Nucleome dynamics in normal and stalled development

by

Bo Hu

Department of Human Genetics
Faculty of Medicine and Health Sciences
McGill University, Montréal

August 2022

A thesis submitted to McGill University
in partial fulfillment of the requirements
of the degree of

Doctor of Philosophy

Abstract

Counted among the most complex machineries known to man, the cells that make up all living organisms lie at the foundation of life itself. Beyond traditional means of grossly assessing cellular morphology and composition, recent advances in sequencing-based assays have fuelled tremendous progress in understanding biological processes across varying scales. In particular, the importance of regulatory mechanisms that does not involve variation in actual genetic sequences – "epigenetics" – has become increasingly evident given their critical function in fine-tuning DNA compaction and folding. The dynamic epigenomic landscape thus not only underlies the diversity between cell types with specialized functions, but also distinguishes healthy and pathological states. Through the assembly of a 3D epigenome atlas of mouse germline development, we found that repressive domains and enhanced insulation maintains transcriptional integrity in the face of global DNA de-methylation and pervasion of enhancer-like signatures during epigenetic reprogramming in primordial germ cells. Subsequently in spermatogonia, these insulatory restraints are then removed en masse as global euchromatization and peripheral detachment of chromatin takes place in the preparation for meiotic entry. In contrast, we leveraged a compendium of 3D epigenomic profiles in brain tumours to reveal that a specific histone mutation, H3K27M, specifically leads to the formation of repressive loop structures via a reader of H3K27me3, canonical PRC1 (cPRC1). Following the validation of H3K27M-associated cPRC1 loops' impact in primary patient tumours, we further pinpointed this process as a therapeutic vulnerability – with the application of a cPRC1 inhibitor demonstrating the capacity to alleviate the oncogenic differentiation blockade. This thesis details how the systematic application of integrative multi-omics can dissect molecular determinants of health and disease as well as provide actionable insights towards the future development of targeted therapeutic strategies.

Résumé

Comptant parmi les machineries les plus complexes connues de l'homme, les cellules qui composent tous les organismes vivants sont à la base de la vie elle-même. Au-delà des moyens traditionnels d'évaluation grossière de la morphologie et de la composition cellulaires, les récentes avancées dans les analyses basées sur le séquençage ont alimenté des progrès considérables dans la compréhension des processus biologiques à différentes échelles. En particulier, l'importance des mécanismes de régulation qui n'impliquent pas de variation dans les séquences génétiques proprement dites - "épigénétique" - est devenue de plus en plus évidente, étant donné leur fonction critique dans le réglage fin de la compaction et du repliement de l'ADN. Le paysage épigénomique dynamique est donc non seulement à la base de la diversité des types de cellules aux fonctions spécialisées, mais il distingue également les états sains et pathologiques. Grâce à l'assemblage d'un atