***). Most of them also displayed H3K27ac enrichment associated with potential activity. The presence of ATRX binding sites at the *ATF7IP2* promoter region and upstream putative active enhancers raised the hypothesis that ATRX may be necessary to maintain the integrity of the chromatin environment at these sites and/or play a role in their interactions as previously observed in mice. The interaction between the promoter region and the putative enhancers was supported by Capture-C experiments showing strong interactions across the sub-TAD region in controls and cases (***Figure 6.15***). However, there were no significant differences between cases and controls in chromatin marks or interactions at these putative enhancers upstream of *ATF7IP2* suggesting they were not responsible for the change in gene expression (***Figure 6.15***).

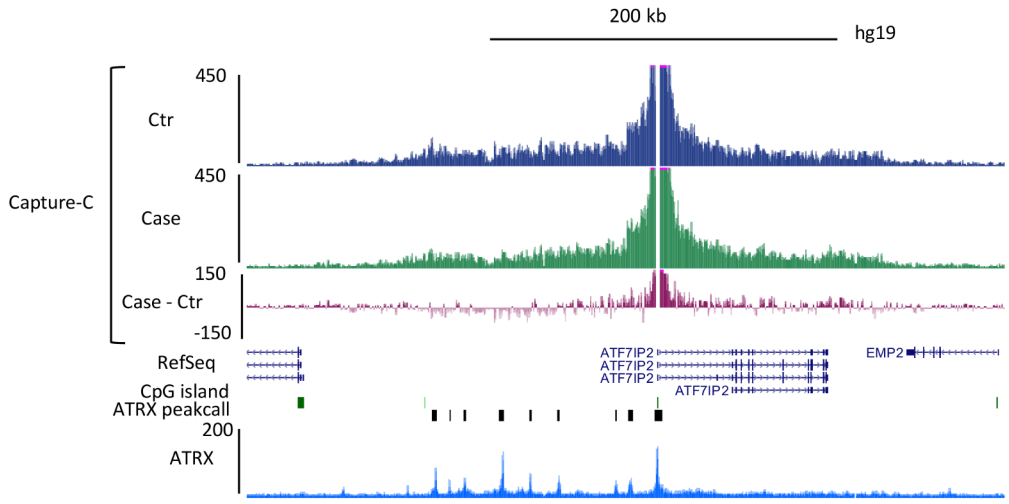


Figure 6.15: Chromatin interactions involving the ATF7IP2 promoter region.

Following these findings, I focussed on the downregulated DEG containing an ATRX binding site at their promoter region. All the downregulated ZNF genes, containing at least one intragenic ATRX binding sites, showed ATRX enrichment at their promoters. This was also the case for at least two other downregulated DEG, *MEIS2* and *AHI1* (***Table 6.2***). Interestingly most of these promoters shared similarities including the presence of predicted CpG islands and PQS. Moreover, for those without a predicted CpG island, Bio-CAP-seq showed the presence of unmethylated CpG clusters as illustrated with the *ZNF506* gene (***Table 6.2***).

Gene Symbol	p-value	logFC	ATRX binding sites spanning promoter region			
			Predicted CpG island	Non-methylated CpG cluster	PQS based on Sahakyan et al. 2017)	PQS based on QGRS Mapper
ZNF595	0,00024	-1,73	yes	?	yes	yes
MEIS2	0,00555	-1,40	yes	yes	yes	yes
ATF7IP2	3,19E-06	-1,09	yes	yes	yes	yes
ZNF718	0,00011	-0,78	yes	yes	yes	yes
AHI1	0,00810	-0,69	yes	yes	yes	yes
ZNF57	0,00465	-0,68	yes	yes	yes	yes
ZNF506	0,0456	-0,66	no	yes	no	yes
ZNF555	6,96E-05	-0,63	yes	yes	yes	yes
ZNF30	0,0354	-0,53	yes	yes	yes	yes
ZNF626	0,0106	-0,52	no	yes	yes	yes

Table 6.2: List of downregulated DEG sharing similar characteristics including ATRX enrichment over their promoter region. Non-methylated CpG clusters were determined based on Bio-CAP-seq data. PQS identified by (Sahakyan et al., 2017) or QGRS Mapper (Kikin et al., 2006)

Due to the background noise associated with mapping problems, it was not possible to investigate the chromatin changes at the ZNF595 gene, which was the most downregulated DEG (***Figure 6.16***.a). Indeed, ZNF595 is located at the subtelomeric extremity of the genome assembly and such regions are prone to mapping artefacts.

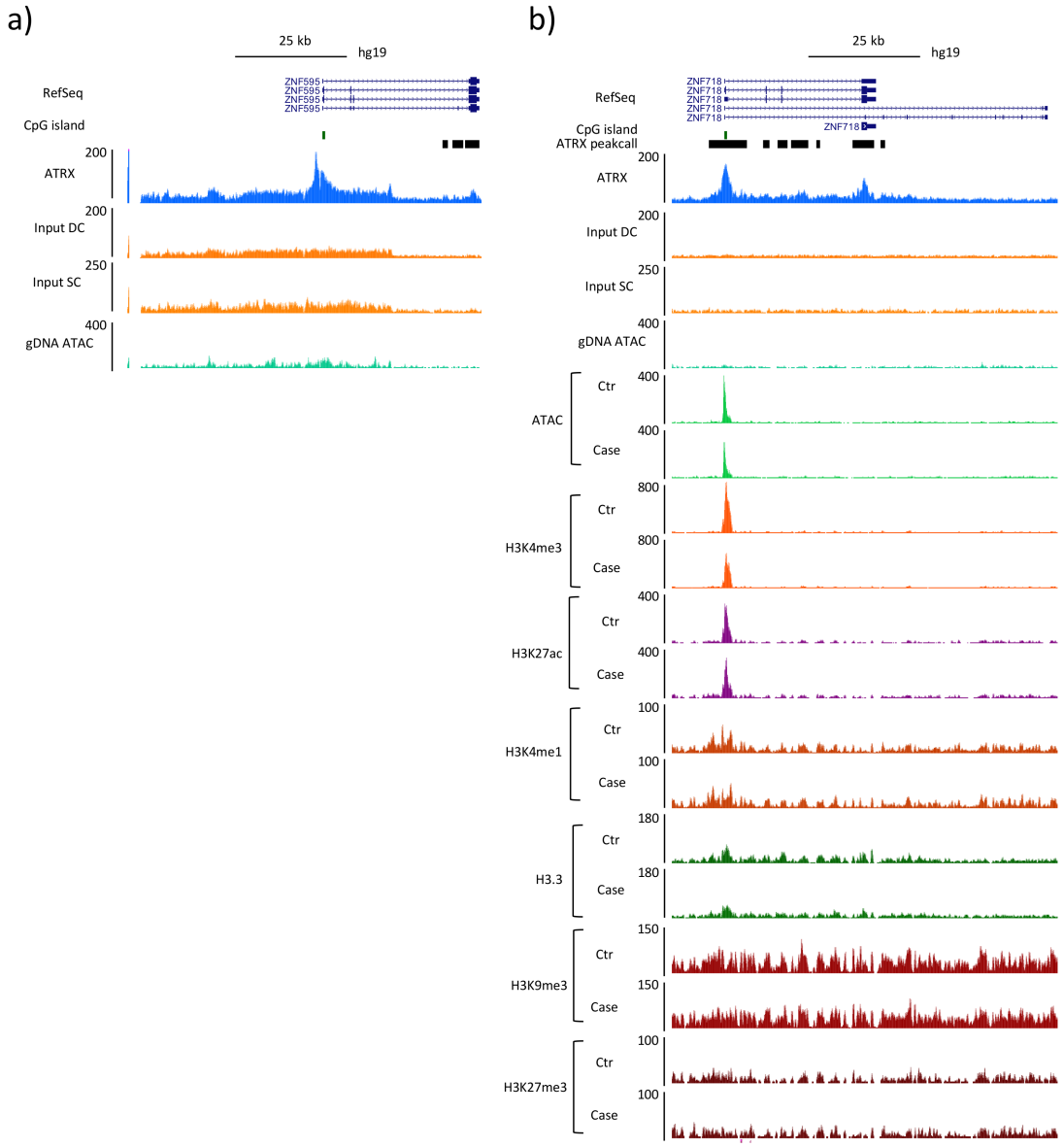


Figure 6.16: Chromatin environment at the ZNF595-ZNF718 subtelomeric region. a. Mapping limitation at the subtelomeric ZNF595 locus detected by input controls and ATAC-seq on gDNA. As a result, the regions presenting artefacts in the inputs were removed from the final peak calling. b. Chromatin environment at the ZNF718 locus.

Interestingly, ZNF718, localised immediately downstream of ZNF595, was also downregulated in patient-derived LCL and a slight reduction in chromatin accessibility and H3K4me3 enrichment could be observed at its promoter region (***Figure 6.16***.b). The subtelomeric region containing ZNF595 and ZNF718 displayed additional ZNF genes further from the telomere that showed a complex pattern of ATRX enrichment (***Figure 6.17***). However, none of the other ZNF genes within this region were significant DEG (all p-values > 0.6). The Capture-C analysis revealed a complex interaction pattern involving the *ZNF718* promoter. Indeed, the *ZNF718* promoter interacted with at least the 3'end of ZNF595 (region with low input background) as well as with other ATRX enriched sites including CTCF sites within *ZNF876P* and the promoter regions of *ZNF732* and *ZNF141* (highlighted in yellow in ***Figure 6.17***). The *ZNF718* promoter and the promoter of another DEG, *PDE6B*, also downregulated in ATRX deficient cells, appeared to share an interacting region located downstream the *PIGG* gene (highlighted by a black box in ***Figure 6.17***). However, the presence of a peak in the input track (related to ATRX ChIP-seq) and in the control track done for ATAC-seq using gDNA may indicate that this signal was an artefact highlighting the importance of considering the mapping limitations even in Capture-C experiments and the necessity of input tracks in interaction studies.

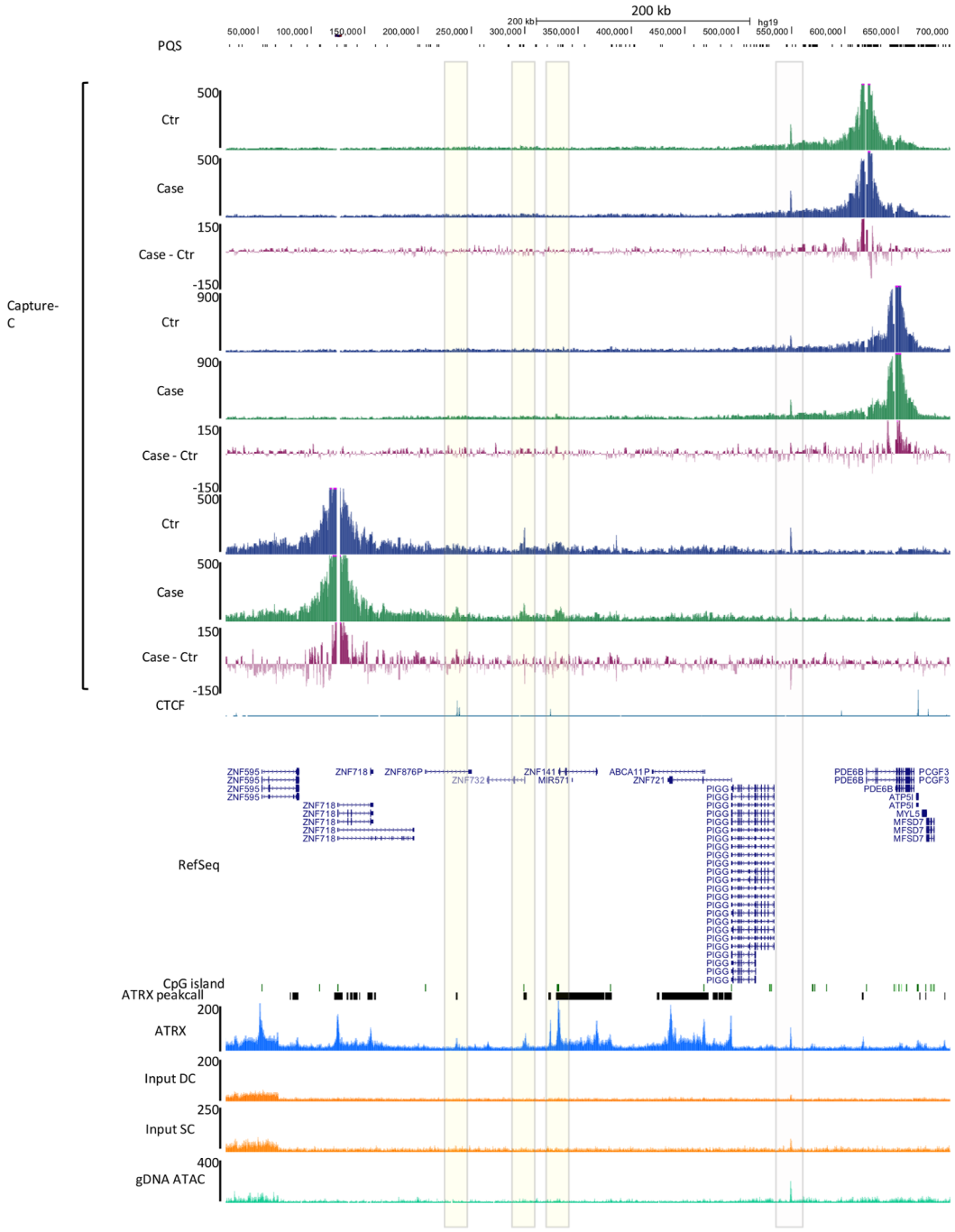


Figure 6.17: Analysis of the chromatin interactions associated with the ZNF718 promoter at the subtelomeric regions of the short arm of chromosome 4.

Similarly to *ZNF595* and *ZNF718*, two other downregulated ZNF DEG, *ZNF57* and *ZNF555* were found within a region containing additional ZNF genes. Like *ZNF595* and *ZNF718*, only *ZNF57* and *ZNF555* were DEG and the surrounding ZNF genes, *ZNF554* and *ZNF77* were not significantly affected by ATRX mutations. Interestingly, ATRX enrichment at ZNF promoters appeared to be associated with gene expression at least at this locus. Indeed, *ZNF556* was the only one that was not expressed and that it did not showed ATRX enrichment at its promoter in contrast with *ZNF57* and *ZNF555* (***Figure 6.18***). However, it displayed an ATRX binding site at its 3' exon.

Both *ZNF555* and *ZNF57* promoters had CpG islands and PQS and were enriched in ATRX. Despite the presence of active chromatin marks and gene expression, the region containing these two ZNF genes was also enriched in H3K9me3. Additionally, some H3K27me3 enrichments were also detected at the promoter region of *ZNF555* (***Figure 6.18*** and ***Figure 6.19***). In ATR-X patient-derived LCL, subtle changes, especially at *ZNF555*, were noticed in chromatin accessibility and H3K4me3 or H3K27ac marks. Additionally, H3K36me3 was reduced at the 3' end of *ZNF555* in cases compare to controls (***Figure 6.18***). Capture-C experiments, focussing on the interactions associated with *ZNF555* and *ZNF57* promoters, showed that these two promoters interact with the other ZNF genes present within the locus and with the 3'end of their own gene. No significant changes in chromatin interactions were noticed between controls and cases (***Figure 6.19***).

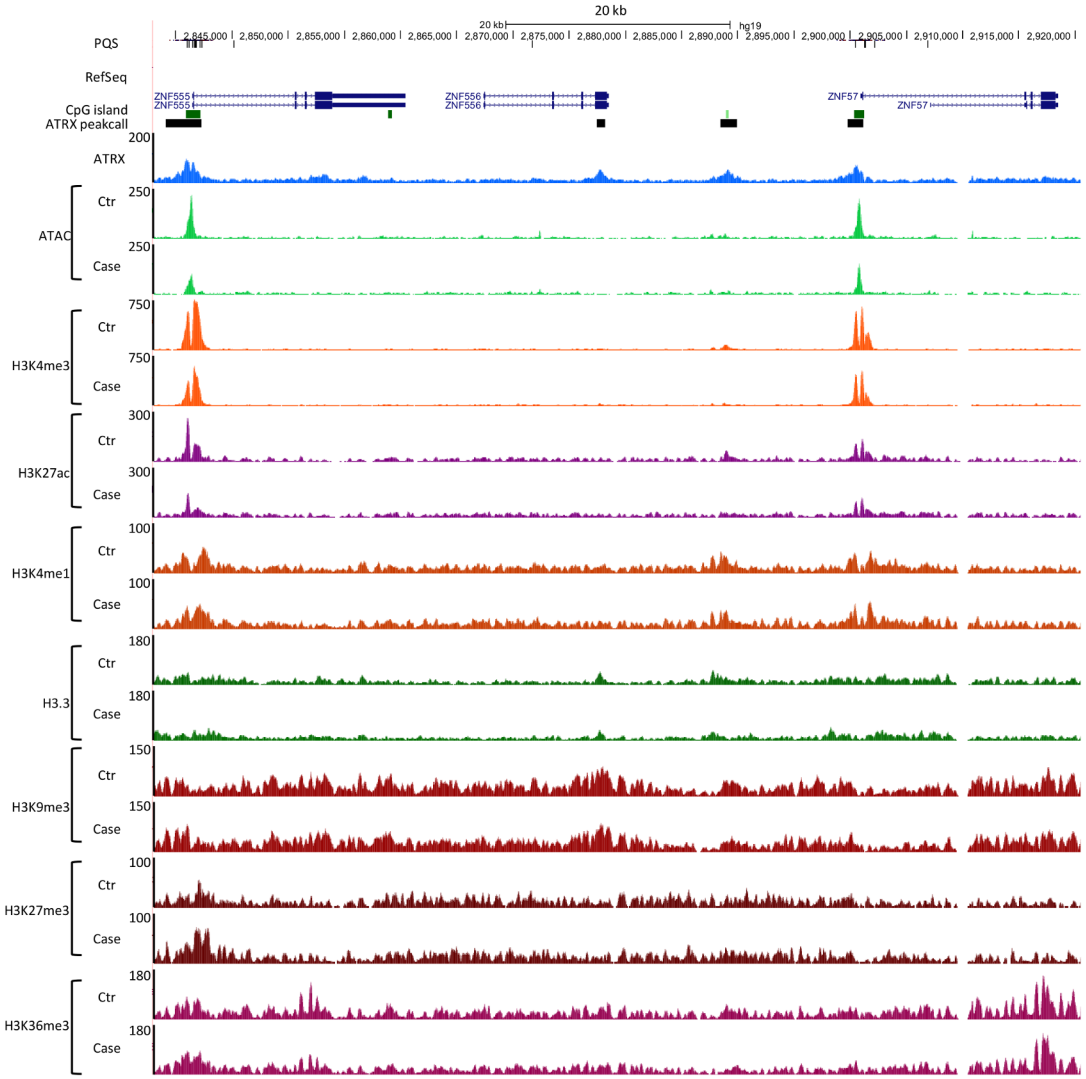


Figure 6.18: Chromatin environment at the ZNF555 - ZNF57 region

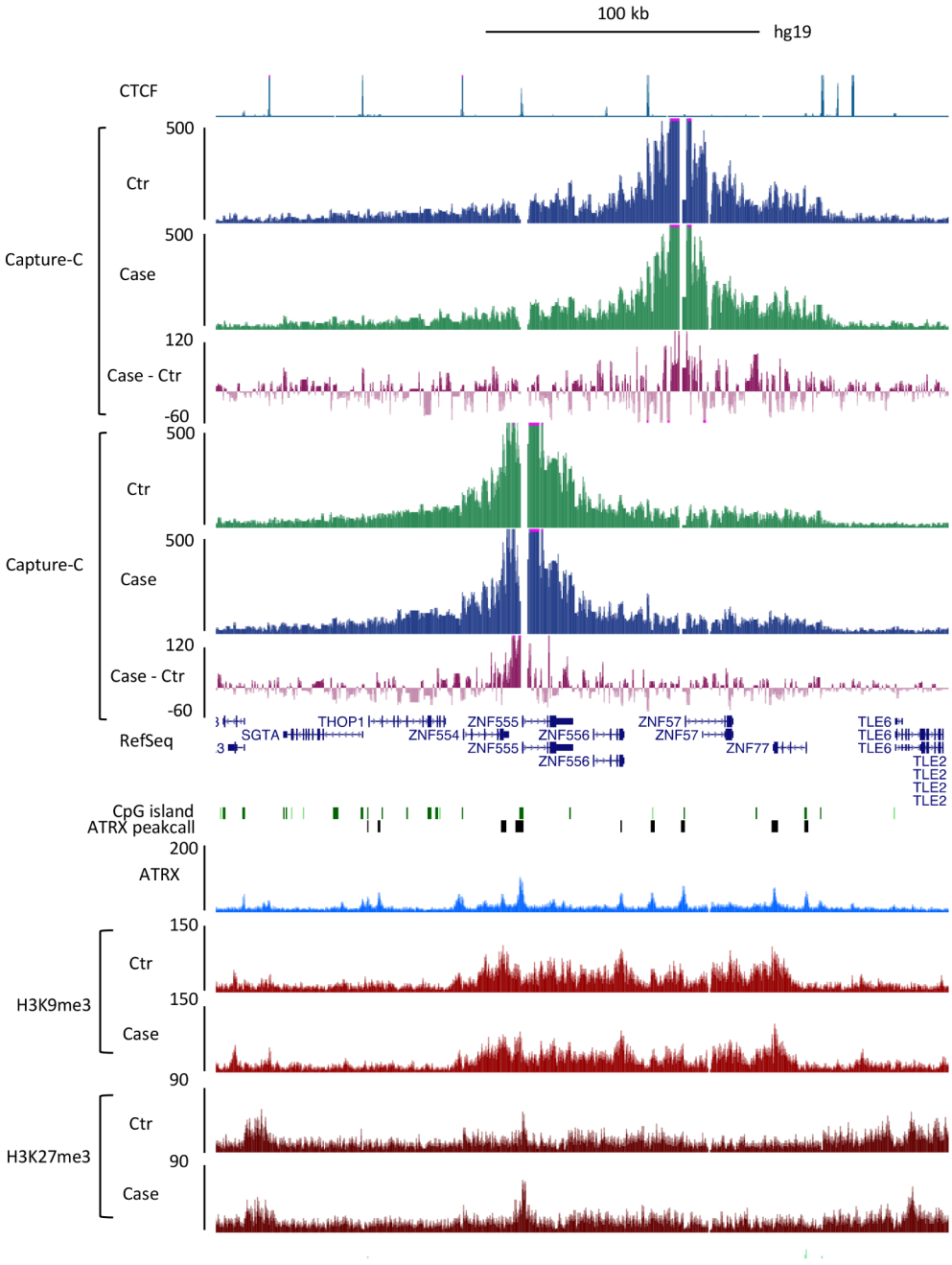


Figure 6.19: Analysis of the interactions associated with the promoters of ZNF555 and ZNF57.

6.2.7. DNA methylation is not sufficient to explain the downregulation observed at certain loci in patient-derived cells.

ATRX deficiency has been associated with aberrant DNA methylation. Furthermore, previous works focussing on NME4 have showed that the downregulation of this gene in ATR-X patients' material and derived LCL is frequently (in 75% of cases) associated with DNA methylation at its promoter region (Karen Lower personal communication). Strikingly, if present, the DNA methylation at the promoter was predominantly associated with the allele containing the largest VNTR. All the DEG mentioned above contained CpG islands/clusters. Therefore, the presence of DNA methylation could explain, at least partially, their downregulation. To test this hypothesis, I performed Southern blot experiments focussing on the ATRX binding sites in the promoter of the three DEG, *ATF7IP2*, *ZNF555* and *ZNF57*, to assess their DNA methylation status and to look for potential VNTRs.

Using two enzymes, one unaffected by nucleotide methylation and the other methylation sensitive, I could detect that the *ATF7IP2* promoter was unmethylated in controls and conversely, partially methylated in ATR-X patient-derived LCL (***Figure 6.20***). Despite the presence of DNA methylation, unmethylated promoter sequences remained in all patient-derived LCL (***Figure 6.20***). As there was no detectable length variation highlighting with the presence of a VNTR between alleles, it was not possible to determine if particular alleles were preferentially methylated. Furthermore, the experiment was carried out using cell population, so it was not possible to determine if the methylated and unmethylated alleles were present in the same or different cells.

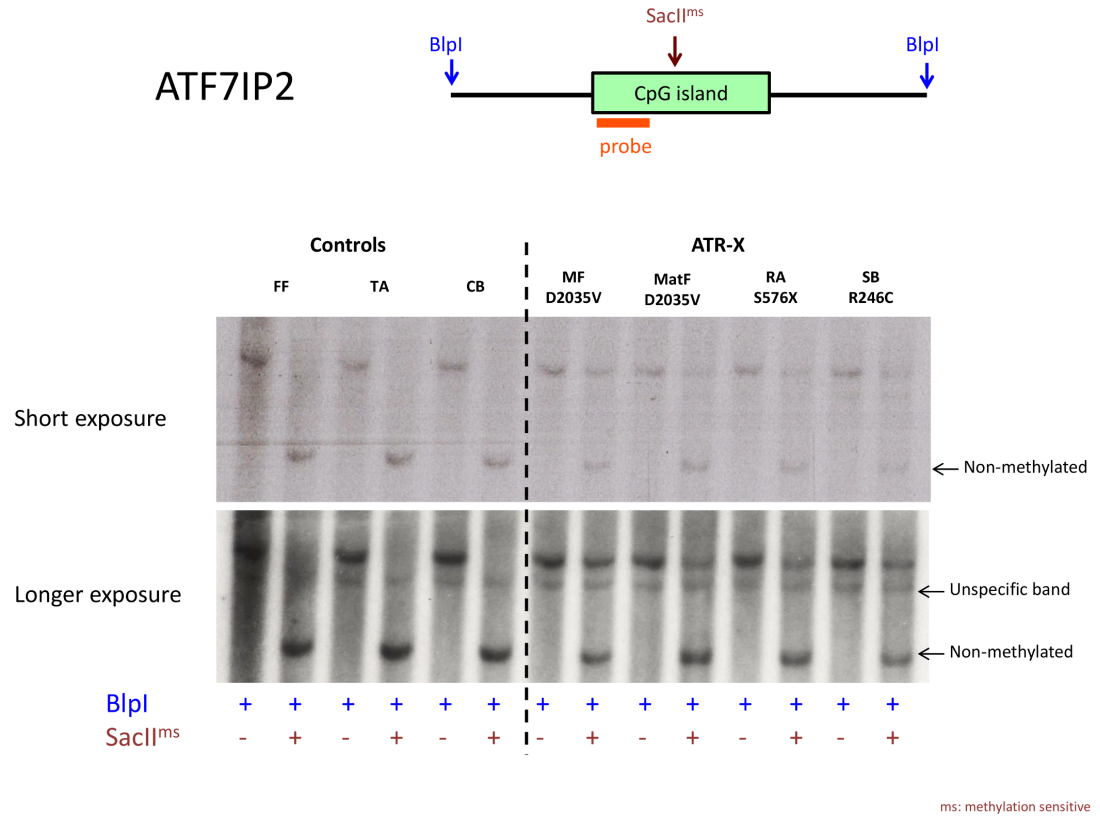


Figure 6.20: Southern blot experiment illustrating the DNA methylation status of the CpG island-containing promoter of ATF7IP2 in LCL.

By contrast with *ATF7IP2*, the subsequent Southern blot experiments focussing on the promoter regions of *ZNF555* and *ZNF57*, did not highlighted the presence of aberrant DNA methylation in ATR-X patient-derived LCL (***Figure 6.21***). In both conditions, cases and controls, the promoter sequences were fully digested by the methylation sensitive enzyme. In the *ZNF555* related experiment, successful double digestion resulted in the detection of two bands of lower molecular weight than the single digest, as the probe covered the methylation sensitive restriction sites (***Figure 6.21***.a).

In parallel, DNAs extracted from venous blood samples were also tested using Southern blots. These samples were derived from the same unaffected individual as in LCL, FF-Ctr, and two ATR-X patients, MF-D2035 and PA-S576X (the RA-S576X's sibling). Samples from venous blood derived from the ATR-X patients confirmed the results

observed in LCL. Unfortunately, the venous blood sample from FF-Ctr failed to provide interpretable results (***Figure 6.21***). No variability between the alleles was detected in either ZNF555 or ZNF57 (LCL and venous blood samples) suggesting that, similarly to ATF7IP2, these regions did not contained VNTRs (***Figure 6.21***). Unlike NME4, the size of the VNTRs could not be associated with the change in gene expression and DNA methylation in ATRX deficient cells. Therefore, ATRX binding sites at the promoter of *ATF7IP2*, *ZNF555* and *ZNF57* probably represent different target types than the ATRX binding sites associated to the *NME4* VNTR.

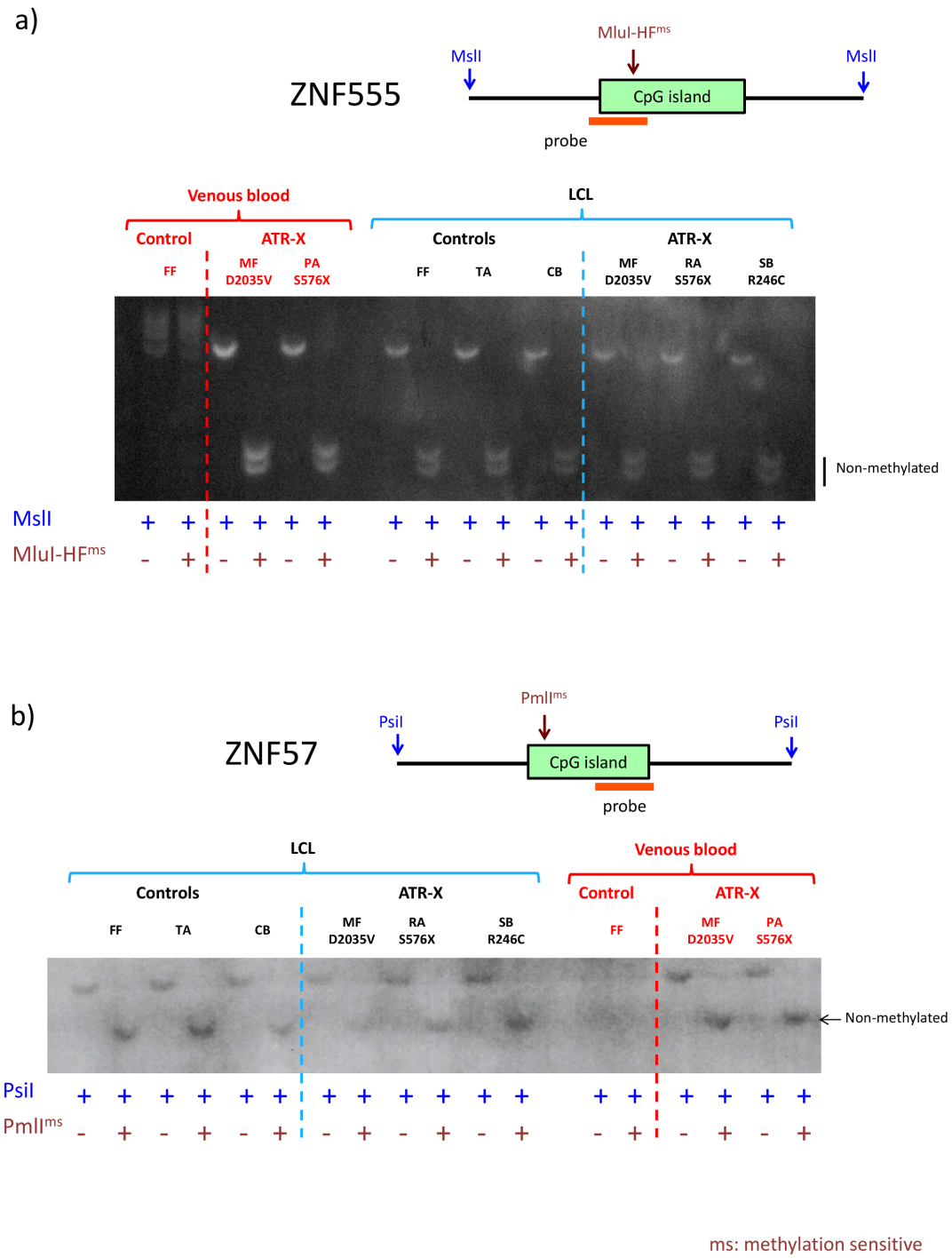


Figure 6.21: Southern blot experiments assessing the DNA methylation status at promoter regions enriched in ATRX at ZNF555 gene (a) and ZNF57 gene (b) showing that, in ATR-X patient derived cells, the decrease in gene expression is not necessarily associated with the presence of aberrant DNA methylation.

To further analyse the DNA methylation status at ATRX binding sites, I performed Bio-CAP-seq experiments in ATR-X patient-derived LCL at the same time than in controls. To optimise the experiment, I first performed Bio-CAP-qPCR at the *ATF7IP2* promoter to assess if I could detect the change in DNA methylation observed by Southern blot. The Bio-CAP-qPCR sensitively detected the difference in DNA methylation at *ATF7IP2* promoter between cases and controls (p-value < 0.001) (***Figure 6.22***.a). I therefore sequenced the Bio-CAP experiments. After normalisation based on the Bio-CAP peak calling, I visualised the variability in DNA methylation between controls and cases. *ATF7IP2* and *ZNF718* promoter regions showed a reduction in Bio-CAP signal in cases (***Figure 6.22***.b and c). By contrast, no obvious differences were observed at the promoter regions of *ZNF555* and *ZNF57*, thus confirming the Southern blot results (***Figure 6.22***.d).

Additionally, I assessed the DNA methylation at *NME4* and the alpha globin loci. Supporting the Southern blot data obtained by former group members, a reduction of the Bio-CAP-seq signal was observed at the *NME4* promoter regions in patient-derived cells (***Figure 6.22***.e). By contrast, the genes within the alpha globin cluster did not showed obvious changes in signal between cases and controls (***Figure 6.22***.f).

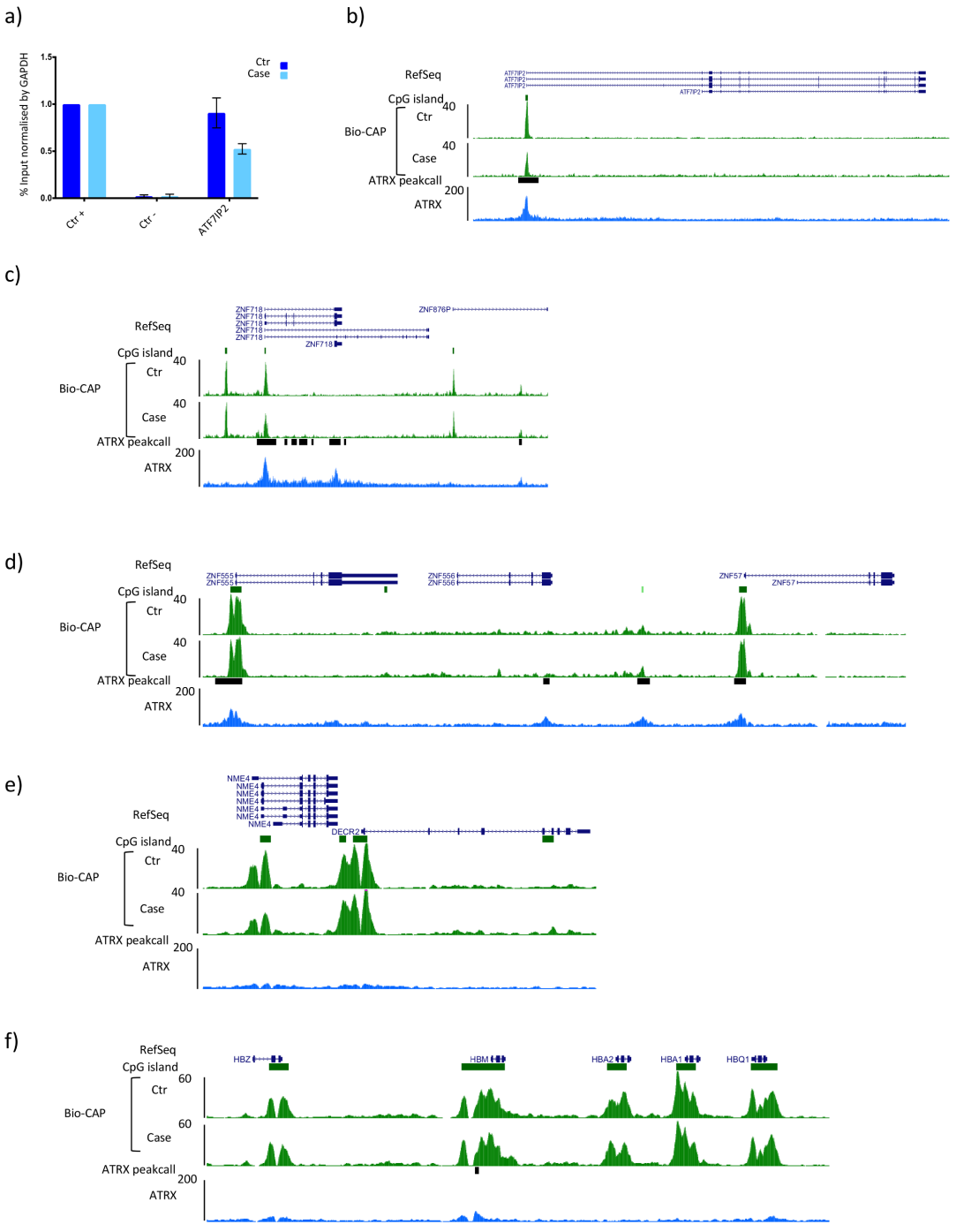


Figure 6.22: Analysis of the DNA methylation status of CpG island-containing promoters by Bio-CAP experiments in LCL. a. Bio-CAP-qPCR showing the reduction of non-methylated CpGs at the ATF7IP2 promoter in cases compared to controls. b. Bio-CAP-seq showing the reduction of non-methylated CpGs at the ATF7IP2 promoter in cases compared to controls. c. Bio-CAP-seq at the ZNF718 locus. d. Bio-CAP-seq at the ZN555/ZNF57 locus. e. Bio-CAP-seq at the NME4 locus. f. Bio-CAP-seq at the alpha globin locus. All the Bio-CAP experiments represent three biological replicates.

Bio-CAP experiment allows the detection of unmethylated CpGs. Thus, the presence of single nucleotide polymorphisms (SNPs) between controls and patients affecting CpGs could theoretically change the Bio-CAP-seq signal at those sites in a similar manner that the presence of DNA methylation. This probability is reduced with the use of unaffected father as controls for each affected related son. However, on analysing potential SNPs in sequences of each individual, no CpG-damaging SNPs were detected at *ATF7IP2* and *ZNF718* promoter sequences. Using the same approach, the analysis of the underlying sequences associated with Bio-CAP signal at *NME4* revealed the presence of a SNP creating a potential novel CpG site in a case. Nevertheless, this signal is reduced in cases compare to patients, which strengthens the hypothesis of the presence of aberrant DNA methylation is cases.

6.2.8. The presence of R-loops is not sufficient to explain ATRX recruitment to the DEG

Aberrant DNA methylation was not sufficient to explain the downregulation and chromatin changes observed at a subset of DEG. An alternative explanation could be found in recent work from the group showing a correlation between ATRX recruitment and R-loop formation at telomeric sequences. Therefore, I performed DNA immunoprecipitation (DIP) to detect R-loop enrichment using an antibody specific to RNA:DNA hybrids. As a control, I carried out the same experiment with a prior treatment with RNaseH to digest R-loops. Therefore, the control should show a basal level similar to the R-loop negative regions. As a positive control region for R-loop formation, I used the characterised R-loop enrichment at the *RPL13A* loci for which I beforehand ensured that ATRX was not bound. I have done the same for the R-loop negative control. Using qPCR, I observed a drop in signal after RNaseH treatment

(***Figure 6.23***). Most of the tested regions, including *ATF7IP2,* showed enrichments similar to the negative control (***Figure 6.23***). *ZNF57* promoter region was the only one that displayed significant signal enrichment for R-loop (***Figure 6.23***). However, no significant difference was detected between cases and controls at any of the tested sites. These data suggested that R-loop is unlikely to explain the changes observed at that sites and the related downregulated DEG.

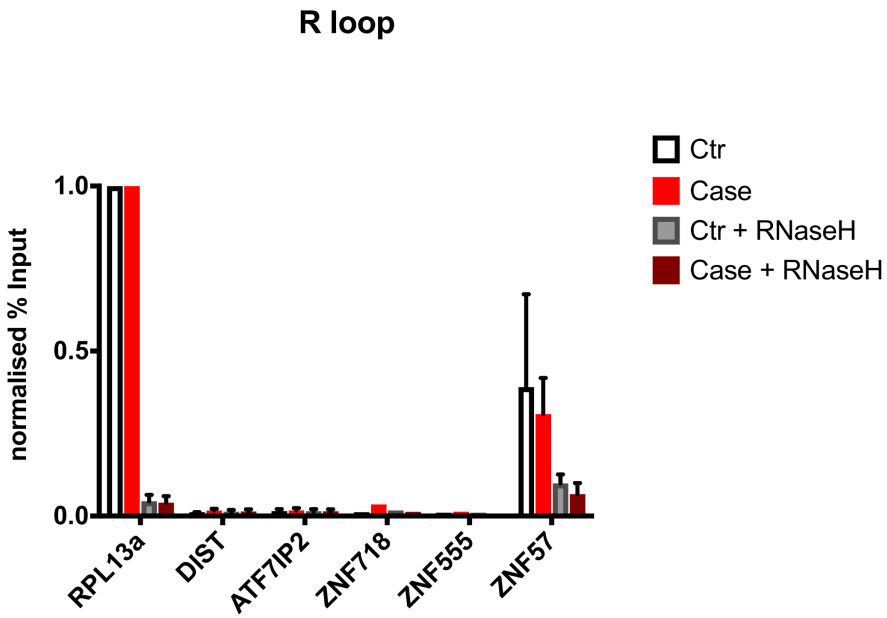


Figure 6.23: Analysis of R-loop formation based on DIP-qPCR at a subset of DEG promoters enriched in ATRX. The experiments were performed with three biological replicates except for ZNF718 and ZNF555, which were the result of a single experiment.

6.2.9. Assessing the presence of G4 *in vivo*

Another hypothesis was whether the presence of G4 played a role in the differential gene expression between cases and controls. Thus, I performed G4 ChIP using BG4 antibody to assess *in vivo* the presence and enrichment in G4 (Hänsel-Hertsch et al., 2016). As a control, I carried out parallel experiment using a pyridostatin treatment to stabilise G4 with the assumption that if the G4 ChIP worked, the signal should be stronger in the presence of pyridostatin. As a negative control for G4 enrichment, I used a region where no G4 was predicted for more than a 2kb (named chr8 NG4 FR1 in ***Figure 6.24***).

Unexpectedly, upon pyridostatin treatment, the signal decreased compare to the DMSO control (***Figure 6.24***.a). Furthermore, the negative control region showed a relatively high enrichment suggesting that the experimental background was very high (***Figure 6.24***.a and b). In three replicates, G4 enrichment was not significantly different between cases and controls at *ATF7IP2* and the level of enrichment was close to the negative control (***Figure 6.24***.b). G4 ChIP-seq did not show a significant enrichment in control or patient even at the G4 forming sequence associated with the Myc promoter (Hänsel-Hertsch et al., 2016) (***Figure 6.24***.c and d). Consequently, the analysis of *in vivo* G4 distribution and enrichment was not pursued further.

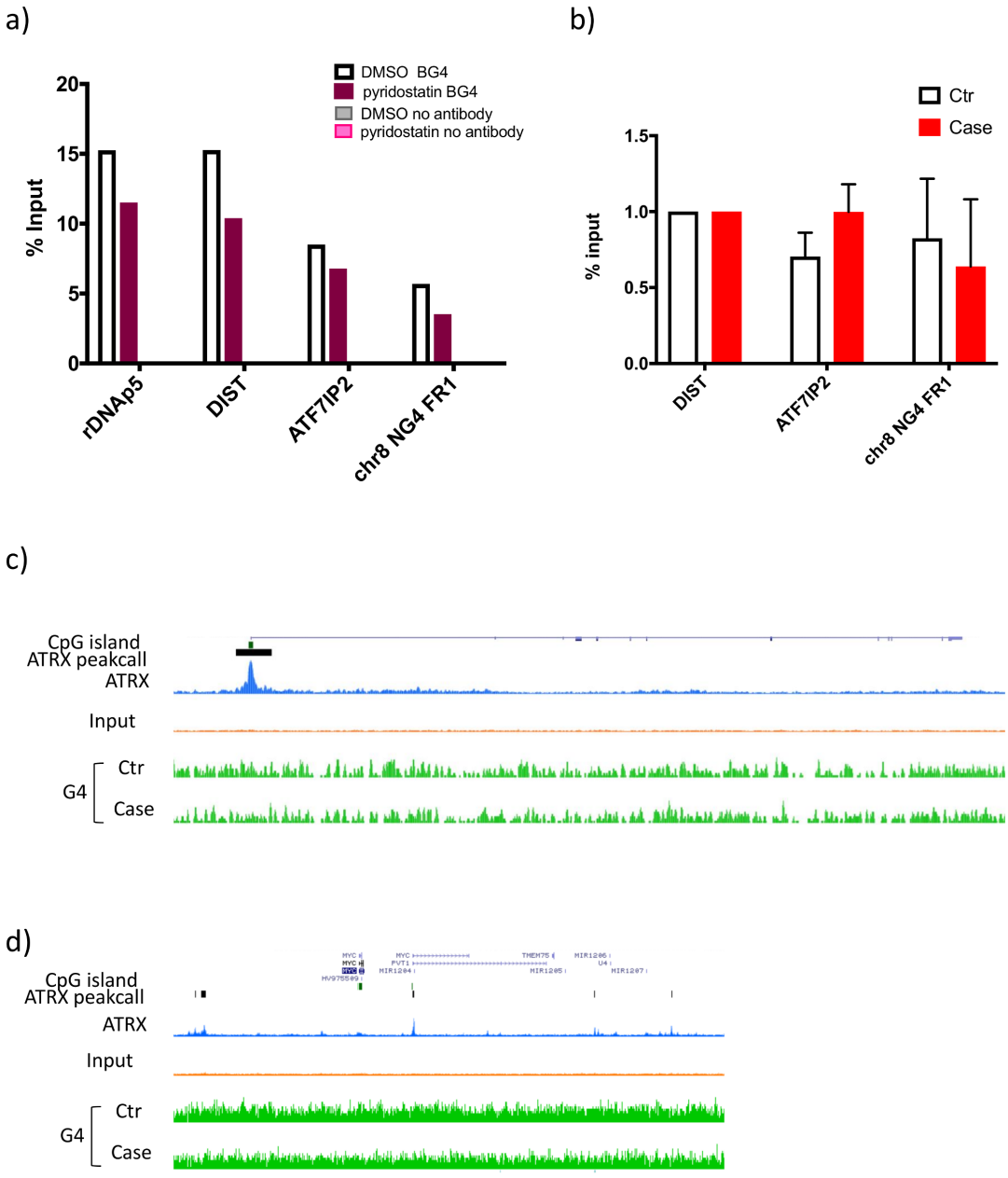


Figure 6.24: Analysis of G4 in vivo. a. G4 ChIP-qPCR done in control LCL using DMSO (control) or pyridostatin (G4 stabiliser) treatment. b. G4 ChIP-qPCR done in three controls and three cases. c. G4 ChIP-seq at the ATF7IP2 locus. d. G4 ChIP-seq at the Myc locus

6.2.10. The ATRX affinity for ZNF genes

ATRX is enriched at a large number of ZNF genes, many of which are clustered. However, only a subset of them were significantly dysregulated in ATRX deficient cells. Even within the same ZNF cluster, only some of them are affected such as at the *ZNF718/ZNF595* and *ZNF555/ZNF57* loci. What is the difference between these ZNF genes and the surrounding others?

I further investigated the surrounding ZNF genes. All the tested ZNF genes belong to the same group as the differentially expressed ZNF genes, the KRAB-containing C2H2-ZNF. They were all enriched in ATRX, with ATRX present at the promoter if they are expressed. Gene expression analysis using the Fluidigm Biomark supported the idea that *ZNF554* expression was not significantly affected by ATRX deficiency, despite its localisation immediatly upstream and in the same orientation as the DEG *ZNF555.* All the tested ZNF genes (DEG or not) displayed PQS around their promoter regions and/or within their first intron. They were all localised within region enriched in H3K9me3 but no significant correlation was observed between ZNF expression, ATRX occupancy at the 3' exon and reduction of H3K9me3 in patient-derived LCL compare to controls.

ZNF genes that display active chromatin marks also tend to show H3.3 enrichment. However, only ZNF718 showed a decrease in H3.3 enrichment in ATRX deficient cells. So, further work is needed to determine how some ZNF genes are affected by ATRX deficiency whereas other are not.

6.2.11. Recapitulation of the effects observed in ATRX deficient cellls via ATRX KO in control LCL and alternatives

With the help of Anthony Cheong, I attempted to KO ATRX in control LCL using CRISPR Cas9, to reveal, on a given genotype, the direct influence of ATRX deficiency on gene expression and chromatin environment. I attempt to transfect LCL using nucleofection with plasmids expressing Cas9 with either GFP or puromycin resistance. However, I failed to obtained suviving cells using both plasmids, even when using a control vector without any single guide RNA, suggesting that the guide or ATRX KO were unlikely to be the cause of cell death. Instead, LCL may not tolerate the nucleofection process, illustrating the difficulty of manipulating some cell types.

As an alternative, I have performed some gene expression analyses on induced pluripotent stem cells (iPSC) from a normal individual, developed by other members of the group and in which ATRX had been KO. Preliminary tests using human iPSC mRNA provided by Yavor Bozhilov, showed that *NME4*, *ZNF555* and *ZNF57* appeared to be downregulated upon ATRX loss (***Figure 6.25***). By contrast, changes in *ZNF595*, *ZNF718* and *ATF7IP2* were not detected. However, these results were produced based on a single experiment, thus replicates are required before any definitive conclusion can be made.

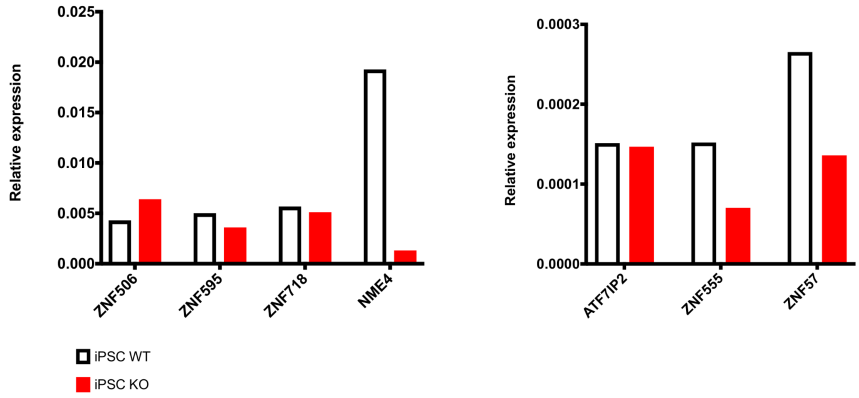


Figure 6.25: Gene expression analysis using RT-qPCR in iPSC either WT or KO for ATRX.

Similarly, I could not detect a difference between WT and ATRX KO in *ATF7IP2* and *ZNF506* expression using mRNA from undifferentiated human umbilical cord blood-derived erythroid progenitor (HUDEP) cells provided by Anthony Cheong (***Figure 6.26***.a Left). However, *ZNF555*, *ZNF57*, *ZNF595* and *ZNF718* were all downregulated in the absence of ATRX, which strenghtens our findings (***Figure 6.26***.a left). Besides, preliminary analyses performed on mRNA from differentiated HUDEP, suggested that *ZNF555*, *ZNF57* and *ZNF595* remained downregulated (***Figure 6.26***.a Rigth). ATAC-seq data (provided by Anthony Cheong), showed a reduced chromatin accessibility at *ZNF555*, *ZNF57* and *ZNF718* in ATRX KO cells (***Figure 6.26***.b and c). Interestingly, *ATF7IP2* and *ZNF506* did not display open chromatin associated with their promoter region by contrast with the data observed in LCL (***Figure 6.26***.d). These may explain why no downregulation was detected at these genes in ATRX KO. Furthermore, no ATRX ChIP data were available for these experiments and it remains to assess if ATRX is present or not at the tested targets. Indeed, an absence of ATRX enrichment may also explain the absence of observed effects upon ATRX loss. All together, these results support the role of ATRX in the regulation of at least a subset of the candidate genes in other cell types.

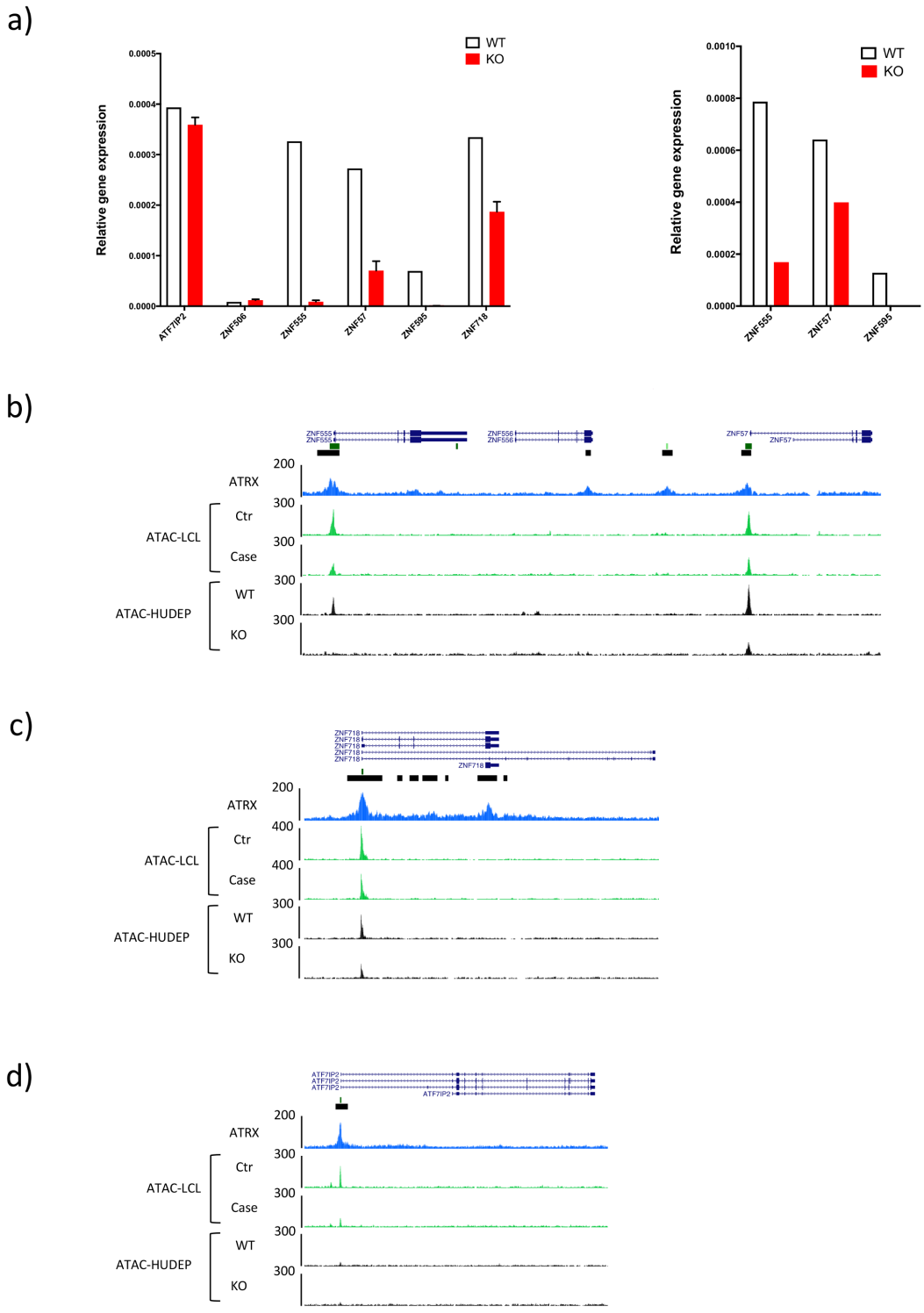


Figure 6.26: Analysis of gene expression and chromatin environment in HUDEP cells with ATRX WT or KO. a. Left: Gene expression analysis based on RT-qPCR in undifferentiatied HUDEP cells. The analysis was performed using four independent ATRX KO clones. Right: Gene expression analysis based on RT-qPCR in differentiated HUDEP cells (only one

replicate). b-d: Analysis of the chromatin accessibility in differentiated HUDEP cells, either WT or KO for ATRX, compared to data obtained in LCL derived from ATR-X patients or unaffected relative. ATRX ChIP-seq data were performed in LCL. ATAC-seq data in differentiated HUDEP were performed by Anthony Cheong

6.3. Discussion

Effects of ATRX impairment on the chromatin landscape have been mainly studied in mouse and studies in human remained limited. Using ATRX ChIP-seq and the DEG identified in chapter 5, I further investigated the chromatin changes that were associated with these loci, as well as the relevance of the findings described in mouse, in the context of human LCL. I first assessed the chromatin changes associated with imprinted DEG in LCL. Four imprinted genes (three upregulated and one downregulated) were present within the DEG including *NDN* and *MKRN3*. These two upregulated DEG were devoid of ATRX binding sites but showed an increase in active chromatin marks in patient-derived cells. In mouse ES cells, de-repression of imprinted genes was observed and it has been suggested that ATRX-dependent H3.3 deposition was important for the establishment of the repressive chromatin environment during the imprinting process (Voon et al., 2015). It is therefore conceivable that ATRX-dependent H3.3 deposition is necessary during this process but becomes dispensable after imprinting establishment. In this scenario, ATRX recruitment to the imprinted loci would become unnecessary after cellular differentiation. However, the aberrant imprinting would have been established early in development and the chromatin landscape would subsequently be maintained in differentiated cells such as LCL. It is therefore plausible that the variations in chromatin environment observed in LCL, between cases and controls, were the residual effects of ATRX dysfunction during development. To explore this possibility future studies would need to focus on imprinting processes in human ES

cells and in differentiated cells in which ATRX would be depleted after the imprinting establishment.

Additional studies performed in mouse confirmed the association between ATRX loss, upregulation of imprinted genes and changes in chromatin interactions (Kernohan et al., 2014). It seems unlikely that such a mechanism occurred at the imprinted locus *LILRB4* (unique DEG potentially fulfilling these criteria) as, despite the presence of an ATRX binding site within 100kb of this gene, no changes in the chromatin environment or interactions were observed in the patient-derived LCL.

Further work in mouse focussing on aPAR genes supported the notion that ATRX was required for transcription through PQS (Levy et al., 2014). Despite the presence of PQS in human homologs, none of the aPAR genes tested in the mouse study were differentially expressed in LCL. Overall, it seems likely that the differences in findings were related to the species or cellular context or the nature of the ATRX mutation (hypomorph vs null). These issues emphasise the importance of Sydney Brenner's observation that if you want to understand human disease you should "study human in a dish".

In human cells, an ATRX binding site was identified within the first intron of *NME4* (Law et al., 2010). This kinase-encoding gene was downregulated in ATRX deficient cells. Although a similar dysregulation of *NME4* was observed in the present study, no significant ATRX enrichment was detected at the *NME4* VNTR, which may be partly due to mapping artefacts. However, clear changes in active chromatin marks were observed between cases and controls.

The discovery that ATRX binding sites were enriched in active regulatory elements raised some questions regarding the role of ATRX at such sites. Interestingly, subsets of ATRX binding sites overlapped with putative enhancers and promoters and were associated with DEG showing alterations of their chromatin environment in ATRX deficient cells. Indeed, at the chemokine receptor *CCR1* locus, I detected a physical interaction between *CCR1* promoter and putative enhancers, which are both enriched in ATRX. In ATRX deficient cells, these regions showed some increase in active chromatin marks and CCR1 was upregulated. Moreover, the chromatin interactions were maintained, even slightly increased between the CCR1 promoter and the enhancer cluster. This suggested that at that locus, ATRX did not appear to work in a similar manner to that seen at the H19 ICR locus in mouse forebrain where fewer interactions across the domain were detected in the ATRX KO (Kernohan et al., 2014). Interestingly, H3.3 enrichment did not appear affected by ATRX deficiency at the CCR1 locus. This observation may suggest an alternative method for H3.3 deposition such as via HIRA complex. Otherwise, the hypomorphic nature of the ATRX mutations may be sufficient to retain this function. Alternatively, regulation of this locus is associated with another function of ATRX unconnected with H3.3 deposition (Kernohan et al., 2014, Hoelper et al., 2017). Although both CCR1 promoter and its upstream putative enhancer cluster contained PQS, no obvious variations in H3K9me3 and H3K27me3 were noticed between control and ATRX deficient cells. This result suggests that de-repression of the CCR1 gene is unlikely to be associated with the loss of epigenetic memory. Changes in DNA methylation were also unlikely as no changes in Bio-CAP profile were noticed at these ATRX binding sites. It is possible that ATRX is involved in transcription repression as noted by Tang et al. when they tethered it to

DNA upstream of the promoter of a reporter gene (Tang et al., 2004); ATRX deficiency may alleviate the repression of these targets.

At the *UST* locus, ATRX interacted with several putative enhancers both intragenic and intergenic. The changes in chromatin environment differ depending on the sites in ATRX deficient cells. Some of the putative enhancers showed an increased enrichment in active chromatin marks in ATRX deficient cells. As at the *CCR1* locus, H3.3 enrichment was not reduced under these conditions, again suggesting some pathway redundancy for H3.3 deposition at these active sites. These observations and hypotheses are in line with the modest effect of HIRA depletion on H3.3 deposition at TFBS as previously observed (Goldberg et al., 2010). The increase in active chromatin marks may suggest that *UST* upregulation may be due to the over activation of some of its interacting enhancers. Alternatively, *UST* dysregulation could be associated with the presence of an intragenic promoter associated with an uncharacterised non-coding RNA gene, LOC105378047. The reduction of active chromatin marks at this putative promoter in ATRX deficient cells could be associated with a decrease in expression of LOC105378047. Could the expression of LOC105378047 affects *UST* expression or vice versa? Unfortunately, LOC105378047 was not present in the microarray probes and additional work may address these questions.

Putative enhancers enriched in ATRX were also associated with downregulated DEG. Indeed, the *PBX4* gene displayed a putative intragenic enhancer enriched in ATRX. In ATRX deficient cells, this site showed a decrease in active chromatin marks and reduced H3.3 enrichment. However, its chromatin interaction profile did not appear particularly affected. By contrast with the sites described above, the decreased in H3.3 enrichment suggested that, at least at this site, ATRX complex may be the main

pathway for the replication-independent deposition of H3.3 and that ATRX impairment could not be compensated by another complex. Similar observations were done for instance in Drosophila in which Xnp and HIRA were showed to have a subset of distinct target sites (Schneiderman et al., 2012). The putative intragenic enhancer at *PBX4* locus contains PQS and showed a decrease in active chromatin marks as well as in chromatin accessibility in patient-derived cells. However, no changes were noticed in Bio-CAP so it was unlikely that the downregulation of the gene was the result of DNA methylation. However, the data could fit with G4-associated loss of epigenetic memory as observed in avian cells upon nucleotide depletion or PrimPol deficiency (Papadopoulou et al., 2015, Schiavone et al., 2016). Indeed, avian cells depleted in nucleotides or PrimPol, leading to replicative stress, showed reduced gene expression and active chromatin marks which were dependent on the presence of G4 (Papadopoulou et al., 2015, Schiavone et al., 2016). However, the presence of aberrant DNA methylation within the studied locus was only detected in a subset of cells where gene expression of the locus was the most severely downregulated (Schiavone et al., 2016).

Further results supported the hypothesis that G4-associated loss of epigenetic memory could occur at a subset of ATRX targets in LCL. Indeed, a subset of downregulated DEG had a CpG island/cluster-containing promoter with PQS and ATRX enrichment. In patient-derived LCL, a reduction of the chromatin accessibility, H3.3 or active chromatin marks, and/or partial DNA methylation were observed at a subset of them. The partial DNA methylation detected at promoters of the DEG *ATF7IP2*, *ZNF718* and *NME4* supported recent results obtained on peripheral blood, which showed abnormal DNA methylation at both *ATF7IP2* and *ZNF718* promoters in ATR-X patients (Schenkel et al., 2017). However, aberrant DNA methylation was observed in only a

subset of this subgroup of downregulated DEG suggesting that DNA methylation was not sufficient to explain the dysregulation and changes in chromatin environment of this subgroup.

ATRX binds G4 structures *in vitro* and associates with PQS sequences including telomeres and the ψζ VNTR (Clynes et al., 2014, Law et al., 2010). Luciferase reporter assays have also supported the role of ATRX in the maintenance of gene expression in the presence of PQS ψζ VNTR (Yue, 2016). Furthermore, a number of studies have associated ATRX loss with DNA replicative stress and DNA damage (Clynes et al., 2014, Leung et al., 2013, Huh et al., 2012). Therefore, it is possible that ATRX deficiency associates with the reduction of H3.3 deposition at the PQS-containing promoters and enhancers of these downregulated DEG and leads to the stabilisation of G4. The persisting G4 would therefore impede the DNA replication, resulting in replication stress, potential DNA damage and impairment in the recycling of the parental histones. As a result, stretches of newly synthesised histones lacking the appropriate PTM (e.g. H3K4me3) would accumulate at the promoters/enhancers of these DEG, which will subsequently reduce transcriptional activity. Interestingly, attempts to produce complete PrimPol KO in pre-existing ATRX KO HUDEP cells were unsuccessful (Anthony Cheong personal communication). Unless further experiments may successfully produce complete PrimPol/ATRX double KO, these preliminary results suggested a potential requirement of at least PrimPol or ATRX for cell survival.

Indubitably further work including additional ChIP-qPCRs and bioinformatic analysis would be valuable to strengthen the relevance of the ChIP-seq data and assess the statistical significance of the data. Some answers could have been brought by the use of ATRX KO cells. Unfortunately, I could not produce ATRX KO in LCL derived from

the unaffected fathers to assess if the phenotypes observed in ATR-X patient-derived LCL could be recapitulated upon ATRX KO. However, the use of ATRX KO human iPSC and HUDEP cells supported the downregulation and the decrease in chromatin accessibility at a restricted number of targets.

Surprisingly, *ATF7IP2* did not appeared affected by ATRX loss in these models. It is possible that the effects of ATRX deficiency at the *ATF7IP2* locus occurred during development and may require a threshold level of expression to occur. Following these events, the *ATF7IP2* promoter may remain hyper-methylated in ATR-X patients, which would result in the maintenance of repressed *ATF7IP2*, independently of ATRX. If this hypothesis was founded, it would suggest that the phenotype observed in ATR-X patients would be a complex composition of both; set events acquired during development and subsequently maintained due to aberrant chromatin environment (*ATF7IP2*), and continuous events due to the present impairment of ATRX (*ZNF555* and *ZNF57*). However, differentiation of either ATRX KO iPSC or HUDEP cells toward the erythroid pathway both failed to recapitulate the alpha-thalassaemia phenotype observed in patients. One hypothesis is that, in these models, the expression of alpha globin may not reach the threshold required to recapitulate what is observed in patients. Moreover, *in vitro* differentiation may be a too simplistic approach, which does not consider the surrounding environment found in individuals that may affect the subsequent gene expression. Similarly, these systems may not be optimum for the study of *ATF7IP2*. Interestingly, the chromatin accessibility observed at the *ATF7IP2* promoter in LCL was not present in HUDEP cells. Thus, this observation may be linked to the differential impact of ATRX deficiency in these distinct cell types. Furthermore, it remains unknown at that stage if *ATF7IP2* promoter is enriched in ATRX in HUDEP cells.

Several putative transcription factors and co-factors have been identified among the DEG and it seems likely that at some sites the effects of ATRX deficiency are indirect rather than direct. This was the case at the *BAIAP2L1* gene. This gene displayed a significant change in gene expression and chromatin interactions in ATRX mutant cells. However, no obvious interactions between *BAIAP2L1* promoter and ATRX binding sites were observed. *BAIAP2L1* is probably indirectly affected by ATRX mutations via the dysregulation of putative transcription factors for instance.

The transcription factor interacting protein ATF7IP2 is supposed to be involved in the MBD1-dependent transcriptional repression through the recruitment SETDB1 therefore bridging DNA methylation recognition and H3K9 methylation (Ichimura et al., 2005). Moreover, ATF7IP2 also interacts with Sp1 transcription factors, which recognise GC-rich motifs at many promoters (Ichimura et al., 2005). Regarding the recent association between ATRX and/or DAXX and SETDB1, and the findings highlighting a high proportion of ATRX binding sites at active promoter regions, it is possible that complex interactions between these diverse proteins are required for the proper regulation of gene expression during the development and beyond. Therefore, disruption of this regulation processes could provoke phenotypic consequences.

The DEG *ZNF57* appeared to be downregulated, not only in LCL, but also in ATRX KO iPSC and HUDEP cells. Furthermore, no DNA hypermethylation was observed in ATR-X patients at this promoter, suggesting that this gene is a strong candidate to be directly affected by ATRX deficiency. Interestingly, ZNF57 has been associated with the methylation integrity of ICR at imprinted loci and loss of ZNF57 triggers hypomethylation (Quenneville et al., 2011). As ATRX loss resulted in reduction of methylation, and de-repression of the repressed imprinted allele, it is tempting to

propose that *ZNF57* downregulation could be associated with the aberrant DNA methylation at such loci in ATRX deficient cells.

Strikingly, despite ATRX affinity for ZNF genes, only a small number of them were differentially expressed and showed changes in chromatin environment. Although ATRX enrichment was observed at the 3' end of some ZNF genes, the differentially expressed ZNF genes showed a significant ATRX enrichment at their promoter and no obvious decreased in H3K9me3 at their 3' end. None of the DEG examined in the present study were specifically described by Valle-Garcia and colleagues (Valle-García et al., 2016). Furthermore, H3K9me3 and H3K36me3 ChIP-seq performed in the current study involved a single cross-linking step whereas native ChIP-seq was used in their study and the cell lines used are different.

Despite attempts to elucidate why certain ZNF genes were affected in ATRX deficient cells but not others, I could not reach a satisfactory conclusion. Indeed, the study of the presence/absence of several features such as R-loop, PQS, chromatins mark, did not highlight a pattern common to all the differentially expressed ZNF genes and sufficiently divergent from the non-DEG tested. Specific and expert analyses, carried out on these genes, are probably required to shed light on the ATRX operation at ZNF genes.

Chapter 7. The ATRX distribution in human differentiated CD34+ cells

7.1. Introduction

The use of immortalised LCL allowed the optimisation of the ATRX ChIP-seq protocol to produce robust and sensitive datasets. These analyses highlighted the presence of ATRX binding sites at unsuspected euchromatin sites, which were associated with active regulatory elements. Further analyses using LCL derived from ATR-X patients supported a role for ATRX in the maintenance of the chromatin environment and gene expression. Despite these findings, the LCL model has limitations for understanding the role of ATRX in the regulation of gene expression.

Indeed, this immortalised cell line, unlike primary cells, has adaptive changes associated with cell culture, which weakens the finding generalisation to other human cells. Furthermore, LCL is not an optimal model to investigate the major phenotypes observed in ATR-X patients such as learning disabilities and alpha-thalassaemia as they are derived from the lymphocyte precursors. Additional studies performed in other cell models would certainly improve the robustness of the results.

Taking advantage of a conditional ATRX KO mouse model that displays brain development defects, several studies have been undertaken to investigate the role of ATRX in brain development and the learning disability in ATR-X patients (Berube et al., 2005, Kernohan et al., 2010). In contrast, no such mouse model exists for the analysis of the alpha-thalassaemia associated with ATR-X syndrome. The common types of alpha-thalassaemia associated with mutations in the alpha globin cluster can be modelled in mouse (Chang et al., 1996). However, in ATR-X patients with alpha-thalassaemia, their alpha globin cluster is normal and to date, no mouse model has

successfully recapitulated the alpha-thalassaemia associated with ATRX deficiency. Our best insight into the role of ATRX in the regulation of the alpha globin cluster emerged from the original ATRX ChIP-seq experiment performed in human erythroid cells (Law et al., 2010). This work has highlighted the presence of ATRX at the PQS ψζ VNTR and suggested a correlation between the size of the VNTR and the severity of the alpha-thalassaemia based on the presence of HbH inclusions (Law et al., 2010). A subsequent study in K562 cells has supported a role for ATRX in the regulation of macroH2A enrichment at the alpha globin genes (Ratnakumar et al., 2012). However, due to the high background associated with the ATRX ChIP-seq in human erythroid cells, it has been impossible to confirm the presence of ATRX enrichment at the alpha globin genes (***Figure 7.1***). Additional studies showed that ATRX alleviates the repressive effect of the ψζ VNTR on the expression of a gene reporter and highlighted the presence of R-loop formation upon transcription activation through the ψζ VNTR (Li et al., 2016, Nguyen et al., 2017). Importantly, in the original study, ATRX enrichment at ψζ VNTR was only detected in erythroid cells but not in cells where the alpha globin was not expressed (Law et al., 2010).

In this context, we opted to re-investigate ATRX distribution in erythroid cells. Therefore, I performed ATRX ChIP-seq experiments in differentiated CD34+ cells extracted from unaffected human donors. The use of this approach would assess the findings seen in LCL and contribute to a better understanding of the role of ATRX at the alpha globin cluster.

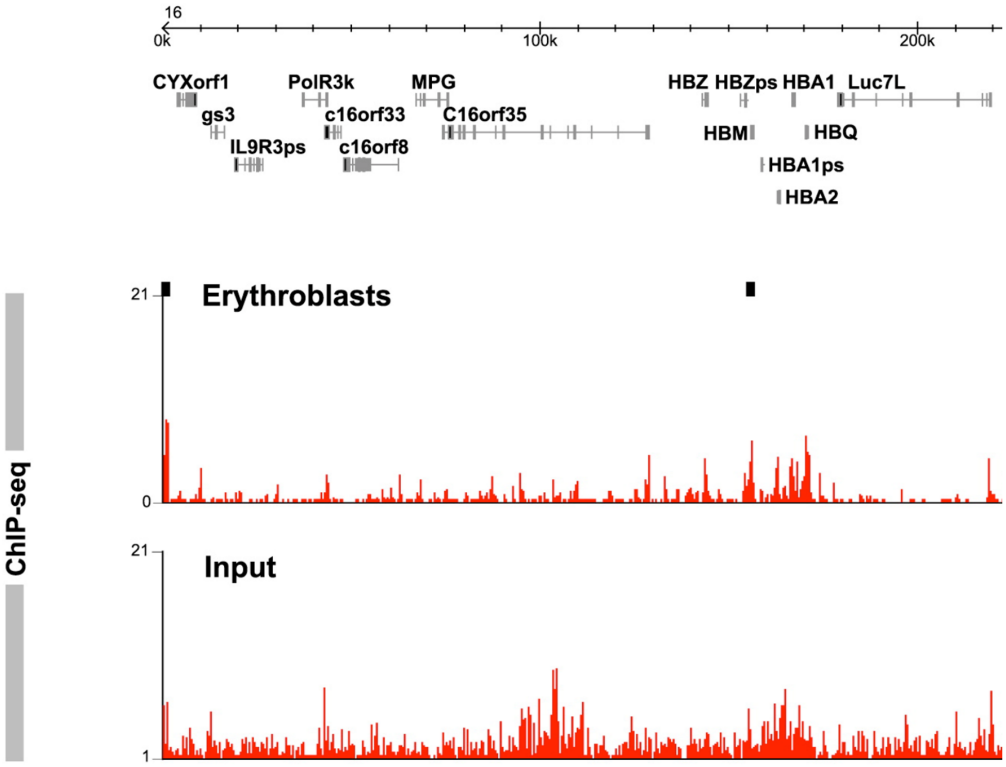


Figure 7.1: Original ATRX ChIP-seq performed in human erythroid cells by Law and colleagues at the alpha globin cluster. Figure modified from (Law et al., 2010).

7.2. Results

7.2.1. Identification of the ATRX binding sites in differentiated CD34+ cells

CD34+ is a marker associated with haematopoietic stem and progenitor cells that can go through erythroid differentiation. The extraction and differentiation of the CD34+ cells was performed by Caroline Scott following her optimised protocol and checked for quality using cytospin and FACS (see Chapter 2). After isolation from both a healthy adult peripheral blood and a leucocyte "cone" (derived from a process of leucocyte reduction filtration from the British Transfusion Service), CD34+ cells were submitted to a differentiation process and harvested at day 10. At that stage, cells express globin genes but are not yet enucleated. At day 10, cells were collected, and I performed ATRX

ChIP-seq experiments. In parallel, Damien Downes performed ATAC-seq experiments to assess the chromatin accessibility. Unfortunately, the number of cells recovered from the sample derived from the healthy adult peripheral blood was insufficient to achieve a high quality ATRX ChIP-seq dataset. By contrast, the sample derived from the leucocyte cone allowed the production of good quality ATRX ChIP-seq datasets that will be used in the present study.

As the number of cells available from the leucocyte cone was limited, I used the remaining cells to perform H3K4me3, H3K4me1, H3K27ac and H3.3 ChIP-seq experiments. The choice of these specific chromatin marks and histone variant was based on (1) the subsequent aim for chromatin annotation of active promoter and enhancers and (2) the data obtained on LCL. Indeed, these chromatin marks were the main ones affected by ATRX deficiency and ATRX has been strongly associated with H3.3. Therefore, all the subsequent analyses will be based on the same genetic background.

The selected ATRX ChIP-seq datasets showed a significant enrichment pattern and a low input background. Due to the low number of replicates compared to LCL ATRX ChIP-seq (two versus six), both peak callings were not equally stringent. Nevertheless, as the two ATRX ChIP-seq in differentiated CD34+ cells were technical replicates, ATRX peaks were retained only if they were present in the two replicates. Using this approach, I identified 9,229 ATRX binding sites in day 10 differentiated CD34+ cells (D10 dCD34+). This new ATRX ChIP-seq dataset obtained in human erythroid cells therefore resulted in the identification of approximately ten times more ATRX binding sites than what was described in Law et al. (917 ATRX binding sites) (Law et al., 2010).

The analyses of ATRX peak sizes and enrichments revealed that the ATRX enrichment across ZNF genes were not as extreme as in LCL. For instance in LCL, the ATRX enrichment across the *ZNF101* gene was so high that it was identified as a single peak. Conversely, the ATRX peak calling in D10 dCD34+ showed a better definition at such sites, which results in the presence of two distinct peaks (***Figure 7.2***). These observations, in addition to the lower stringency criteria, suggested that for a same locus, more ATRX peaks could be detected in D10 dCD34+ ATRX ChIP-seq. Thus, it would contribute to increase the number of detected peaks in this cell types compared to LCL.

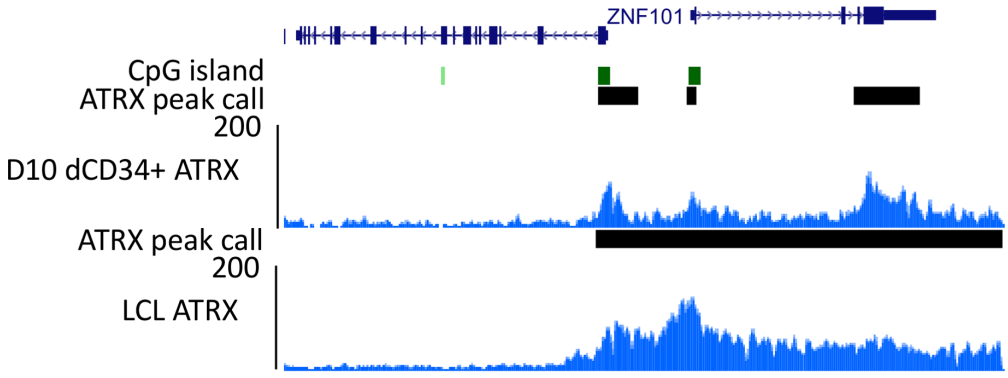


Figure 7.2: Illustration of the different resolutions of the ATRX peak calls obtained between D10 dCD34+ and LCL

7.2.2. ATRX is particularly enriched at certain repetitive elements

The analysis of the ATRX enrichment at repeats in D10 dCD34+ cells by direct mapping of the reads to the human repeatome showed that ATRX was enriched at a subset of simple repeats including (CATTT)n, (TAGA)n, (TGAG)n and (TGAA)n (***Figure 7.3***). ATRX enrichments at the large and small subunits of rRNA were even

more pronounced than in LCL, likewise for TAR1 subtelomeric repeats (***Figure 7.3***). A subset of tRNA was also enriched in ATRX over input. A slight enrichment at telomeric repeat (CCCTAA)n was detected compared to LCL with approximately twice more ATRX ChIP reads than in the input (***Figure 7.3***).

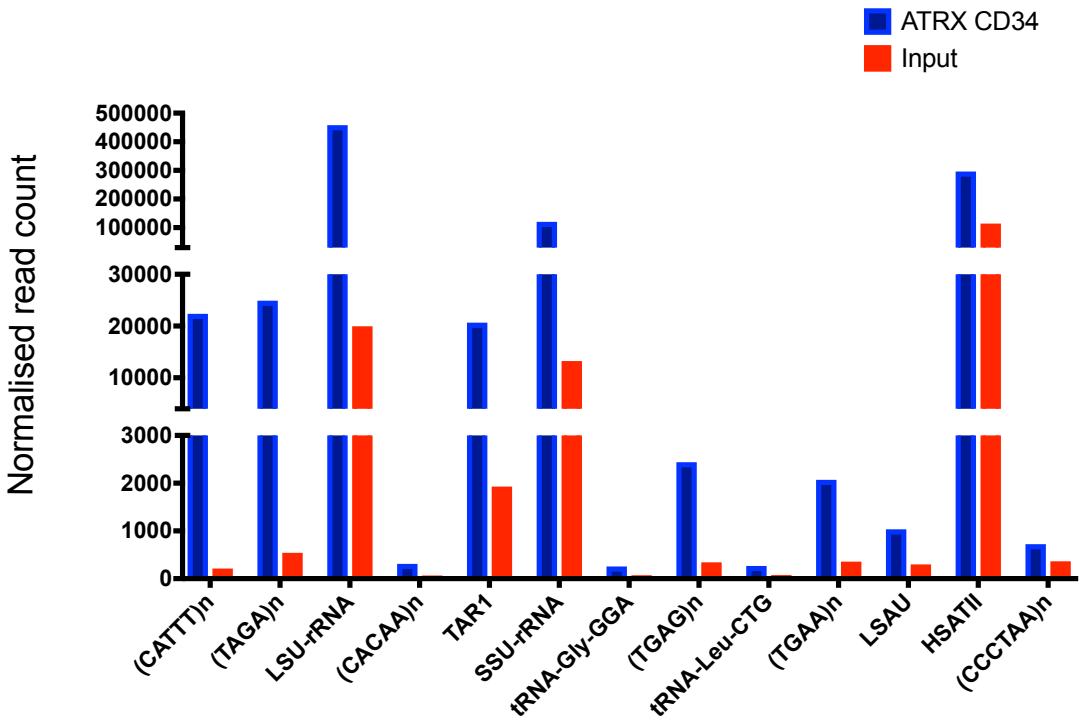


Figure 7.3: Selected representative repetitive elements that show ATRX enrichment using direct mapping of the ATRX ChIP-seqs and related inputs.

7.2.3. ATRX is enriched at CpG island-containing promoters and gene bodies

I analysed the distribution of the ATRX binding sites based on RefSeq datasets. As described in Chapter 4 for LCL, almost two third of the ATRX binding sites were intragenic (~ 21% at promoters and ~41% within the gene body) (***Figure 7.4***.a).

I then investigated the representation of predicted CpG islands at ATRX binding sites. More than 90% of the ATRX binding sites spanning promoters contained CpG islands. CpG islands were also detected at a subset of gene bodies and intergenic ATRX binding

sites (***Figure 7.4***.b). In total, almost 30% of the ATRX binding sites in D10 dCD34+ cells contained at least one CpG island (***Figure 7.4***.b). These results are very similar to the data observed at ATRX binding sites in LCL. Reciprocally, less than 10% of the predicted CpG islands spanned at least one ATRX binding sites in D10 dCD34+ cells (***Figure 7.4***.c).

Due to insufficient cell number, Bio-CAP-seq experiments were not performed in D10 dCD34+ cells.

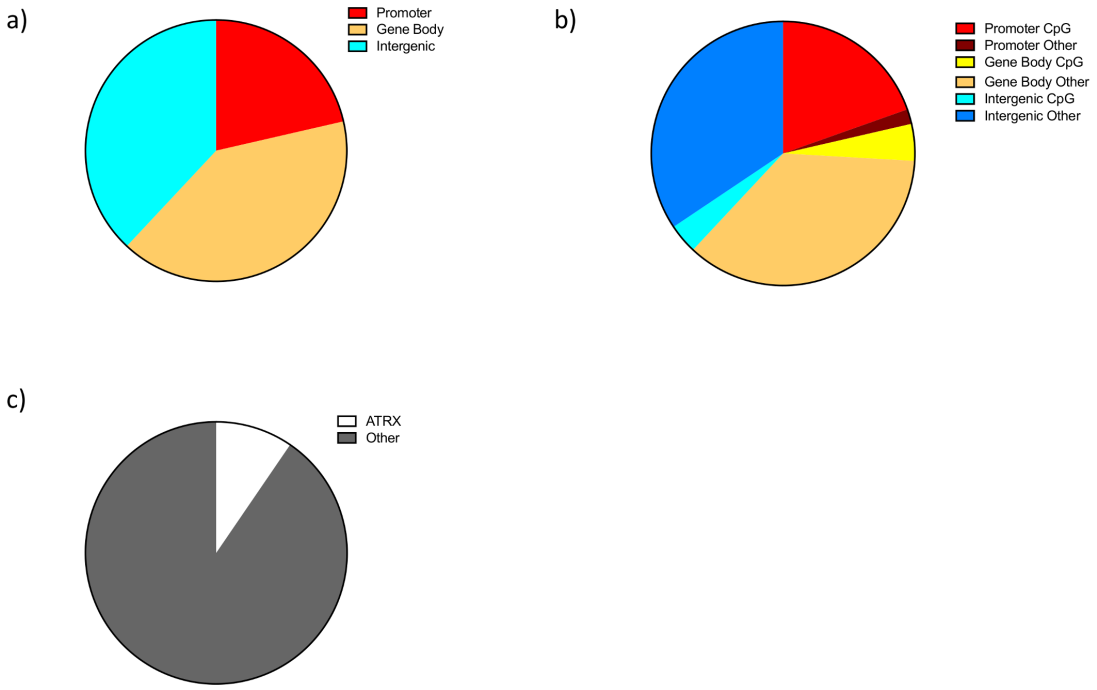


Figure 7.4: Distribution of the ATRX binding sites in D10 dCD34+ cells. a. Analysis of ATRX binding sites using the RefSeq annotation. b. Distribution of annotated CpGs islands at ATRX binding sites. c. Representation of ATRX binding sites at annotated CpG islands.

7.2.4. A third of the ATRX binding sites contains putative G4 forming sequences

An enrichment analysis for PQS showed that more that 35% of the ATRX binding sites contained at least one PQS, which represents 9,568 PQS (***Figure 7.5***.a). Almost 80% of the ATRX binding sites spanning the promoter regions contained PQS whereas it was only 25% for those in the gene body and in intergenic binding sites (***Figure 7.5***.a). Moreover, as observed in LCL, ATRX binding sites located at promoter regions contained, on average, more PQS per peak than the other two subgroups (***Figure 7.5***.b and c).

There were approximately 10% less ATRX binding sites containing PQS in D10 dCD34+ cells than in LCL (36.7% versus 46.5%, respectively). However, PQS were still present in more than a third of the ATRX binding sites in D10 dCD34+ cells and the similar pattern were observed with a higher number, on average, of PQS per ATRX peaks spanning promoters than the other subgroups (***Figure 7.5***.b). Moreover, as in LCL, ATRX occupancy was observed to only a very small number of PQS (~1.5%) (***Figure 7.5***.d).

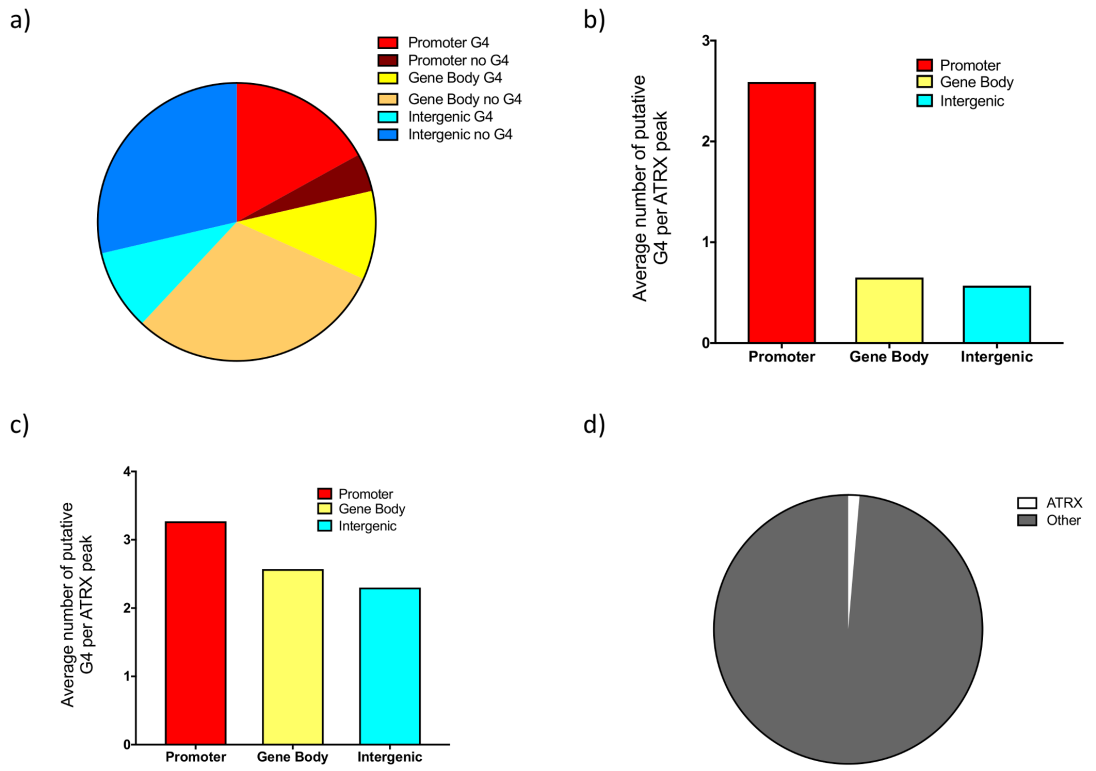


Figure 7.5: Analysis of putative G4-forming sequences at ATRX binding sites. a. Distribution of the PQS and non-PQS containing sequences at ATRX binding sites. b. Average number of PQS per ATRX binding site present at each subgroup: promoter, gene body and intergenic ATRX binding sites. c. Average number of PQS per ATRX binding site present at each subgroup only considering the ATRX binding sites spanning PQS. d. Proportion of ATRX binding sites at PQS genome-wide.

7.2.5. More than half of the ATRX binding sites contain nucleosome-depleted regions

Work done in LCL revealed that a very high proportion (~89%) of ATRX binding sites contained open chromatin regions. These results were unexpected for a protein that is mainly described for its association with heterochromatin regions. To ensure that these observations were not specific to LCL, I performed the same analysis in D10 dCD34+ cells. In contrast with the sequence-based analysis using Refseq, predicted CpG islands and PQS datasets, which remained unchanged across the different cell types of a same

species, chromatin accessibility based on ATAC-seq varies between cell types even within a same individual.

The number of open chromatin regions based on ATAC-seq datasets was more than twice lower in D10 dCD34+ cells than in LCL. However, only two biological replicates where used in the dataset from D10 dCD34+ cells instead of eight used in LCL. Besides, the proportion of shared open chromatin sites between the two cell types differed. Approximately 16,000 ATAC-seq positive regions were common to both cell types representing more than 60% of the open chromatin regions present in D10 dCD34+ cells (***Figure 7.6***.a) but only 25% of those detected in LCL (***Figure 7.6***.b).

Although less overrepresented than in LCL, more than 50% of the ATRX binding sites contained at least one open chromatin region in D10 dCD34+ cells. Open chromatin regions were present at most of the ATRX binding sites spanning promoter regions, half of the sites within gene bodies and less than half of the intergenic sites (***Figure 7.6***.c).

By contrast, less than 20% of the open chromatin regions identified in D10 dCD34+ cells were enriched in ATRX (***Figure 7.6***.d). These results suggest that the enrichment in open chromatin regions within the ATRX binding sites is unlikely to be due to an artefact of the ChIP-seq technique and that ATRX is specifically targeted at a subset of nucleosome-depleted regions.

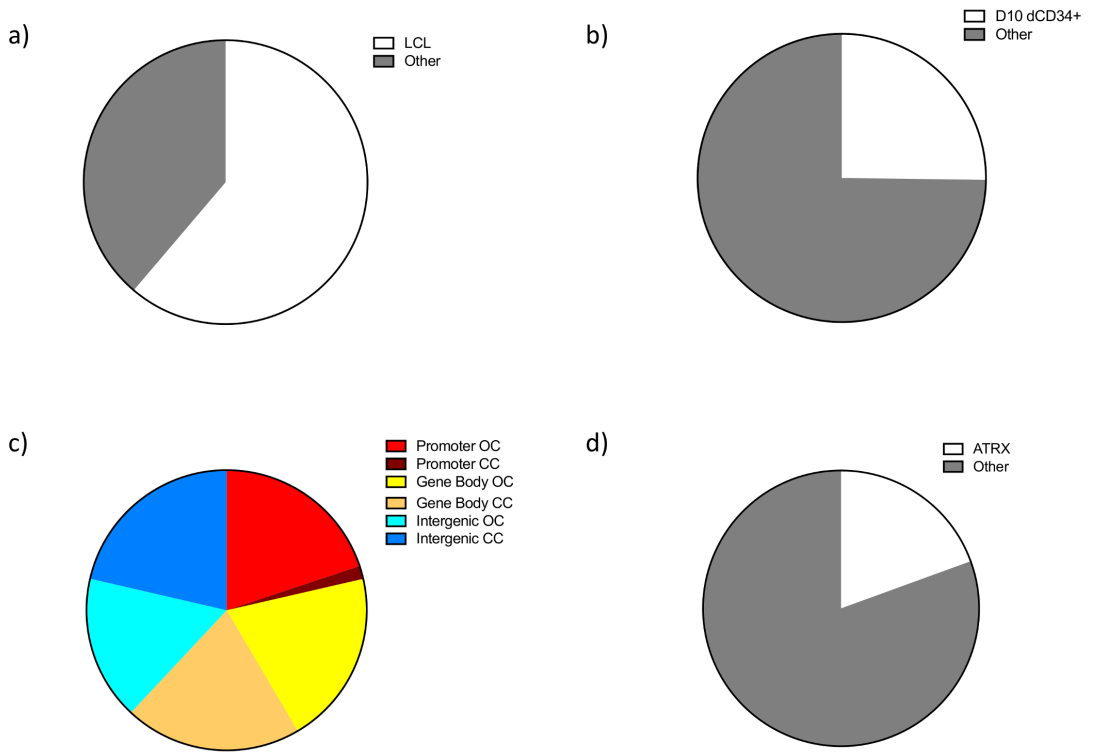


Figure 7.6: Analysis of chromatin accessibility at ATRX binding sites. a. Pie chart representing all the open chromatin regions identified by ATAC-seq in D10 dCD34+ cells. In white (LCL): proportion of open chromatin regions also present in LCL. b. Pie chart representing all the open chromatin regions identified by ATAC-seq LCL. In white (D10 dCD34+): proportion of open chromatin regions also present in D10 dCD34+ cells. c. Representation of open chromatin sites at ATRX binding sites. OC: open chromatin; CC: closed chromatin. d. Proportion of ATRX binding sites at open chromatin sites identified by ATAC-seq in D10 dCD34+.

7.2.6. Open chromatin regions that contain ATRX binding sites are enriched in active regulatory elements

The ChIP-seq experiments performed for H3K4me3, H3K4me1 and H3K27ac histone marks in D10 dCD34+ cells were sufficient to perform a superficial annotation of the ATRX binding sites and open chromatin sites using a GenoSTAN analysis.

However, only four states could be predicted: active promoters, active enhancers and enhancers (unknown state) and background. Under these conditions, approximately 33% of active promoters, 7% active enhancers and 18% enhancers were annotated leaving more than 40% of background (***Figure 7.7***.a). When I restricted the analysis to ATRX binding sites, a similar pattern was observed, with the exception of active enhancers, which showed a higher proportion (***Figure 7.7***.b). More than 40% background was observed as expected based on the proportion of ATRX binding sites without an open chromatin region (***Figure 7.7***.b).

The analysis of the open chromatin regions exclusively, did not strongly affect the annotation pattern either (***Figure 7.7***.c). Interestingly, the analysis of the ATRX binding sites that spanned at open chromatin regions showed a significant reduction in background (<10%) whereas all regulatory elements significantly increased. Indeed more than a third of the ATAC enriched ATRX binding sites were enriched in active regulatory elements (***Figure 7.7***.d).

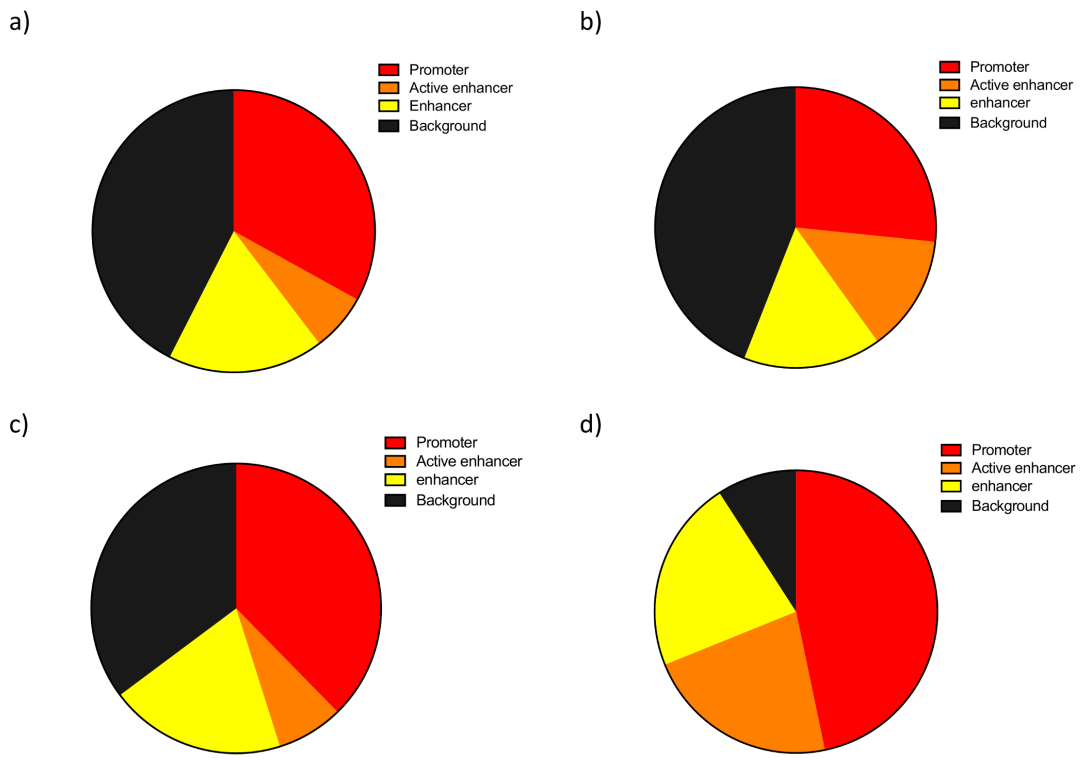


Figure 7.7: Identification of regulatory elements at ATRX binding sites in D10 dCD34+ cells using GenoSTan. a. Pie chart representing the distribution of the identified chromatin states at all the sites included in the dataset. b. Pie chart representing the distribution of the identified chromatin states at the ATRX binding sites only. c. Pie chart representing the distribution of the identified chromatin states at all the open chromatin sites included in the dataset. d. Pie chart representing the distribution of the identified chromatin states at the ATRX binding sites spanning open chromatin region only.

7.2.7. Most ATRX binding sites conserved between both cell types share a similar chromatin accessibility pattern

The pattern of ATRX binding sites is known to differ between cell types (Law et al., 2010, Nguyen et al., 2017). Therefore, I determined the proportion of ATRX binding sites conserved between the D10 dCD34+ cells and LCL. Almost 20% of them were conserved between both cells types representing 1715 peaks in D10 dCD34+ cells and 1687 peaks in LCL (***Figure 7.8***. a). The difference in peak numbers is explained by the

different sensitivities of the peak calling and ATRX enrichment between both cell types as discussed in the section "7.2.1. Identification of the ATRX binding sites in differentiated CD34+ cells". Thus, some large ATRX binding sites present in LCL spanned more than one ATRX binding sites called in D10 dCD34+ cells.

Interestingly, approximately half of the ATRX binding sites conserved between the two cell lines spanned promoter regions (***Figure 7.8***. b). The proportion of predicted CpG islands was twice larger (~60%) than in each individual cell type (~30%) (***Figure 7.8***.b). This enrichment in predicted CpG islands was not only due to the large proportion of promoter regions present at the conserved ATRX binding sites. Actually, gene body and intergenic CpG islands were also proportionally enriched at conserved ATRX binding sites compared to in each individual cell type (***Figure 7.8***.b). Thus, the ATRX binding sites that were cell type-specific were mainly categorised as within gene body or intergenic and approximately 20% of them contained predicted CpG islands (***Figure 7.8***.c).

Among the conserved ATRX binding sites between both cell types, 63% of them contain PQS (***Figure 7.8***.d).

Interestingly, almost 65% of the conserved ATRX binding sites were overlapping with conserved open chromatin regions and a minority of them showed cell type-specific open chromatin sites (< 7%) (***Figure 7.8***.e). Surprisingly, even among the conserved open chromatin regions, only a minor proportion of them (<10%) overlapped with conserved ATRX binding sites (***Figure 7.8***.f).

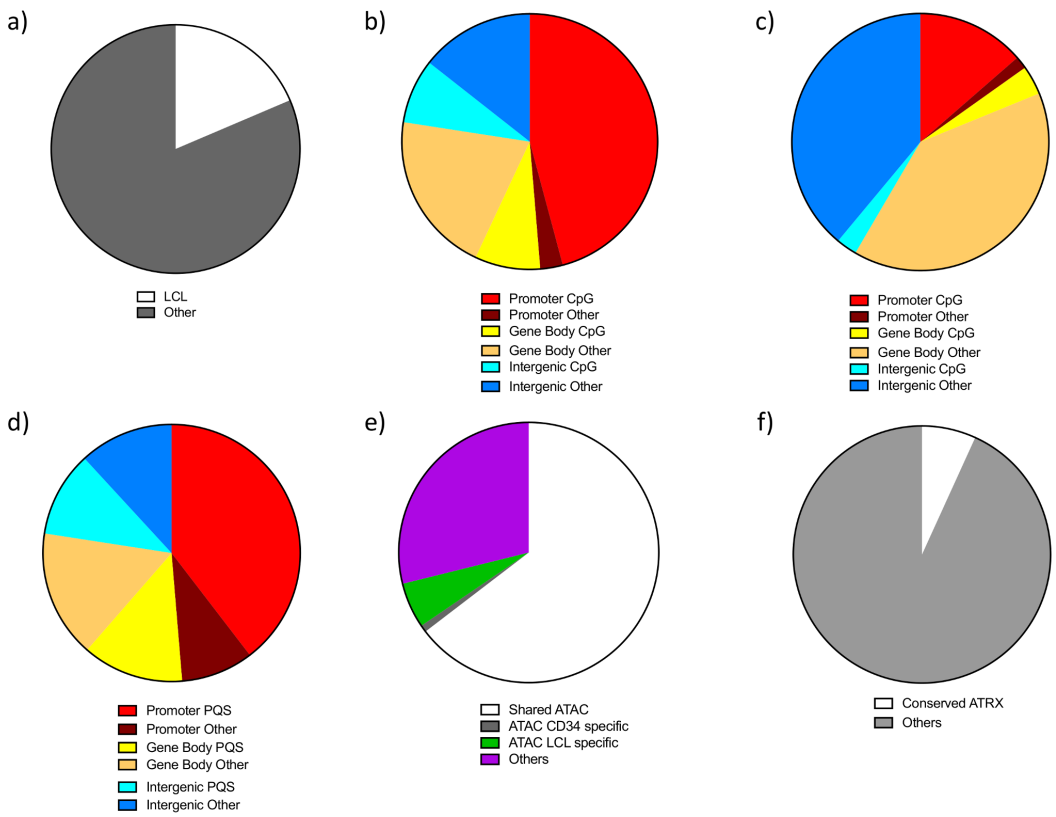


Figure 7.8: Analysis of the ATRX binding sites conserved between LCL and D10 dCD34+ cells. a. Pie chart representing the ATRX binding sites identified in D10 dCD34+ cells and highlighting the proportion of them that were conserved in LCL (white). b. Distribution of annotated CpGs islands at the conserved ATRX binding sites based on their position relative to genes. c. Distribution of annotated CpGs islands at the D10 dCD34+ cell type-specific ATRX binding sites based on their position relative to genes. d. Distribution of the PQS at the conserved ATRX binding sites based on their position relative to genes. e. Pie chart of the conserved ATRX binding sites showing the proportion of sites spanning conserved open chromatin region (Shared ATAC) and sites spanning open chromatin regions that are cell type-specific (ATAC CD34 specific and ATAC LCL specific). f. Pie chart representing all the conserved open chromatin sites and highlighting the proportion of them spanning conserved ATRX binding sites (white).

7.2.8. ATRX enrichment at ZNF genes is conserved across cell types

Similar to LCL, gene ontology analysis based on the ATRX binding sites in D10 dCD34+ cells showed a strong enrichment in terms associated with ZNF genes.

Repeating the analysis with conserved ATRX binding sites between D10 dCD34+ cells and LCL confirmed the ATRX affinity for ZNF genes in both cell types. Moreover, this affinity was not dependent of the open chromatin status as both subgroups of conserved ATRX binding sites, which overlapped with conserved open chromatin or not, showed enrichment in Znf-C2H2 ZNF genes.

7.2.9. ATRX binding sites display distinct motifs according to the cell type

Motif analysis, based on the ATRX binding sites identified in D10 dCD34+ cells, highlighted significant enrichment in GATA-like motifs (***Table 7.1***). RUNX associated motifs that were enriched at ATRX binding sites in LCL were not found among the best hits in D10 dCD34+ cells.

Rank	Motif	Name	P-value	% of Targets Sequences with Motif	% of Background Sequences with Motif
1		Gata2	1e-531	20.43%	5.47%
2		Gata1	1e-521	18.96%	4.81%
3		Gata4	1e-506	25.65%	8.64%
4		GATA3	1e-416	32.24%	14.27%

Rank	Motif	Name	P-value	% of Targets Sequences with Motif	% of Background Sequences with Motif
1		Elk4	1.00E-12	13.47%	8.35%
2		Klf4	1.00E-10	10.73%	6.54%
3		Sp1	1.00E-09	15.63%	10.62%

Rank	Motif	Name	P-value	% of Targets Sequences with Motif	% of Background Sequences with Motif
1		Elk4	1.00E-11	18.23%	11.16%
2		Fli1	1.00E-10	22.29%	14.88%

Rank	Motif	Name	p-value	% of Targets Sequences with Motif	% of Background Sequences with Motif
1		Egr2	1.00E-07	5.60%	1.91%
2		Rbpj1	1.00E-04	19.77%	13.83%

Table 7.1: Motif analysis of ATRX binding sites to identify enrichment in know motifs. Analysis performed on: all ATRX binding sites identified in D10 dCD34+ cells (First table), all conserved ATRX binding sites (Second table), the conserved ATRX binding sites spanning conserved open chromatin regions only (Third table) and the conserved ATRX binding sites that do not overlap with conserved open chromatin region (Forth table).

I further investigated the motifs found at ATRX binding sites that were conserved between both cell lines. The top enriched motifs were associated with Elk4 and the CG rich Klf4 and Sp1 (***Table 7.1***). Nevertheless, the p-values associated with this analysis were drastically higher in comparison to the previous one. Then, I focused only on the conserved ATRX binding sites overlapping with open chromatin regions present in both cell type tested. These parameters maintained the Elk4 motif as top candidate (***Table 7.1***). Finally, I repeat the analysis using only the conserved ATRX binding sites that did not overlap with conserved open chromatin sites. The results showed the G-rich Egr2 motif as top candidate (***Table 7.1***). Interestingly, G-rich ERG associated motifs were also the top candidate of the ATRX binding sites that did not overlap with open chromatin region in LCL. These results suggest that the motifs enriched at the ATRX binding sites very between cell types.

7.2.10. The ATRX enrichment at the globin clusters is cell type-dependent

Previous work showed that ATRX was enriched at the human alpha globin cluster in a cell-type dependent manner. Thus, ATRX enrichment was detected only in cells that expressed alpha globin. The main ATRX binding site was located at the ψζ-VNTR. By contrast, ATRX enrichment was detected at the NME4 VNTR in all cell types (Law et al., 2010). I could not detect a significant ATRX enrichment at NME4 in LCL partly due to mapping limitations. However, despite identical bioinformatic issues, ATRX enrichment could be identified at that same site in D10 dCD34+ cells (***Figure 7.9***). Furthermore, the associated active marks were also more enriched compared to LCL (***Figure 7.9***).

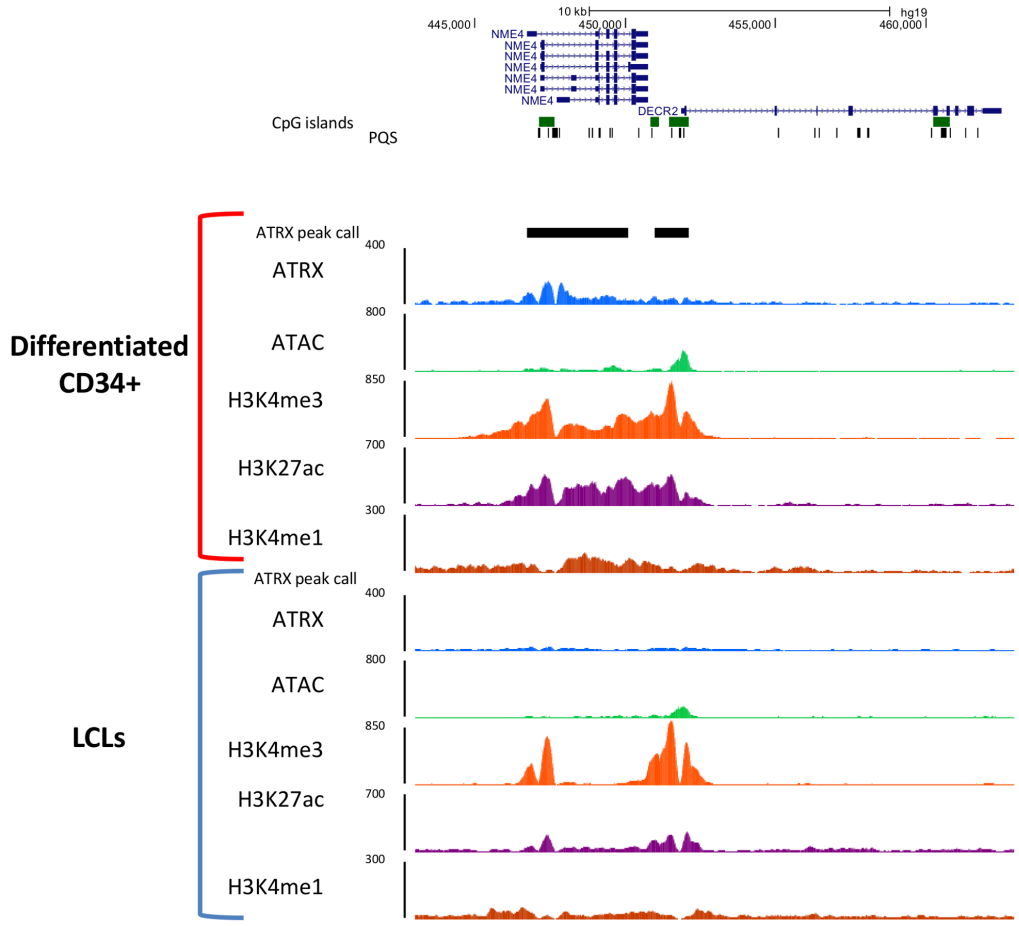


Figure 7.9: Variations in chromatin environment between D10 dCD34+ cells and LCL at the NME4 locus

The human alpha globin gene cluster stretches over a 30kb region located at the subtelomeric region of the short arm of chromosome 16. The cluster contains five genes (5'-*HBZ-HBM-HBA2-HBA1-HBQ*-3') including one embryonic ζ-gene and two alpha globin genes, *HBA1* and *HBA2*. The alpha globin genes are regulated by cis-acting regulatory elements composed of four multispecies conserved sequence enhancers R1, R2, R3 and R4 located 30-70kb upstream of the globin genes. In mouse, an extra enhancer, Rm is present (Hay et al., 2016). In erythroid-specific tissue, these enhancers present characteristic marks of active enhancers with hypersensitive sites and H3K4me1 and H3K27ac enrichments and interact with the alpha globin promoters.

Interestingly, these elements do not show equal enhancing capacity and R2 has been suggested to be the most potent regulatory element (Higgs and Wood, 2008, Hay et al., 2016).

I analysed the ATRX enrichment at the different components of the alpha globin cluster and its associated regulatory elements. In D10 dCD34+ cells, a clear ATRX enrichment at the ψζ-VNTR of the alpha globin locus was identified (***Figure 7.10***). Interestingly, ATRX enrichment was also detected across the two alpha globin genes as well as at the R1, R2 and R4 regulatory elements (***Figure 7.10***). Strikingly, the highest ATRX enrichment was observed at the R2 enhancer and spanned a PQS (***Figure 7.10***). Indeed, PQS containing regions were not limited to the ψζ-VNTR and could be found at most of the ATRX binding sites at the alpha globin locus together with its associated enhancer except R1. An H3.3 enrichment was also observed at most of the ATRX binding sites in this region with a particular enrichment at R1, R2, *HBM*, *HBA2* and *HBA1* (***Figure 7.10***).

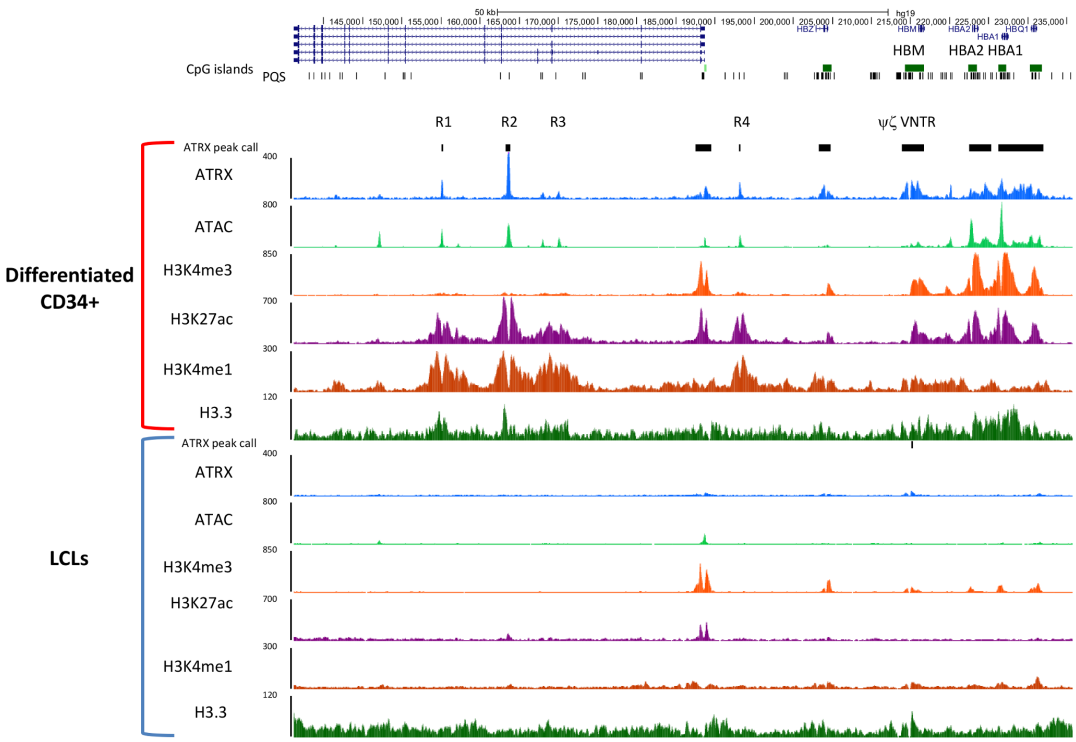


Figure 7.10: ATRX enrichment across the alpha globin cluster in D10 dCD34+ cells and LCL.

Previous studies of ATRX enrichment in human erythroid cells focused on the subtelomeric region of the short arm of the chromosome 16. This region has been of particular interest due to the alpha-thalassaemia phenotype observed in ATR-X patients. By contrast, the analysis of ATRX occupancy at the beta globin cluster, located on the short arm of chromosome 11, remains largely under described.

The beta globin cluster is regulated by five regulatory elements composing the beta locus control region (LCR). Unlike the duplicated alpha globin genes, there is only one beta globin gene per allele (Higgs et al., 2005).

In LCL, the beta globin cluster was largely depleted in ATRX, however, the ATRX occupancy pattern was very different in D10 dCD34+ cells. Indeed, in D10 dCD34+ cells, significant ATRX enrichments at the beta globin gene and within its LCR were detected (***Figure 7.11***). These regions were also associated with some H3.3 enrichment

and active chromatin marks (***Figure 7.11***). Although PQS were detected, their enrichment was not comparable to the one observed at the alpha globin cluster. Similarly, none of the enhancers contained within the LCR region showed ATRX enrichment comparable to R2 at the alpha globin cluster.

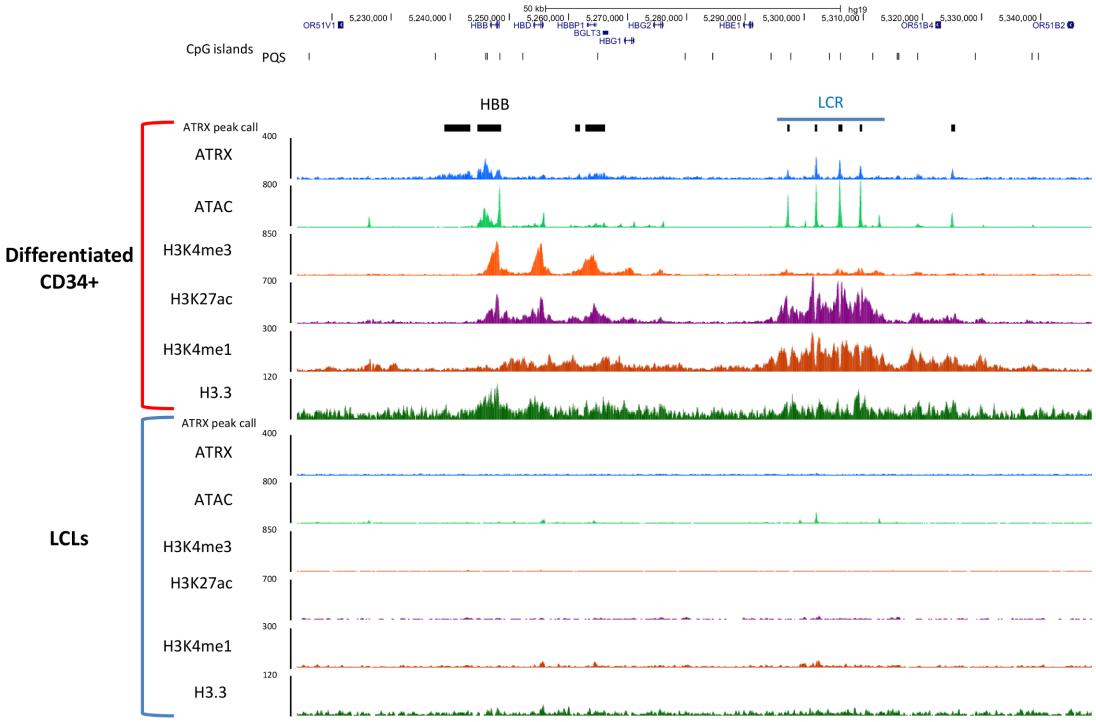


Figure 7.11: ATRX enrichment across the beta globin cluster in D10 dCD34+ cells and LCL.

7.3. Discussion

The genome-wide analysis of ATRX binding sites in an erythroid model derived from human primary cells has been of particular relevance as mutations in ATRX can cause alpha-thalassaemia. Previous work has highlighted the presence of ATRX enrichment at one particular site of the alpha globin cluster, the ψζ VNTR (Law et al., 2010). Using a newly optimised ATRX ChIP-seq protocol offering a great increase in sensitivity, I investigated the genome-wide distribution of ATRX in human differentiated CD34+ cells. I have supplemented the study by performing ChIP-seq experiments of histone marks associated with active chromatin in the same genomic background. The optimised ATRX ChIP-seq protocol increased by a factor ten the number of ATRX binding sites identified in human erythroid cells resulting in more than 9, 000 sites genome-wide.

The data obtained in human differentiated CD34+ cells supported most of the observations linked to the analyses in LCL. Indeed, the affinity of ATRX for ZNF genes was confirmed across cell types. These data further support the recent work done by Valle-Garcia and colleagues using two other cell lines (Valle-García et al., 2016). The ATRX affinity for rDNA repeats and subtelomeric TAR1 repeats were also conserved across cell types. A third of the ATRX binding sites spanned either the promoter regions or gene bodies. In both LCL and D10 dCD34+ cells, the majority of the promoters enriched in ATRX contained CpG islands and in total 30% of the ATRX binding sites contained predicted CpG islands. Although lower than in LCL, more than 35% of ATRX binding sites in D10 dCD34+ cells contained PQS. Finally, in both cell types, only a small percentage of the total PQS were enriched in ATRX (~1.5%). All these similarities between these two cell types reinforce the findings of the study.

The low propensity for ATRX enrichment at PQS may be explain by the different DNA structure and putative G4 stability. Indeed, the fact that a sequence was predicted to be a PQS does not guarantee the effective presence of G4 structures at that site in any particular cell type. G4 structures may form in a cell-type dependent manner (Hänsel-Hertsch et al., 2016). Furthermore, some PQS may not be able to form G4 *in vivo* or may not produce stable G4. If ATRX were indeed recruited to such non-canonical DNA structures, its target sites would be restricted to regions where stable G4 could be found, therefore representing only a restricted proportion of the PQS. It is also conceivable that ATRX may be sensitive to only a subset of G4 structures. Nevertheless, such hypotheses would not explain why only a subset of ATRX binding sites contained PQS. If G4 structures were the main ATRX target, one might expect the presence of PQS at most of the ATRX binding sites. As this is not the case, G4 structures might not be the only factor involved in ATRX recruitment/stabilisation. Indeed, it is known that ATRX recruitment/stabilisation can be dependent on ATRX's interaction with H3K9me3 for instance (Eustermann et al., 2011). Alternatively PQS predictions may not be sensitive enough to detect specific G4 structures recognised by ATRX.

Data obtained with D10 dCD34+ cells also supported the observation that ATRX binding sites were enriched in open chromatin regions and active regulatory elements suggesting that ATRX is a new player to consider at euchromatin sites. Moreover, most of the ATRX binding sites conserved across the two cell types spanned open chromatin regions shared by both cell types and contained a majority of ATRX binding sites spanning promoters. Considering only the ATRX binding sites conserved between LCL and D10 dCD34+ cells, approximately double the proportion of ATRX binding sites containing PQS and maintain the enrichment in ZNF genes. However, motif analyses

performed in D10 dCD34+ cells revealed that the detected motifs were associated with the cell types. Thus, motif analyses performed on ATRX binding sites identified in LCL showed enrichment in RUNX associated motifs. RUNX factors, especially RUNX3 expression, are important for the regulation of LCL growth (Spender et al., 2005, Whiteman and Farrell, 2006). By contrast, similar analyses performed in D10 dCD34+ cells highlight the enrichment of GATA-like motifs. GATA transcription factors are involved in development and differentiation of diverse tissues including the haematopoietic system (Lentjes et al., 2016). Besides, GATA1 is one of the key transcription factors recruited to the alpha and beta globin enhancers during erythroid differentiation. These observations suggest that ATRX could be recruited to a subset of cell type-specific enhancers when they would be active. Motif analyses focusing on all conserved ATRX binding sites or conserved ATRX binding sites that contain conserved open chromatin regions were not really conclusive and the p-values were not highly significant. However, motif analysis based on conserved ATRX binding sites that were depleted in conserved open chromatin regions gave consistent results as those observed in LCL. Although the p-values were not highly significant, they showed the presence of GC-rich ERG-associated motifs. All together, these observations suggest that (1) ATRX may be recruited/stabilised through distinct mechanisms between open and closed chromatin regions and (2) ATRX recruitment/stabilisation at open chromatin regions is unlikely to be associated with the recognition of a single motif conserved across cell types.

Despite the presence of mapping artefacts, ATRX ChIP-seq in D10 dCD34+ cells highlighted the presence of the protein at the *NME4* VNTR as previously reported. Beyond the mapping complexity of this region, the detection of ATRX enrichment at

NME4 VNTR in erythroid cells, but not in LCL, may be associated with the increase of *NME4* expression in this cell type. The increase in *NME4* expression was also supported by the increase in active chromatin marks and previous work (Lower et al., 2009).

Importantly, using the optimised ATRX ChIP-seq in D10 dCD34+ cells, I confirmed the ATRX enrichment in a cell type-dependent manner at the ψζ VNTR within the alpha globin cluster. Moreover, the increased sensitivity resulted in the clear enrichment of ATRX across the alpha globin genes and at their R1, R2 and R4 enhancers. The strongest ATRX enrichment was detected at the R2 enhancer, which is supposed to have the greatest influence on alpha globin expression (Higgs and Wood, 2008, Hay et al., 2016). This result may have a bearing on why the ATR-X mutations lead to alpha-thalassaemia. In addition, the regions were highly enriched in PQS including at the R2 and alpha globin genes. Further work on differentiated CD34+ cells from ATR-X patients could lead to important discoveries. Additionally to these new findings, ATRX enrichment was also observed at the beta globin cluster including the beta globin gene and the LCR region. Thus these new observations raise further questions regarding the difference in regulation of these two clusters in ATRX deficient cells.

Alpha and beta globin clusters are located at different genomic regions. Actually, the human alpha globin gene cluster is located within the gene rich, early replicative subtelomeric region of the short arm of chromosome 16. The cluster contains five genes including two alpha globin genes, HBA1 and HBA2. Despite the presence of CpG island promoters, the alpha globin genes remain unmethylated in non-erythroid cells and are repressed by the polycomb repressive complex (PRC) 2 associated with H3K27me3 enrichment (Garrick et al., 2008). By contrast, the beta globin locus is found in a late replicative, relatively gene sparse region on chromosome 11. Unlike the

duplicated alpha globin genes, there is only one copy of the beta globin gene. In non-erythroid cells, the beta globin promoter is methylated and does not show PRC2 and H3K27me3 enrichments (Garrick et al., 2008). In addition, the beta globin cluster was not as enriched in PQS as the alpha globin cluster. Analyses in LCL showed that ATRX binding sites were not systematically associated with changes in gene expression and in chromatin environment in ATRX deficient cells suggesting the presence of pathway redundancy. Supporting this hypothesis, the LCL derived from ATR-X patients maintained H3.3 enrichment similar to normal at many ATRX binding sites. Furthermore, the experiments carried out in LCL showed that ATRX was enriched at a number of target genes and/or associated enhancers, which appeared unaffected in their expression in ATRX deficient cells such as MYC, TAB2 and ZNF554 for instance. Therefore, It is possible that the beta globin cluster remains unaffected in ATRX deficient cells due to the presence of a redundant pathway. This may not be the case at all ATRX binding sites at the alpha globin cluster. The comparison of the macroH2A occupancy at these sites may also be highly informative in this context.

Chapter 8. General discussion and conclusions

In 1995, *ATRX* mutations were identified as the underlying cause of the ATR-X syndrome. Ever since then, they have been shown to be associated with defects in essential cellular processes including DNA replication, transcription and maintenance of chromatin integrity (Gibbons et al., 1995, Clynes et al., 2014, Wong et al., 2010, Berube et al., 2005, Garrick et al., 2006). Over two decades of investigations have emphasised the association of ATRX with heterochromatic regions including pericentric heterochromatin, telomeres and rDNA repeats (McDowell et al., 1999, Gibbons et al., 2000, Wong et al., 2010, Udugama et al., 2018). Despite the increasing interest in this chromatin remodeling factor, our knowledge of the distribution of ATRX and its function within euchromatin in the human genome remains very limited.

In this context, the aims of this thesis have been to (1) overcome the limitations of the ATRX ChIP-seq protocol in order to produce a reliable analysis of the genome-wide distribution of ATRX in human derived cells, (2) increase the panel of DEG associated with *ATRX* mutations taking advantage of LCL derived from ATR-X patients and their unaffected relatives, (3) assess the changes in chromatin environment at DEG linked to ATRX deficiency and (4) gain insights into the role of ATRX in the regulation of gene expression.

8.1. Optimisation of the ATRX ChIP-seq protocol places ATRX in euchromatin

The optimisation of the ATRX ChIP-seq protocol (Truch et al., 2018) has provided a new high-resolution map of the genome-wide distribution of the ATRX binding sites in LCL and erythroid cells. Beyond confirming the presence of ATRX at repetitive

DNA and PQS sequences, these datasets uncovered the presence of a significant number of ATRX binding sites associated with regulatory elements. Moreover, most of these elements, including alpha globin enhancers in erythroid cells, display active chromatin marks. Although ATRX has been mainly associated with heterochromatin for years, the present study identified that almost 90% of its binding sites contained nucleosome-depleted regions in LCL. As previously noted, the distribution of ATRX varies between cell types (Law et al., 2010, Voon et al., 2015). Indeed, the enrichment of ATRX across the alpha globin cluster in erythroid cells strongly contrasts with the pattern in LCL. Although more analyses are required to achieve firm conclusions, our data support the hypothesis that ATRX occupancy, at least at a subset of targets, is associated with the activity of the locus. Interestingly, a large majority of the ATRX binding sites conserved across erythroid cells and LCL appeared to share the same chromatin accessibility characteristics; either spanning open chromatin regions present in both cell types or spanning regions completely depleted in open chromatin in both cell types. By contrast, most of the ATRX binding sites associated with cell type-specific open chromatin regions, were also cell type-specific. These results suggest that the chromatin state plays a role in ATRX occupancy. Furthermore, analysis of these high-quality ATRX ChIP-seq data highlights a clear affinity of ATRX for a particular class of zinc finger, the C2H2-ZNF genes, in different cell lines, supporting a recent study from Valle-García and colleagues (Valle-García et al., 2016).

Taken together these results place ATRX at both heterochromatin and euchromatin, which significantly increases the repertoire of ATRX binding sites in human.

8.2. Pathogenic ATRX mutations are associated with limited changes in gene expression

Gene expression analysis performed in LCL shows that only a restricted number of genes (234 in total) are differentially expressed between ATR-X patients and unaffected individuals and only about a hundred of these show an absolute fold change greater than 0.5. Interestingly, no significant differences in gene expression are noticed between the ATRX mutations affecting the ADD or the ATPase domain, with the exception of the *NME4* gene. However, the reasons underlying the observed susceptibility of *NME4* to mutations associated with the ATPase domain remain unknown.

The DEG are enriched in putative transcription factors and co-factors, including KRAB-containing C2H2-ZNF genes, which may result in indirect effects on gene expression.

One interesting observation is that the special relationship between ATRX binding sites and the DEG appears to be different in downregulated versus upregulated DEG. Unlike upregulated DEG, the closest ATRX binding site associated with each downregulated DEG is more likely to be within the DEG rather than in its flanking regions. Surprisingly, despite ATRX's affinity for ZNF genes, only seven of them are differentially expressed with an absolute fold change > 0.5. These are all downregulated in patient-derived LCL and all belong to the KRAB-containing C2H2-ZNF family. Comparative analyses between differentially expressed and unaffected KRAB-containing C2H2-ZNF genes failed to determine the underlining cause of this phenomenon. Nevertheless, these data show that a high level of ATRX enrichment at a locus does not necessarily mean its expression will be perturbed when ATRX is mutated.

8.3. ATRX mutations are associated with changes in chromatin environment at enhancers and promoters linked to DEG

The relevance of the findings that associate ATRX with a subset of active regulatory elements is further supported by the analysis of the chromatin environment at DEG. Indeed, a subset of DEG display ATRX enrichment at their promoter and/or associated putative enhancers. In patient-derived cells, some of these sites show changes in chromatin environment including chromatin accessibility and enrichment in H3K4me3, H3K27ac and H3.3. However, Capture-C experiments indicated only few changes in the three dimensional chromatin interactions at these sites. Curiously, some regulatory elements, enriched in ATRX and associated with upregulated DEG, display similar (or even increased) levels of H3.3 in ATRX mutant cells. These observations suggest that at some sites either ATRX is not associated with H3.3 deposition or there is some pathway redundancy.

A subset of downregulated DEG shares similar characteristics including a promoter enriched in ATRX and containing CpG clusters and PQS. In ATRX deficient cells, some of these DEG have reduced marks of active chromatin and in some cases, decreased level of H3.3. Some of them (e.g. *ATF7IP2*) also show partial DNA methylation at the promoter as previously observed at the *NME4* locus. However, unlike *NME4*, no VNTR was detected at the tested DEG. At some downregulated DEG (e.g. *PBX4*), ATRX is enriched at the PQS containing enhancer rather than the promoter and in ATR-X cases there is a reduction in active chromatin marks and in H3.3 enrichment at these enhancers. These results suggest that a similar underlying mechanism may occur at these promoters and enhancers leading to the downregulation of these DEG in patients.

8.4. ATRX is enriched at the ψζ VNTR, *HBA1* and *HBA2,* and their associated enhancers in erythroid cells

The optimisation of ATRX ChIP-seq reveals a much more complex pattern of ATRX enrichment across the alpha globin locus than previously noticed in erythroid cells (Law et al., 2010). Indeed, ATRX occupancy across *HBM*, *HBA1* and *HBA2* as well as at the alpha globin enhancers, including R2, indicates several new sites where ATRX may play a role in the regulation of the alpha globin expression. The presence of PQS at the ψζ VNTR, *HBA1*, *HBA2* and R2 associated with downregulation of alpha globin expression in most of the ATR-X patients correlates with the patterns observed at the downregulated DEG described above. These results raise the possibility that the downregulation of the alpha globin might involve features associated with the enhancer, promoter and/or gene body in addition to the presence of the ψζ VNTR. Future studies using D10 dCD34+ derived from patients with ATR-X syndrome should help solve the conundrum of why they develop alpha-thalassaemia.

8.5. Does ATRX deficiency lead to the loss of epigenetic memory?

ATRX enrichment at PQS-containing regulatory elements was found to be associated with downregulation of DEG in some instances. An attractive hypothesis, which potentially explains this association, would be G4-induced loss of epigenetic memory in cases of ATRX deficiency. Studies by the Sale group suggest that, in specific circumstances such as nucleotide depletion or loss of G4 unwinding helicases, G4 leads to replicative stress and the subsequent perturbation of parental histones recycling resulting in the loss of epigenetic memory. Furthermore, the perturbation of the epigenetic memory associates with aberrant gene expression (Papadopoulou et al., 2015, Sarkies et al., 2012). ATRX has been showed to bind G4 *in vitro* and ATRX

deficiency has been associated with defects in DNA replication and the DNA damage response (Clynes et al., 2014, Leung et al., 2013, Law et al., 2010). Thus, it is plausible that the downregulation of DEG displaying ATRX enrichment at their PQS, may be the consequence of replicative stress caused by G4 stabilisation in ATRX deficient cells. Under normal circumstances, ATRX in association with DAXX may prevent/resolve G4, potentially via H3.3 deposition/nucleosome remodeling, therefore avoiding replicative stress. In ATRX deficient cells, G4 may persist and impede DNA replication. The resulting replication stress and uncoupling of replicative helicase and polymerase may impede the recycling of the parental histones, which are carriers of active chromatin marks such as H3K4me3 or H3K27ac. Upon replication, the locus would therefore display a reduction of active chromatin marks that may associate with downregulation of DEG. A recent study showed that PrimPol deficiency was associated with G4-induced loss of epigenetic memory. Interestingly, the severity of the gene repression at the studied locus varied and in the most severe cases was associated with DNA methylation (Schiavone et al., 2016). These observations echo the findings of the present study and might explain the variation in alpha globin and *NME4* expressions observed in patients (Law et al., 2010). Furthermore, preliminary experiments performed in our group by Anthony Cheong failed to produce ATRX/PrimPol double mutants potentially suggesting that at least one of these two proteins may be required for cell survival. Alternatively, it could be also possible that the decreased expression of the DEG may be the cause and not the result of the reduced enrichment in active chromatin marks and H3.3 observed in patient-derived LCL.

Most of the upregulated DEG in ATRX deficient cells are not associated with H3.3 depletion or changes in repressive chromatin marks, which hints at the possibility that this effect may be unrelated to loss of epigenetic memory. Therefore, it is plausible that ATRX may not be required to solve/avoid the persistence of G4 at such sites. Indeed, ATRX alone is not able to resolve G4 structures *in vitro* so it is often hypothesised that ATRX-DAXX-H3.3 complex may be involved in the process (Clynes et al., 2013, Clynes et al., 2014). Several studies have suggested the existence of ATRX functions that are independent of DAXX or not associated with H3.3 deposition. Thus, gene expression analysis comparing ATRX and DAXX deficiencies highlighted a subgroup of ATRX specific gene targets for which expression was not affected by DAXX deficiency (Hoelper et al., 2017). Work focusing on ZNF genes suggested that ATRX, but not DAXX, was involved in the maintenance of the chromatin integrity at the ZNF 3' exons (Valle-García et al., 2016). Other studies focussing on mouse forebrain suggested that ATRX is involved in nucleosome positioning rather than H3.3 deposition (Kernohan et al., 2014). ATRX has already been associated with transcription repression and thus ATRX may act as a moderate repressor, which would explain the de-repression of these DEG in patient-derived cells (Tang et al., 2004).

8.6. ATRX-dependent deposition of H3.3 at euchromatin regions?

Previous work in mouse ES cells implied that ATRX was not required for the deposition of H3.3 at transcribed genes (Goldberg et al 2010). Nevertheless, some studies have associated ATRX with H3.3 deposition in at least a subset of genes, including aPar genes and imprinted loci (Levy et al., 2014, Voon et al., 2015). In the current study, a change in H3.3 enrichment was detected at a subset of promoter regions (e.g. ATF7IP2) and putative enhancer sites (e.g. PBX4) of downregulated DEG in patient-derived LCL.

However, in most other DEG, H3.3 enrichment was maintained or increased at their associated regulatory elements. Together with the presence of ATRX at putative active enhancers and TFBS, these observations suggest a role of ATRX in the deposition of H3.3 at some euchromatin sites, which is critical at some of them, but at many others, pathway redundancy for H3.3 deposition should be present. Pathway redundancy between HIRA and ATRX complexes might explain why the TFBS remained unaffected upon losses of either HIRA or ATRX (Goldberg et al., 2010). In that case, the impairment of either of these complexes individually would not result in significant changes at redundant target sites and only depletion of both, would lead to observable effects. Supporting this hypothesis, a study in Drosophila demonstrated redundancy between the ATRX homolog Xnp and HIRA. Additionally, the Xnp/HIRA double mutants were lethal (Schneiderman et al., 2012).

8.7. ATRX recruitment: different targets, different mechanisms?

Our results support previous observations indicating that ATRX is recruited to different loci by a number of different mechanisms. Indeed, the H3K9me3/H3K4me0 combinatorial readout appears unlikely to be the signal that explains binding of ATRX to active regulatory elements (Iwase et al., 2011, Eustermann et al., 2011). A large number of ATRX binding sites contained nucleosome-depleted regions; nevertheless, open chromatin is not sufficient to explain the recruitment/stabilisation of ATRX, as ATRX is absent from most regions of open chromatin. Likewise, for PQS, only a proportion (between 35% and 50% depending of the cell type) of the ATRX binding sites contained such sequences. Furthermore, it is clear that ATRX is not targeted to most PQS as only 1.5% of them were enriched in ATRX, independently of the cell type.

The ATRX ADD domain was shown to display a combinational readout selecting for H3K9me3/H3K4me0 (Eustermann et al., 2011). Therefore, ATRX occupancy at nucleosome-depleted regions, enriched in H3K4me3 marks, appears paradoxical. However, works in Drosophila showed that Xnp interacts with nucleosome-depleted regions and is displaced upon nucleosome deposition (Schneiderman et al., 2012, Schneiderman et al., 2009). Although Xnp does not contain an ADD domain, these data offer a supportive insight regarding ATRX dynamics at such regions. Furthermore, ATRX binds to G4, has DNA translocation activity and displays stimulation of its ATPase activity by DNA suggesting that ATRX can interact directly with DNA (Argentaro et al., 2007, Xue et al., 2003, Law et al., 2010). It is likely that ATRX binding at such regions would be very dynamic. Based on the studies on Xnp in Drosophila, an exciting hypothesis might be that ATRX is recruited to a subset of nucleosome-depleted regions and is involved in the protection of DNA exposure and potentially associated with the deposition of H3.3 (***Figure 8.1***). Several means of recruitment may be considered such as the presence of non-canonical DNA structures and DNA-RNA hybrids including G4 or R-loops. Interactions with transcription factors or co-factors could be also considered regarding the high enrichment in TFBS at ATRX binding sites. It is possible that ATRX binding at nucleosome-depleted regions serves as a binding platform for the recruitment of DAXX-H3.3 complex or alternatively, DAXX may be already associated with ATRX. Alternatively, H3.3 deposition may also be linked with HIRA complex (***Figure 8.1***).

Once newly nucleosomes are assembled and/or display the PTM associated with their chromatin environment (e.g. H3K4me3 at promoters), ATRX may be displaced until the next round of nucleosome-depletion associated for instance with transcription or other features of DNA (***Figure 8.1***) (Schneiderman et al., 2009, Schneiderman et al.,

2012). Transcription may result in G4 or R-loops formation or simply expose DNA to the nuclear environment resulting in ATRX recruitment/stabilisation, and a novel cycle of H3.3 deposition and so on (***Figure 8.1***). This process would be independent of HIRA complex recruitment and function and may result in pathway redundancy at a subset of genomic regions as identified by works in Drosophila (Schneiderman et al., 2009, Schneiderman et al., 2012).

This hypothesis is not incompatible with the association of ATRX with H3K9me3/H3K4me0 or other factors such as HP1 at other sites. However, it paints a picture of ATRX recruitment and enrichment at diverse spatial and temporal regions of the genome and emphasises how the structural features of ATRX might cooperate to target the protein to various chromatin environments (***Figure 8.1***).

8.8. Conclusions

In this study, I have produced an optimised ATRX ChIP-seq protocol providing high-quality datasets. Using this resource, I identified new ATRX binding sites in human samples and unveiled the presence of ATRX at a subset of promoters and enhancers, most of which contain active chromatin marks. Thus, this work reveals the distribution of ATRX at euchromatin, including rapid nucleosome turnover regions, in human cells. Furthermore, I have clarified the distribution of ATRX enrichment at the alpha and beta globin clusters in erythroid and non-erythroid cells. In addition to the ψζ VNTR, ATRX was also enriched across the *HBA1* and *HBA2* genes and at their associated enhancers, including R2, in erythroid cells. Additionally, I have enlarged the pool of identified genes affected by pathogenic ATRX mutations in LCL and characterised the associated changes in the chromatin environment. These results suggest a role for ATRX in the deposition of H3.3 at a subset of regulatory elements. Ultimately, this work enlarges

our understanding of ATRX targets and recruitment and provides new insights regarding the role of ATRX in the maintenance of the epigenetic integrity at a subset of targets.

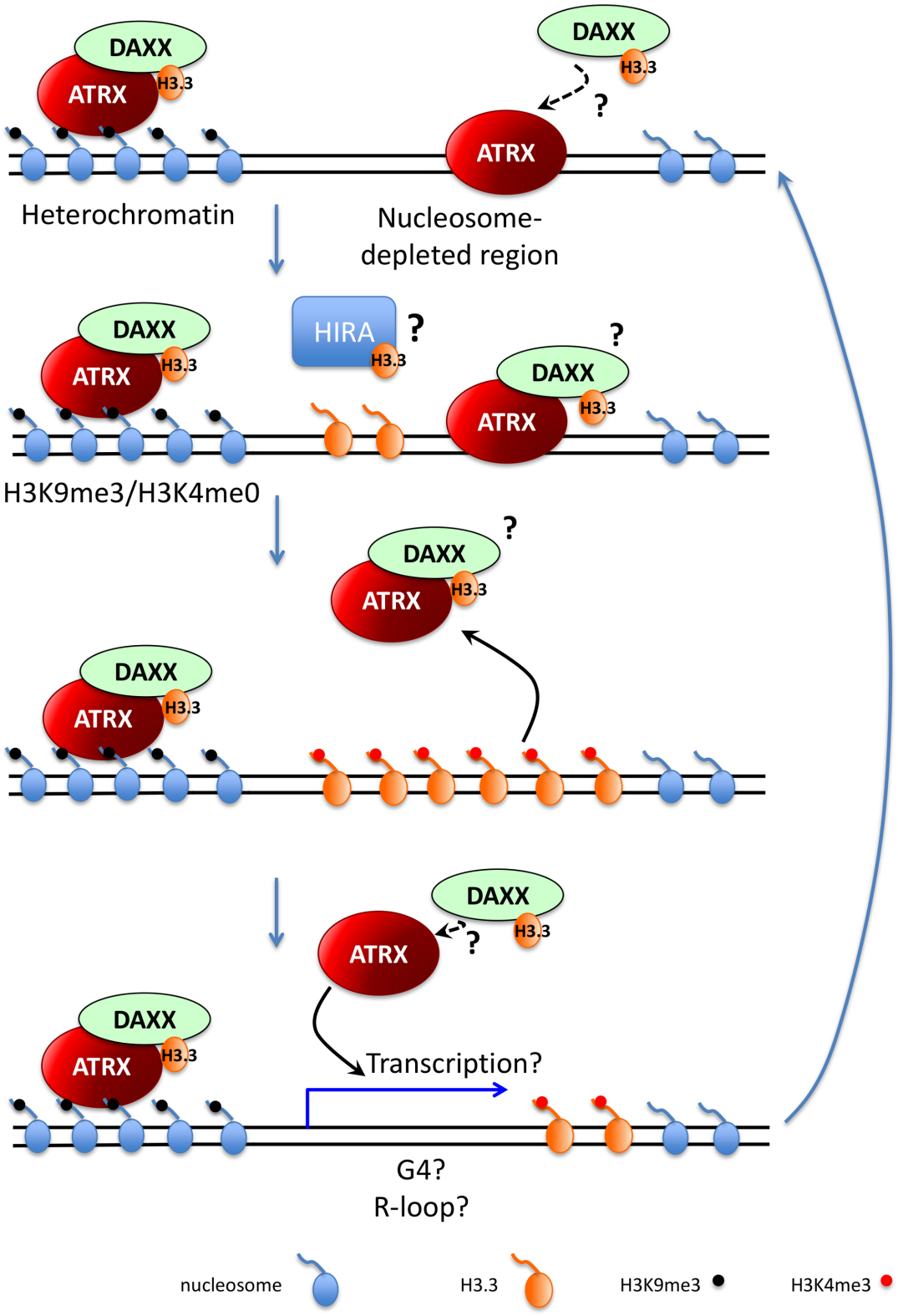


Figure 8.1: Hypothesis of ATRX interactions and dynamics depending on the change in chromatin environment.

References

ABREU, A. P., MACEDO, D. B., BRITO, V. N., KAISER, U. B. & LATRONICO, A. C. 2015. A new pathway in the control of the initiation of puberty: the MKRN3 gene. *Journal of molecular endocrinology,* 54**,** R131-R139.

AGOSTON, Z., LI, N., HASLINGER, A., WIZENMANN, A. & SCHULTE, D. 2012. Genetic and physical interaction of Meis2, Pax3 and Pax7 during dorsal midbrain development. *BMC developmental biology,* 12**,** 10.

ANDREWS, S. 2010. FastQC: a quality control tool for high throughput sequence data.

ANISIMOV, S. V. 2012. A prevalence of imprinted genes within the total transcriptomes of human tissues and cells. *Molecular biology international,* 2012.

ARGENTARO, A., YANG, J.-C., CHAPMAN, L., KOWALCZYK, M. S., GIBBONS, R. J., HIGGS, D. R., NEUHAUS, D. & RHODES, D. 2007. Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD) domain of the chromatin-associated protein ATRX. *Proceedings of the National Academy of Sciences,* 104**,** 11939-11944.

AYYANATHAN, K., LECHNER, M. S., BELL, P., MAUL, G. G., SCHULTZ, D. C., YAMADA, Y., TANAKA, K., TORIGOE, K., RAUSCHER, F. J. J. G. & DEVELOPMENT 2003. Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. 17**,** 1855-1869.

BADENS, C., MARTINI, N., COURRIER, S., DESPORTES, V., TOURAINE, R., LEVY, N. & EDERY, P. J. A. J. O. M. G. P. A. 2006. ATRX syndrome in a girl with a heterozygous mutation in the ATRX Zn finger domain and a totally skewed X-inactivation pattern. 140**,** 2212-2215.

BAGHERI-FAM, S., ARGENTARO, A., SVINGEN, T., COMBES, A. N., SINCLAIR, A. H., KOOPMAN, P. & HARLEY, V. R. J. H. M. G. 2011. Defective survival of proliferating Sertoli cells and androgen receptor function in a mouse model of the ATR-X syndrome. 20**,** 2213-2224.

BANERJI, J., RUSCONI, S. & SCHAFFNER, W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell,* 27**,** 299-308.

BARTHOLOMEW, B. J. A. R. O. B. 2014. Regulating the chromatin landscape: structural and mechanistic perspectives. 83**,** 671-696.

BAUMANN, C. & DE LA FUENTE, R. J. C. 2009. ATRX marks the inactive X chromosome (Xi) in somatic cells and during imprinted X chromosome inactivation in trophoblast stem cells. 118**,** 209-222.

BENOIST, C. & CHAMBON, P. J. N. 1981. In vivo sequence requirements of the SV40 early promoter region. 290**,** 304-310.

BÉRUBÉ, N. G., HEALY, J., MEDINA, C. F., WU, S., HODGSON, T., JAGLA, M. & PICKETTS, D. J. 2008. Patient mutations alter ATRX targeting to PML nuclear bodies. *European Journal of Human Genetics,* 16**,** 192-201.

BERUBE, N. G., MANGELSDORF, M., JAGLA, M., VANDERLUIT, J., GARRICK, D., GIBBONS, R. J., HIGGS, D. R., SLACK, R. S. & PICKETTS, D. J. 2005. The chromatin-remodeling protein ATRX is critical for neuronal survival during corticogenesis. *J Clin Invest,* 115**,** 258-67.

BLACK, B. E. & CLEVELAND, D. W. J. C. 2011. Epigenetic centromere propagation and the nature of CENP-a nucleosomes. 144**,** 471-479.

BLACK, D. L., CHABOT, B. & STEITZ, J. A. J. C. 1985. U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. 42**,** 737-750.

BLACKLEDGE, N. P., LONG, H. K., ZHOU, J. C., KRIAUCIONIS, S., PATIENT, R. & KLOSE, R. J. 2011. Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic acids research,* 40**,** e32-e32.

BOCHMAN, M. L., PAESCHKE, K. & ZAKIAN, V. A. 2012. DNA secondary structures: stability and function of G-quadruplex structures. *Nature Reviews Genetics,* 13**,** 770-780.

BÖNISCH, C. & HAKE, S. B. J. N. A. R. 2012. Histone H2A variants in nucleosomes and chromatin: more or less stable? 40**,** 10719-10741.

BORTZ, P. D. S. & WAMHOFF, B. R. 2011. Chromatin immunoprecipitation (ChIP): revisiting the efficacy of sample preparation, sonication, quantification of sheared DNA, and analysis via PCR. *PLoS One,* 6**,** e26015.

BOURGON, R., GENTLEMAN, R. & HUBER, W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences,* 107**,** 9546-9551.

BRÁZDA, V., LAISTER, R. C., JAGELSKÁ, E. B. & ARROWSMITH, C. 2011. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC molecular biology,* 12**,** 33.

BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C., CHANG, H. Y. & GREENLEAF, W. J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods,* 10**,** 1213-1218.

CAI, H., HU, J., SONG, P. & GONG, W. 2006. PSM2, a novel protein similar to MCAF2, is involved in the mouse embryonic and adult male gonad development. *Molecular biology reports,* 33**,** 159-166.

CHAMBERS, V. S., MARSICO, G., BOUTELL, J. M., DI ANTONIO, M., SMITH, G. P. & BALASUBRAMANIAN, S. 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature biotechnology,* 33**,** 877-881.

CHANG, F. T., MCGHIE, J. D., CHAN, F. L., TANG, M. C., ANDERSON, M. A., MANN, J. R., CHOO, K. A. & WONG, L. H. 2013. PML bodies provide an important platform for the maintenance of telomeric chromatin integrity in embryonic stem cells. *Nucleic acids research,* 41**,** 4447-4458.

CHANG, J., LU, R., XU, S., MENESES, J., CHAN, K., PEDERSEN, R. & KAN, Y. J. B. 1996. Inactivation of mouse alpha-globin gene by homologous recombination: mouse model of hemoglobin H disease. 88**,** 1846-1851.

CLYNES, D., HIGGS, D. R. & GIBBONS, R. J. 2013. The chromatin remodeller ATRX: a repeat offender in human disease. *Trends in biochemical sciences,* 38**,** 461-466.

CLYNES, D., JELINSKA, C., XELLA, B., AYYUB, H., SCOTT, C., MITSON, M., TAYLOR, S., HIGGS, D. R. & GIBBONS, R. J. 2015. Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRX. *Nature communications,* 6.

CLYNES, D., JELINSKA, C., XELLA, B., AYYUB, H., TAYLOR, S., MITSON, M., BACHRATI, C. Z., HIGGS, D. R. & GIBBONS, R. J. 2014. ATRX dysfunction induces replication defects in primary mouse cells. *PloS one,* 9**,** e92915.

COLLAS, P. & DAHL, J. A. 2008. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci,* 13**,** 929-943.

CONSORTIUM, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

COSTANZI, C. & PEHRSON, J. R. 2001. MACROH2A2, a new member of the MACROH2A core histone family. *Journal of Biological Chemistry,* 276**,** 21776-21784.

CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M. & SHARP, P. A. J. P. O. T. N. A. O. S. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. 107**,** 21931-21936.

CUTTER, A. R. & HAYES, J. J. 2015. A brief review of nucleosome structure. *FEBS letters,* 589**,** 2914-2922.

DA WEI HUANG, B. T. S., STEPHENS, R., BASELER, M. W., LANE, H. C. & LEMPICKI, R. A. 2008. DAVID gene ID conversion tool. *Bioinformation,* 2**,** 428.

DAVIES, J. O., TELENIUS, J. M., MCGOWAN, S. J., ROBERTS, N. A., TAYLOR, S., HIGGS, D. R. & HUGHES, J. R. 2015. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature methods,* 13**,** 74.

DAY, D. S., LUQUETTE, L. J., PARK, P. J. & KHARCHENKO, P. V. 2010. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol,* 11**,** R69.

DE LA FUENTE, R., VIVEIROS, M. M., WIGGLESWORTH, K. & EPPIG, J. J. 2004. ATRX, a member of the SNF2 family of helicase/ATPases, is required for chromosome alignment and meiotic spindle organization in metaphase II stage mouse oocytes. *Developmental biology,* 272**,** 1-14.

DELLAIRE, G. & BAZETT-JONES, D. P. J. B. 2004. PML nuclear bodies: dynamic sensors of DNA damage and cellular stress. 26**,** 963-977.

DENIZIO, J. E., ELSÄSSER, S. J. & BLACK, B. E. J. N. A. R. 2014. DAXX co-folds with H3. 3/H4 using high local stability conferred by the H3. 3 variant recognition residues. 42**,** 4318-4331.

DRANÉ, P., OUARARHNI, K., DEPAUX, A., SHUAIB, M. & HAMICHE, A. 2010. The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3. 3. *Genes & development,* 24**,** 1253-1265.

EHRLICH, M., GAMA-SOSA, M. A., HUANG, L.-H., MIDGETT, R. M., KUO, K. C., MCCUNE, R. A. & GEHRKE, C. J. N. A. R. 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. 10**,** 2709-2721.

ELSÄSSER, S. J., HUANG, H., LEWIS, P. W., CHIN, J. W., ALLIS, C. D. & PATEL, D. J. J. N. 2012. DAXX envelops a histone H3. 3–H4 dimer for H3. 3-specific recognition. 491**,** 560.

ELSÄSSER, S. J., NOH, K.-M., DIAZ, N., ALLIS, C. D. & BANASZYNSKI, L. A. 2015. Histone H3. 3 is required for endogenous retroviral element silencing in embryonic stem cells. *Nature,* 522**,** 240-244.

ERNST, J., KHERADPOUR, P., MIKKELSEN, T. S., SHORESH, N., WARD, L. D., EPSTEIN, C. B., ZHANG, X., WANG, L., ISSNER, R. & COYNE, M. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43.

EUSTERMANN, S., YANG, J.-C., LAW, M. J., AMOS, R., CHAPMAN, L. M., JELINSKA, C., GARRICK, D., CLYNES, D., GIBBONS, R. J., RHODES, D. J. N. S. & BIOLOGY, M. 2011. Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. 18**,** 777.

FINN, R. D., ATTWOOD, T. K., BABBITT, P. C., BATEMAN, A., BORK, P., BRIDGE, A. J., CHANG, H.-Y., DOSZTÁNYI, Z., EL-GEBALI, S. & FRASER, M. 2016. InterPro in 2017—beyond protein family and domain annotations. *Nucleic acids research,* 45**,** D190-D199.

FINN, R. D., BATEMAN, A., CLEMENTS, J., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., HEGER, A., HETHERINGTON, K., HOLM, L. & MISTRY, J. 2013. Pfam: the protein families database. *Nucleic acids research,* 42**,** D222-D230.

FISCHLE, W., WANG, Y., JACOBS, S. A., KIM, Y., ALLIS, C. D., KHORASANIZADEH, S. J. G. & DEVELOPMENT 2003. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. 17**,** 1870-1881.

FRANKLIN, S. G. & ZWEIDLER, A. 1977. Non-allelic variants of histones 2a, 2b and 3 in mammals. *Nature,* 266**,** 273.

FRIED, V. 1992. MacroH2A, a core histone containing a large nonhistone region. *Science,* 257**,** 1398-1400.

FRIETZE, S., O'GEEN, H., BLAHNIK, K. R., JIN, V. X. & FARNHAM, P. J. J. P. O. 2010. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. 5**,** e15082.

FUKAMACHI, H. & ITO, K. J. O. 2004. Growth regulation of gastric epithelial cells by Runx3. 23**,** 4330.

FUREY, T. S. 2012. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics,* 13**,** 840-852.

GARRICK, D., DE GOBBI, M., SAMARA, V., RUGLESS, M., HOLLAND, M., AYYUB, H., LOWER, K., SLOANE-STANLEY, J., GRAY, N. & KOCH, C. 2008. The role of the polycomb complex in silencing α-globin gene expression in nonerythroid cells. *Blood,* 112**,** 3889-3899.

GARRICK, D., SAMARA, V., MCDOWELL, T. L., SMITH, A. J., DOBBIE, L., HIGGS, D. R. & GIBBONS, R. J. 2004. A conserved truncated isoform of the ATR-X syndrome protein lacking the SWI/SNF-homology domain. *Gene,* 326**,** 23-34.

GARRICK, D., SHARPE, J. A., ARKELL, R., DOBBIE, L., SMITH, A. J., WOOD, W. G., HIGGS, D. R. & GIBBONS, R. J. 2006. Loss of Atrx affects trophoblast development and the pattern of X-inactivation in extraembryonic tissues. *PLoS genetics,* 2**,** e58.

GAUTIER, L., COPE, L., BOLSTAD, B. M. & IRIZARRY, R. A. 2004. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics,* 20**,** 307-315.

GENTLEMAN, R., CAREY, V., HUBER, W. & HAHNE, F. 2016. genefilter: genefilter: methods for filtering genes from high-throughput experiments. 2016. *R package version,* 1.

GHOSAL, G. & MUNIYAPPA, K. J. N. A. R. 2005. Saccharomyces cerevisiae Mre11 is a high-affinity G4 DNA-binding protein and a G-rich DNA-specific endonuclease: implications for replication of telomeric DNA. 33**,** 4692-4703.

GIBBONS, R. J., MCDOWELL, T. L., RAMAN, S., O'ROURKE, D. M., GARRICK, D., AYYUB, H. & HIGGS, D. R. 2000. Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. *Nature genetics,* 24**,** 368-371.

GIBBONS, R. J., PELLAGATTI, A., GARRICK, D., WOOD, W. G., MALIK, N., AYYUB, H., LANGFORD, C., BOULTWOOD, J., WAINSCOAT, J. S. & HIGGS, D. R. 2003. Identification of acquired somatic mutations in the gene encoding chromatin-remodeling factor ATRX in the α-thalassemia myelodysplasia syndrome (ATMDS). *Nature genetics,* 34**,** 446-449.

GIBBONS, R. J., PICKETTS, D. J., VILLARD, L. & HIGGS, D. R. 1995. Mutations in a putative global transcriptional regulator cause X-linked mental retardation with α-thalassemia (ATR-X syndrome). *Cell,* 80**,** 837-845.

GIBBONS, R. J., WADA, T., FISHER, C. A., MALIK, N., MITSON, M. J., STEENSMA, D. P., FRYER, A., GOUDIE, D. R., KRANTZ, I. D. & TRAEGER-SYNODINOS, J. 2008. Mutations in the chromatin-associated protein ATRX. *Human mutation,* 29**,** 796-802.

GIBBONS, R. J. O. J. O. R. D. 2006. Alpha thalassaemia-mental retardation, X linked. 1**,** 15.

GILBERT, N., BOYLE, S., FIEGLER, H., WOODFINE, K., CARTER, N. P. & BICKMORE, W. A. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell,* 118**,** 555-566.

GOLDBERG, A. D., BANASZYNSKI, L. A., NOH, K.-M., LEWIS, P. W., ELSAESSER, S. J., STADLER, S., DEWELL, S., LAW, M., GUO, X. & LI, X. 2010. Distinct factors control histone variant H3. 3 localization at specific genomic regions. *Cell,* 140**,** 678-691.

HALDER, R., HALDER, K., SHARMA, P., GARG, G., SENGUPTA, S. & CHOWDHURY, S. 2010. Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Molecular bioSystems,* 6**,** 2439-2447.

HÄNSEL-HERTSCH, R., BERALDI, D., LENSING, S. V., MARSICO, G., ZYNER, K., PARRY, A., DI ANTONIO, M., PIKE, J., KIMURA, H. & NARITA, M. 2016. G-quadruplex structures mark human regulatory chromatin. *Nature genetics*.

HÄNSEL-HERTSCH, R., DI ANTONIO, M. & BALASUBRAMANIAN, S. J. N. R. M. C. B. 2017. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. 18**,** 279.

HANSSEN, L. L., KASSOUF, M. T., OUDELAAR, A. M., BIGGS, D., PREECE, C., DOWNES, D. J., GOSDEN, M., SHARPE, J. A., SLOANE-STANLEY, J. A. & HUGHES, J. R. 2017. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature cell biology,* 19**,** 952.

HARTEVELD, C. L. & HIGGS, D. R. J. O. J. O. R. D. 2010. α-thalassaemia. 5**,** 13.

HAY, D., HUGHES, J. R., BABBS, C., DAVIES, J. O., GRAHAM, B. J., HANSSEN, L. L., KASSOUF, M. T., OUDELAAR, A. M., SHARPE, J. A. & SUCIU, M. C. 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nature genetics,* 48**,** 895.

HEAPHY, C. M., DE WILDE, R. F., JIAO, Y., KLEIN, A. P., EDIL, B. H., SHI, C., BETTEGOWDA, C., RODRIGUEZ, F. J., EBERHART, C. G. & HEBBAR, S. 2011. Altered telomeres in tumors with ATRX and DAXX mutations. *Science,* 333**,** 425-425.

HEGYI, H. 2015. Enhancer-promoter interaction facilitated by transiently forming G-quadruplexes. *Scientific reports,* 5.

HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell,* 38**,** 576-589.

HELLMAN, A. & CHESS, A. J. S. 2007. Gene body-specific methylation on the active X chromosome. 315**,** 1141-1143.

HIGGS, D., GARRICK, D., ANGUITA, E., DE GOBBI, M., HUGHES, J., MUERS, M., VERNIMMEN, D., LOWER, K., LAW, M. & ARGENTARO, A. 2005. Understanding α-globin gene regulation: aiming to improve the management of Thalassemia. *Annals of the New York Academy of Sciences,* 1054**,** 92-102.

HIGGS, D. R. & WOOD, W. G. 2008. Long-range regulation of α globin gene expression during erythropoiesis. *Current opinion in hematology,* 15**,** 176-183.

HOELPER, D., HUANG, H., JAIN, A. Y., PATEL, D. J. & LEWIS, P. W. 2017. Structural and mechanistic insights into ATRX-dependent and-independent functions of the histone chaperone DAXX. *Nature Communications,* 8**,** 1193.

HONG, L., SCHROTH, G., MATTHEWS, H., YAU, P. & BRADBURY, E. 1993. Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4" tail" to DNA. *Journal of Biological Chemistry,* 268**,** 305-314.

HUANG, Y., GU, L. & LI, G.-M. J. J. O. B. C. 2018. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. jbc. RA118. 002839.

HUH, M. S., O'DEA, T. P., OUAZIA, D., MCKAY, B. C., PARISE, G., PARKS, R. J., RUDNICKI, M. A. & PICKETTS, D. J. 2012. Compromised genomic integrity impedes muscle growth after Atrx inactivation. *The Journal of clinical investigation,* 122**,** 4412.

HUISINGA, K. L., BROWER-TOLAND, B. & ELGIN, S. C. J. C. 2006. The contradictory definitions of heterochromatin: transcription and silencing. 115**,** 110-122.

HUPPERT, J. L. & BALASUBRAMANIAN, S. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic acids research,* 35**,** 406-413.

HUYHN, K., RENFREE, M. B., GRAVES, J. A. & PASK, A. J. 2011. ATRX has a critical and conserved role in mammalian sexual differentiation. *BMC developmental biology,* 11**,** 39.

HYUN, K., JEON, J., PARK, K., KIM, J. J. E. & MEDICINE, M. 2017. Writing, erasing and reading histone lysine methylations. 49**,** e324.

ICHIMURA, T., WATANABE, S., SAKAMOTO, Y., AOTO, T., FUJITA, N. & NAKAO, M. 2005. Transcriptional repression and heterochromatin formation by MBD1 and MCAF/AM family proteins. *Journal of Biological Chemistry,* 280**,** 13928-13935.

IMBEAULT, M., HELLEBOID, P.-Y. & TRONO, D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature,* 543**,** 550.

ISHOV, A. M., VLADIMIROVA, O. V. & MAUL, G. G. 2004. Heterochromatin and ND10 are cell-cycle regulated and phosphorylation-dependent alternate nuclear sites of the transcription repressor Daxx and SWI/SNF protein ATRX. *Journal of cell science,* 117**,** 3807-3820.

IVANAUSKIENE, K., DELBARRE, E., MCGHIE, J. D., KÜNTZIGER, T., WONG, L. H. & COLLAS, P. J. G. R. 2014. The PML-associated protein DEK regulates the balance of H3. 3 loading on chromatin and is important for telomere integrity. gr. 173831.114.

IWASE, S., XIANG, B., GHOSH, S., REN, T., LEWIS, P. W., COCHRANE, J. C., ALLIS, C. D., PICKETTS, D. J., PATEL, D. J. & LI, H. 2011. ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. *Nature structural & molecular biology,* 18**,** 769-776.

IYER, B. V., KENWARD, M. & ARYA, G. 2011. Hierarchies in eukaryotic genome organization: insights from polymer theory and simulations. *BMC biophysics,* 4**,** 8.

JIANG, W.-Q., NGUYEN, A., CAO, Y., CHANG, A. C.-M. & REDDEL, R. R. J. P. O. 2011. HP1-mediated formation of alternative lengthening of telomeres-associated PML bodies requires HIRA but not ASF1a. 6**,** e17036.

JIN, C., FELSENFELD, G. J. G. & DEVELOPMENT 2007. Nucleosome stability mediated by histone variants H3. 3 and H2A. Z. 21**,** 1519-1529.

JONES, P. A. J. N. R. G. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. 13**,** 484.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome research,* 12**,** 996-1006.

KERNOHAN, K. D., JIANG, Y., TREMBLAY, D. C., BONVISSUTO, A. C., EUBANKS, J. H., MANN, M. R. & BÉRUBÉ, N. G. 2010. ATRX partners with cohesin and MeCP2 and contributes to developmental silencing of imprinted genes in the brain. *Developmental cell,* 18**,** 191-202.

KERNOHAN, K. D., VERNIMMEN, D., GLOOR, G. B. & BÉRUBÉ, N. G. 2014. Analysis of neonatal brain lacking ATRX or MeCP2 reveals changes in nucleosome density, CTCF binding and chromatin looping. *Nucleic acids research***,** gku564.

KOSCHMANN, C., CALINESCU, A.-A., NUNEZ, F. J., MACKAY, A., FAZAL-SALOM, J., THOMAS, D., MENDEZ, F., KAMRAN, N., DZAMAN, M. & MULPURI, L. J. S. T. M. 2016. ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma. 8**,** 328ra28-328ra28.

KRUEGER, F. 2015. Trim Galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*.

KUNDAJE, A. 2013. A comprehensive collection of signal artifact blacklist regions in the human genome. *ENCODE.[hg19-blacklist-README. doc-EBI]. Available online at: https://sites. google. com/site/anshulkundaje/projects/blacklists*.

LACHNER, M., O'CARROLL, D., REA, S., MECHTLER, K. & JENUWEIN, T. J. N. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. 410**,** 116.

LAI, W. K. & PUGH, B. F. 2017. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Reviews Molecular Cell Biology,* 18**,** 548.

LALLEMAND-BREITENBACH, V. J. C. S. H. P. I. B. 2010. PML nuclear bodies. a000661.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M. & FITZHUGH, W. J. N. 2001. Initial sequencing and analysis of the human genome. 409**,** 860-922.

LANE, A. N., CHAIRES, J. B., GRAY, R. D. & TRENT, J. O. J. N. A. R. 2008. Stability and kinetics of G-quadruplex structures. 36**,** 5482-5515.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods,* 9**,** 357-359.

LAW, M. J., LOWER, K. M., VOON, H. P. J., HUGHES, J. R., GARRICK, D., VIPRAKASIT, V., MITSON, M., DE GOBBI, M., MARRA, M., MORRIS, A., ABBOTT, A., WILDER, S. P., TAYLOR, S., SANTOS, G. M., CROSS, J., AYYUB, H., JONES, S., RAGOUSSIS, J., RHODES, D., DUNHAM, I., HIGGS, D. R. & GIBBONS, R. J. 2010. ATR-X Syndrome Protein Targets Tandem Repeats and Influences Allele-Specific Expression in a Size-Dependent Manner. *Cell,* 143**,** 367-378.

LECHNER, M. S., SCHULTZ, D. C., NEGOREV, D., MAUL, G. G., RAUSCHER III, F. J. J. B. & COMMUNICATIONS, B. R. 2005. The mammalian heterochromatin protein 1 binds diverse nuclear proteins through a common motif that targets the chromoshadow domain. 331**,** 929-937.

LEE, T. I., JOHNSTONE, S. E. & YOUNG, R. A. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature protocols,* 1**,** 729.

LEHNERTZ, B., UEDA, Y., DERIJCK, A. A., BRAUNSCHWEIG, U., PEREZ-BURGOS, L., KUBICEK, S., CHEN, T., LI, E., JENUWEIN, T. & PETERS, A. H. J. C. B. 2003. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. 13**,** 1192-1200.

LENTJES, M. H., NIESSEN, H. E., AKIYAMA, Y., DE BRUINE, A. P., MELOTTE, V. & VAN ENGELAND, M. J. E. R. I. M. M. 2016. The emerging role of GATA transcription factors in development and disease. 18.

LEUNG, J. W.-C., GHOSAL, G., WANG, W., SHEN, X., WANG, J., LI, L. & CHEN, J. 2013. Alpha thalassemia/mental retardation syndrome X-linked gene product ATRX is required for proper replication restart and cellular resistance to replication stress. *Journal of Biological Chemistry,* 288**,** 6342-6350.

LEVANON, D., BETTOUN, D., HARRIS-CERRUTI, C., WOOLF, E., NEGREANU, V., EILAM, R., BERNSTEIN, Y., GOLDENBERG, D., XIAO, C. & FLIEGAUF, M. J. T. E. J. 2002. The Runx3 transcription factor regulates development and survival of TrkC dorsal root ganglia neurons. 21**,** 3454-3463.

LEVY, M. A., FERNANDES, A. D., TREMBLAY, D. C., SEAH, C. & BÉRUBÉ, N. G. 2008. The SWI/SNF protein ATRX co-regulates pseudoautosomal genes that have translocated to autosomes in the mouse genome. *BMC genomics,* 9**,** 468.

LEVY, M. A., KERNOHAN, K. D., JIANG, Y. & BÉRUBÉ, N. G. 2014. ATRX promotes gene expression by facilitating transcriptional elongation through guanine-rich coding regions. *Human molecular genetics,* 24**,** 1824-1835.

LEWIS, P. W., ELSAESSER, S. J., NOH, K.-M., STADLER, S. C. & ALLIS, C. D. 2010. Daxx is an H3. 3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres. *Proceedings of the National Academy of Sciences,* 107**,** 14075-14080.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

LI, H., ILIN, S., WANG, W., DUNCAN, E. M., WYSOCKA, J., ALLIS, C. D. & PATEL, D. J. J. N. 2006. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. 442**,** 91.

LI, Y. & LI, H. J. A. B. B. S. 2012. Many keys to push: diversifying the 'readership'of plant homeodomain fingers. 44**,** 28-39.

LI, Y., SYED, J., SUZUKI, Y., ASAMITSU, S., SHIODA, N., WADA, T. & SUGIYAMA, H. 2016. Effect of ATRX and G-Quadruplex Formation by the VNTR Sequence on α-Globin Gene Expression. *ChemBioChem,* 17**,** 928-935.

LIANG, J., ZHOU, H., GERDT, C., TAN, M., COLSON, T., KAYE, K. M., KIEFF, E. & ZHAO, B. 2016. Epstein–Barr virus super-enhancer eRNAs are essential for MYC oncogene expression and lymphoblast proliferation. *Proceedings of the National Academy of Sciences***,** 201616697.

LONG, H. K., SIMS, D., HEGER, A., BLACKLEDGE, N. P., KUTTER, C., WRIGHT, M. L., GRÜTZNER, F., ODOM, D. T., PATIENT, R. & PONTING, C. P. J. E. 2013. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. 2**,** e00348.

LOSSI, A., MILLAN, J., VILLARD, L., ORELLANA, C., CARDOSO, C., PRIETO, F., FONTES, M. & MARTINEZ, F. 1999. Mutation of the XNP/ATR-X gene in a family with severe mental retardation, spastic paraplegia and skewed pattern of X inactivation: demonstration that the mutation is involved in the inactivation bias. *American journal of human genetics,* 65**,** 558.

LOVEJOY, C. A. & CORTEZ, D. 2009. Common mechanisms of PIKK regulation. *DNA repair,* 8**,** 1004-1008.

LOVEJOY, C. A., LI, W., REISENWEBER, S., THONGTHIP, S., BRUNO, J., DE LANGE, T., DE, S., PETRINI, J. H., SUNG, P. A. & JASIN, M. 2012. Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of telomeres pathway. *PLoS genetics,* 8**,** e1002772.

LOWER, K. M., HUGHES, J. R., DE GOBBI, M., HENDERSON, S., VIPRAKASIT, V., FISHER, C., GORIELY, A., AYYUB, H., SLOANE-STANLEY, J. & VERNIMMEN, D. J. P. O. T. N. A. O. S. 2009. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. 106**,** 21771-21776.

LUEDI, P. P., DIETRICH, F. S., WEIDMAN, J. R., BOSKO, J. M., JIRTLE, R. L. & HARTEMINK, A. J. 2007. Computational and experimental identification of novel human imprinted genes. *Genome research,* 17**,** 000-000.

LUGER, K. & RICHMOND, T. J. 1998. The histone tails of the nucleosome. *Current opinion in genetics & development,* 8**,** 140-146.

LUND, R. J., RAHKONEN, N., MALONZO, M., KAUKO, L., EMANI, M. R., KIVINEN, V., NÄRVÄ, E., KEMPPAINEN, E., LAIHO, A. & SKOTTMAN, H. J. S. C. R. 2017. RNA Polymerase III Subunit POLR3G Regulates Specific Subsets of PolyA+ and SmallRNA Transcriptomes and Splicing in Human Pluripotent Stem Cells. 8**,** 1442-1454.

LUPO, A., CESARO, E., MONTANO, G., ZURLO, D., IZZO, P. & COSTANZO, P. 2013. KRAB-zinc finger proteins: a repressor family displaying multiple biological functions. *Current genomics,* 14**,** 268-278.

MAZE, I., NOH, K.-M., SOSHNEV, A. A. & ALLIS, C. D. J. N. R. G. 2014. Every amino acid matters: essential contributions of histone variants to mammalian development and disease. 15**,** 259.

MCDOWELL, T., GIBBONS, R., SUTHERLAND, H., O'ROURKE, D., BICKMORE, W., POMBO, A., TURLEY, H., GATTER, K., PICKETTS, D. & BUCKLE, V. J. P. O. T. N. A. O. S. 1999. Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes. 96**,** 13983-13988.

MCGOWAN, S. J., HUGHES, J. R., HAN, Z.-P. & TAYLOR, S. 2013. MIG: Multi-Image Genome viewer. *Bioinformatics,* 29**,** 2477-2478.

MEISSNER, A., MIKKELSEN, T. S., GU, H., WERNIG, M., HANNA, J., SIVACHENKO, A., ZHANG, X., BERNSTEIN, B. E., NUSBAUM, C. & JAFFE, D. B. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature,* 454**,** 766.

MEYER, C. A. & LIU, X. S. 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics,* 15**,** 709-721.

MINKOVSKY, A., SAHAKYAN, A., RANKIN-GEE, E., BONORA, G., PATEL, S. & PLATH, K. 2014. The Mbd1-Atf7ip-Setdb1 pathway contributes to the maintenance of X chromosome inactivation. *Epigenetics & chromatin,* 7**,** 1.

MIRSKY, A. E., SILVERMAN, B. & PANDA, N. C. J. P. O. T. N. A. O. S. 1972. Blocking by histones of accessibility to DNA in chromatin: addition of histones. 69**,** 3243-3246.

MITSON, M., KELLEY, L. A., STERNBERG, M. J., HIGGS, D. R. & GIBBONS, R. J. 2011. Functional significance of mutations in the Snf2 domain of ATRX. *Human molecular genetics,* 20**,** 2603-2610.

MOHYUDDIN, A., AYUB, Q., SIDDIQI, S., CARVALHO-SILVA, D. R., MAZHAR, K., REHMAN, S., FIRASAT, S., DAR, A., TYLER-SMITH, C. & MEHDI, S. Q. 2004. Genetic instability in EBV-transformed lymphoblastoid cell lines. *Biochimica et Biophysica Acta (BBA)-General Subjects,* 1670**,** 81-83.

MOROZOV, V. M., GIOVINAZZI, S., ISHOV, A. M. J. E. & CHROMATIN 2017. CENP-B protects centromere chromatin integrity by facilitating histone deposition via the H3. 3-specific chaperone Daxx. 10**,** 63.

NAN, X., HOU, J., MACLEAN, A., NASIR, J., LAFUENTE, M. J., SHU, X., KRIAUCIONIS, S. & BIRD, A. 2007. Interaction between chromatin proteins MECP2 and ATRX is disrupted by mutations that cause inherited mental retardation. *Proceedings of the National Academy of Sciences,* 104**,** 2709-2714.

NASHUN, B., YUKAWA, M., LIU, H., AKIYAMA, T. & AOKI, F. J. D. 2010. Changes in the nuclear deposition of histone H2A variants during pre-implantation development in mice. dev. 051805.

NGUYEN, D. T., VOON, H. P. J., XELLA, B., SCOTT, C., CLYNES, D., BABBS, C., AYYUB, H., KERRY, J., SHARPE, J. A. & SLOANE-STANLEY, J. A. 2017. The chromatin remodelling factor ATRX suppresses R-loops in transcribed telomeric repeats. *EMBO reports,* 18**,** 914-928.

NOLIS, I. K., MCKAY, D. J., MANTOUVALOU, E., LOMVARDAS, S., MERIKA, M. & THANOS, D. J. P. O. T. N. A. O. S. 2009. Transcription factors mediate long-range enhancer–promoter interactions. 106**,** 20222-20227.

O'LEARY, N. A., WRIGHT, M. W., BRISTER, J. R., CIUFO, S., HADDAD, D., MCVEIGH, R., RAJPUT, B., ROBBERTSE, B., SMITH-WHITE, B. & AKO-ADJEI, D. J. N. A. R. 2015. Reference sequence

(RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. 44**,** D733-D745.

O'SULLIVAN, R. J., ARNOULT, N., LACKNER, D. H., OGANESIAN, L., HAGGBLOM, C., CORPET, A., ALMOUZNI, G., KARLSEDER, J. J. N. S. & BIOLOGY, M. 2014. Rapid induction of alternative lengthening of telomeres by depletion of the histone chaperone ASF1. 21**,** 167.

OKANO, M., BELL, D. W., HABER, D. A. & LI, E. J. C. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. 99**,** 247-257.

OLIVA-RICO, D. & HERRERA, L. A. 2017. Regulated expression of the lncRNA TERRA and its impact on telomere biology. *Mechanisms of ageing and development*.

PAESCHKE, K., BOCHMAN, M. L., GARCIA, P. D., CEJKA, P., FRIEDMAN, K. L., KOWALCZYKOWSKI, S. C. & ZAKIAN, V. A. J. N. 2013. Pif1 family helicases suppress genome instability at G-quadruplex motifs. 497**,** 458.

PALMER, D. K., O'DAY, K., WENER, M. H., ANDREWS, B. S. & MARGOLIS, R. L. 1987. A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. *The Journal of cell biology,* 104**,** 805-815.

PAPADOPOULOU, C., GUILBAUD, G., SCHIAVONE, D. & SALE, J. E. 2015. Nucleotide pool depletion induces G-quadruplex-dependent perturbation of gene expression. *Cell reports,* 13**,** 2491-2503.

PARK, P. J. 2009. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics,* 10**,** 669-680.

PERTEA, M. & SALZBERG, S. L. J. G. B. 2010. Between a chicken and a grape: estimating the number of human genes. 11**,** 206.

PEUKERT, K., STALLER, P., SCHNEIDER, A., CARMICHAEL, G., HÄNEL, F. & EILERS, M. J. T. E. J. 1997. An alternative pathway for gene regulation by Myc. 16**,** 5672-5686.

PHILLIPS, D. 1963. The presence of acetyl groups in histones. *Biochemical Journal,* 87**,** 258.

PHILLIPS-CREMINS, J. E. 2014. Unraveling architecture of the pluripotent genome. *Current opinion in cell biology,* 28**,** 96-104.

PICKETTS, D. J., HIGGS, D. R., BACHOO, S., BLAKE, D. J., QUARRELL, O. W. & GIBBONS, R. J. 1996. ATRX encodes a novel member of the SNF2 family of proteins: mutations point to a common mechanism underlying the ATR-X syndrome. *Human molecular genetics,* 5**,** 1899-1907.

POGO, B., ALLFREY, V. & MIRSKY, A. 1966. RNA synthesis and histone acetylation during the course of gene activation in lymphocytes. *Proceedings of the National Academy of Sciences,* 55**,** 805-812.

PUTO, L. A., BENNER, C. & HUNTER, T. 2015. The DAXX co-repressor is directly recruited to active regulatory elements genome-wide to regulate autophagy programs in a model of human prostate cancer. *Oncoscience,* 2**,** 362.

QUENNEVILLE, S., VERDE, G., CORSINOTTI, A., KAPOPOULOU, A., JAKOBSSON, J., OFFNER, S., BAGLIVO, I., PEDONE, P. V., GRIMALDI, G. & RICCIO, A. 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Molecular cell,* 44**,** 361-372.

RABINOVITZ, S., KAUFMAN, Y., LUDWIG, G., RAZIN, A. & SHEMER, R. 2012. Mechanisms of activation of the paternally expressed genes by the Prader-Willi imprinting center in the Prader-Willi/Angelman syndromes domains. *Proceedings of the National Academy of Sciences***,** 201116661.

RAMÍREZ, F., DÜNDAR, F., DIEHL, S., GRÜNING, B. A. & MANKE, T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research,* 42**,** W187-W191.

RANSOM, M., DENNEHEY, B. K. & TYLER, J. K. J. C. 2010. Chaperoning histones during DNA replication and repair. 140**,** 183-195.

RATNAKUMAR, K., DUARTE, L. F., LEROY, G., HASSON, D., SMEETS, D., VARDABASSO, C., BÖNISCH, C., ZENG, T., XIANG, B. & ZHANG, D. Y. 2012. ATRX-mediated chromatin association of histone variant macroH2A1 regulates α-globin expression. *Genes & development,* 26**,** 433-438.

RHODES, D. & LIPPS, H. J. 2015. G-quadruplexes and their regulatory roles in biology. *Nucleic acids research,* 43**,** 8627-8637.

RIBEYRE, C., LOPES, J., BOULÉ, J.-B., PIAZZA, A., GUÉDIN, A., ZAKIAN, V. A., MERGNY, J.-L. & NICOLAS, A. 2009. The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS genetics,* 5**,** e1000475.

RICHMOND, T., FINCH, J., RUSHTON, B., RHODES, D. & KLUG, A. 1984. Structure of the nucleosome core particle at 7 Å resolution. *Nature,* 311**,** 532.

RITCHIE, K., SEAH, C., MOULIN, J., ISAAC, C., DICK, F. & BÉRUBÉ, N. G. 2008. Loss of ATRX leads to chromosome cohesion and congression defects. *The Journal of cell biology,* 180**,** 315-324.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* 43**,** e47-e47.

ROBERTS, C., SUTHERLAND, H. F., FARMER, H., KIMBER, W., HALFORD, S., CAREY, A., BRICKMAN, J. M., WYNSHAW-BORIS, A., SCAMBLER, P. J. J. M. & BIOLOGY, C. 2002. Targeted mutagenesis of the Hira gene results in gastrulation defects and patterning abnormalities of mesoendodermal derivatives prior to early embryonic lethality. 22**,** 2318-2328.

RYAN, D. P. & OWEN-HUGHES, T. J. C. O. I. C. B. 2011. Snf2-family proteins: chromatin remodellers for any occasion. 15**,** 649-656.

SADIC, D., SCHMIDT, K., GROH, S., KONDOFERSKY, I., ELLWART, J., FUCHS, C., THEIS, F. J. & SCHOTTA, G. 2015. Atrx promotes heterochromatin formation at retrotransposons. *EMBO reports***,** e201439937.

SAHA, A., WITTMEYER, J. & CAIRNS, B. R. J. N. R. M. C. B. 2006. Chromatin remodelling: the industrial revolution of DNA around histones. 7**,** 437.

SAHAKYAN, A. B., CHAMBERS, V. S., MARSICO, G., SANTNER, T., DI ANTONIO, M. & BALASUBRAMANIAN, S. 2017. Machine learning model for sequence-driven DNA G-quadruplex formation. *Scientific reports,* 7**,** 14535.

SAKSOUK, N., SIMBOECK, E., DÉJARDIN, J. J. E. & CHROMATIN 2015. Constitutive heterochromatin formation and transcription in mammals. 8**,** 3.

SANTOS-ROSA, H., SCHNEIDER, R., BANNISTER, A. J., SHERRIFF, J., BERNSTEIN, B. E., EMRE, N. T., SCHREIBER, S. L., MELLOR, J. & KOUZARIDES, T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature,* 419**,** 407.

SANZ, L. A., HARTONO, S. R., LIM, Y. W., STEYAERT, S., RAJPURKAR, A., GINNO, P. A., XU, X. & CHÉDIN, F. 2016. Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Molecular cell,* 63**,** 167-178.

SARKIES, P., MURAT, P., PHILLIPS, L. G., PATEL, K., BALASUBRAMANIAN, S. & SALE, J. E. J. N. A. R. 2012. FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. 40**,** 1485-1498.

SARKIES, P., REAMS, C., SIMPSON, L. J. & SALE, J. E. 2010. Epigenetic instability due to defective replication of structured DNA. *Molecular cell,* 40**,** 703-713.

SAXONOV, S., BERG, P. & BRUTLAG, D. L. J. P. O. T. N. A. O. S. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. 103**,** 1412-1417.

SCHENKEL, L. C., KERNOHAN, K. D., MCBRIDE, A., REINA, D., HODGE, A., AINSWORTH, P. J., RODENHISER, D. I., PARE, G., BÉRUBÉ, N. G. & SKINNER, C. 2017. Identification of epigenetic signature associated with alpha thalassemia/mental retardation X-linked syndrome. *Epigenetics & chromatin,* 10**,** 10.

SCHIAVONE, D., JOZWIAKOWSKI, S. K., ROMANELLO, M., GUILBAUD, G., GUILLIAM, T. A., BAILEY, L. J., SALE, J. E. & DOHERTY, A. J. 2016. PrimPol is required for replicative tolerance of G quadruplexes in vertebrate cells. *Molecular cell,* 61**,** 161-169.

SCHNEIDERMAN, J. I., ORSI, G. A., HUGHES, K. T., LOPPIN, B. & AHMAD, K. 2012. Nucleosome-depleted chromatin gaps recruit assembly factors for the H3. 3 histone variant. *Proceedings of the National Academy of Sciences,* 109**,** 19721-19726.

SCHNEIDERMAN, J. I., SAKAI, A., GOLDSTEIN, S. & AHMAD, K. 2009. The XNP remodeler targets dynamic chromatin in Drosophila. *Proceedings of the National Academy of Sciences,* 106**,** 14472-14477.

SCHÜBELER, D., MACALPINE, D. M., SCALZO, D., WIRBELAUER, C., KOOPERBERG, C., VAN LEEUWEN, F., GOTTSCHLING, D. E., O'NEILL, L. P., TURNER, B. M. & DELROW, J. 2004. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes & development,* 18**,** 1263-1271.

SCHULTZ, D. C., AYYANATHAN, K., NEGOREV, D., MAUL, G. G., RAUSCHER, F. J. J. G. & DEVELOPMENT 2002. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. 16**,** 919-932.

SCHWARTZENTRUBER, J., KORSHUNOV, A., LIU, X.-Y., JONES, D. T., PFAFF, E., JACOB, K., STURM, D., FONTEBASSO, A. M., QUANG, D.-A. K. & TÖNJES, M. 2012. Driver mutations in histone H3. 3 and chromatin remodelling genes in paediatric glioblastoma. *Nature,* 482**,** 226.

SHAH, R. R. & BIRD, A. P. J. G. M. 2017. MeCP2 mutations: progress towards understanding and treating Rett syndrome. 9**,** 17.

SHANBHAG, N. M., RAFALSKA-METCALF, I. U., BALANE-BOLIVAR, C., JANICKI, S. M. & GREENBERG, R. A. J. C. 2010. ATM-dependent chromatin changes silence transcription in cis to DNA double-strand breaks. 141**,** 970-981.

SHLYUEVA, D., STAMPFEL, G. & STARK, A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics,* 15**,** 272-286.

SIE, L., LOONG, S. & TAN, E. 2009. Utility of lymphoblastoid cell lines. *Journal of neuroscience research,* 87**,** 1953-1959.

SONG, L., ZHANG, Z., GRASFEDER, L. L., BOYLE, A. P., GIRESI, P. G., LEE, B.-K., SHEFFIELD, N. C., GRÄF, S., HUSS, M. & KEEFE, D. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research*.

SPENDER, L. C., WHITEMAN, H. J., KARSTEGL, C. E. & FARRELL, P. J. J. O. 2005. Transcriptional cross-regulation of RUNX1 by RUNX3 in human B cells. 24**,** 1873.

SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics,* 13**,** 613.

STARK, R. & BROWN, G. 2011. DiffBind: differential binding analysis of ChIP-Seq peak data. *R package version,* 100**,** 4.3.

STEVENS, A. J., STUFFREIN-ROBERTS, S., CREE, S. L., GIBB, A., MILLER, A. L., DOUDNEY, K., AITCHISON, A., ECCLES, M. R., JOYCE, P. R. & FILICHEV, V. V. J. P. O. 2014. G-quadruplex structures and CpG methylation cause drop-out of the maternal allele in polymerase chain reaction amplification of the imprinted MEST gene promoter. 9**,** e113955.

STUMPF, M., WASKOW, C., KRÖTSCHEL, M., VAN ESSEN, D., RODRIGUEZ, P., ZHANG, X., GUYOT, B., ROEDER, R. G. & BORGGREFE, T. J. P. O. T. N. A. O. S. 2006. The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. 103**,** 18504-18509.

SZENKER, E., RAY-GALLET, D. & ALMOUZNI, G. J. C. R. 2011. The double face of the histone variant H3. 3. 21**,** 421.

TANG, J., WU, S., LIU, H., STRATT, R., BARAK, O. G., SHIEKHATTAR, R., PICKETTS, D. J. & YANG, X. 2004. A novel transcription regulatory complex containing death domain-associated protein and the ATR-X syndrome protein. *Journal of Biological Chemistry,* 279**,** 20369-20377.

TING, D. T., LIPSON, D., PAUL, S., BRANNIGAN, B. W., AKHAVANFARD, S., COFFMAN, E. J., CONTINO, G., DESHPANDE, V., IAFRATE, A. J. & LETOVSKY, S. J. S. 2011. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. 331**,** 593-596.

TRUCH, J., TELENIUS, J., HIGGS, D. R. & GIBBONS, R. J. 2018. How to Tackle Challenging ChIP-Seq, with Long-Range Cross-Linking, Using ATRX as an Example. *Histone Variants.* Springer.

TUZ, K., HSIAO, Y.-C., JUÁREZ, O., SHI, B., HARMON, E. Y., PHELPS, I. G., LENNARTZ, M. R., GLASS, I. A., DOHERTY, D. & FERLAND, R. J. 2013. The Joubert Syndrome Associated Missense Mutation (V443D) in the Abelson-helper Integration Site 1 (AHI1) Protein Alters Its Localization and Protein-Protein Interactions. *Journal of Biological Chemistry***,** jbc. M112. 420786.

UDUGAMA, M., SANIJ, E., VOON, H. P., SON, J., HII, L., HENSON, J. D., CHAN, F. L., CHANG, F. T., LIU, Y. & PEARSON, R. B. J. P. O. T. N. A. O. S. 2018. Ribosomal DNA copy loss and repeat instability in ATRX-mutated cancers. 115**,** 4737-4742.

URRUTIA, R. 2003. KRAB-containing zinc-finger repressor proteins. *Genome biology,* 4**,** 231.

VALLE-GARCÍA, D., QADEER, Z. A., MCHUGH, D. S., GHIRALDINI, F. G., CHOWDHURY, A. H., HASSON, D., DYER, M. A., RECILLAS-TARGA, F. & BERNSTEIN, E. 2016. ATRX binds to atypical chromatin domains at the 3' exons of zinc finger genes to preserve H3K9me3 enrichment. *Epigenetics***,** 1-17.

VENOLIA, L. & GARTLER, S. M. J. N. 1983. Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. 302**,** 82.

VLACHAKIS, N., CHOE, S.-K. & SAGERSTROM, C. G. 2001. Meis3 synergizes with Pbx4 and Hoxb1b in promoting hindbrain fates in the zebrafish. *Development,* 128**,** 1299-1312.

VOON, H. P., HUGHES, J. R., RODE, C., INTI, A., JENUWEIN, T., FEIL, R., HIGGS, D. R. & GIBBONS, R. J. 2015. ATRX Plays a Key Role in Maintaining Silencing at Interstitial Heterochromatic Loci and Imprinted Genes. *Cell reports*.

WADA, T., BAN, H., MATSUFUJI, M., OKAMOTO, N., ENOMOTO, K., KUROSAWA, K. & AIDA, N. J. A. J. O. N. 2013. Neuroradiologic features in X-linked α-thalassemia/mental retardation syndrome.

WALDMANN, T., BAACK, M., RICHTER, N. & GRUSS, C. 2003. Structure-specific binding of the proto-oncogene protein DEK to DNA. *Nucleic acids research,* 31**,** 7003-7010.

WASĄG, P. & LENARTOWSKI, R. J. P. H. I. M. D. 2016. Nuclear matrix-structure, function and pathogenesis. 70**,** 1206-1219.

WATSON, L. A., SOLOMON, L. A., LI, J. R., JIANG, Y., EDWARDS, M., SHIN-YA, K., BEIER, F. & BÉRUBÉ, N. G. J. T. J. O. C. I. 2013. Atrx deficiency induces telomere dysfunction, endocrine defects, and reduced life span. 123**,** 2049-2063.

WHITEMAN, H. J. & FARRELL, P. J. 2006. RUNX expression and function in human B cells. *Critical Reviews™ in Eukaryotic Gene Expression,* 16.

WONG, L. H., MCGHIE, J. D., SIM, M., ANDERSON, M. A., AHN, S., HANNAN, R. D., GEORGE, A. J., MORGAN, K. A., MANN, J. R. & CHOO, K. A. 2010. ATRX interacts with H3. 3 in maintaining telomere structural integrity in pluripotent embryonic stem cells. *Genome research,* 20**,** 351-360.

WONG, L. H., REN, H., WILLIAMS, E., MCGHIE, J., AHN, S., SIM, M., TAM, A., EARLE, E., ANDERSON, M. A. & MANN, J. J. G. R. 2009. Histone H3. 3 incorporation provides a unique and functionally essential telomeric chromatin in embryonic stem cells.

XUE, Y., GIBBONS, R., YAN, Z., YANG, D., MCDOWELL, T. L., SECHI, S., QIN, J., ZHOU, S., HIGGS, D. & WANG, W. 2003. The ATRX syndrome protein forms a chromatin-remodeling complex with Daxx and localizes in promyelocytic leukemia nuclear bodies. *Proceedings of the National Academy of Sciences,* 100**,** 10635-10640.

YAMANE, K., TOUMAZOU, C., TSUKADA, Y.-I., ERDJUMENT-BROMAGE, H., TEMPST, P., WONG, J. & ZHANG, Y. J. C. 2006. JHDM2A, a JmjC-containing H3K9 demethylase, facilitates transcription activation by androgen receptor. 125**,** 483-495.

YUH, C.-H., RANSICK, A., MARTINEZ, P., BRITTEN, R. J. & DAVIDSON, E. H. J. M. O. D. 1994. Complexity and organization of DNA-protein interactions in the 5'-regulatory region of an endoderm-specific marker gene in the sea urchin embryo. 47**,** 165-186.

ZACHER, B., MICHEL, M., SCHWALB, B., CRAMER, P., TRESCH, A. & GAGNEUR, J. 2017. Accurate promoter and enhancer identification in 127 ENCODE and Roadmap Epigenomics cell types and tissues by GenoSTAN. *PloS one,* 12**,** e0169249.

ZENG, P., VAKOC, C. R., CHEN, Z., BLOBEL, G. A. & BERGER, S. L. 2006. In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. *Biotechniques,* 41**,** 694.

ZENG, W., BALL JR, A. R. & YOKOMORI, K. J. E. 2010. HP1: heterochromatin binding proteins working the genome. 5**,** 287-292.

ZENTNER, G. E., HENIKOFF, S. J. N. S. & BIOLOGY, M. 2013. Regulation of nucleosome dynamics by histone modifications. 20**,** 259.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M. & LI, W. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology,* 9**,** R137.

ZHANG, Y., WONG, C.-H., BIRNBAUM, R. Y., LI, G., FAVARO, R., NGAN, C. Y., LIM, J., TAI, E., POH, H. M. & WONG, E. J. N. 2013. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. 504**,** 306.

ZHAO, J., BACOLLA, A., WANG, G. & VASQUEZ, K. M. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences,* 67**,** 43-62.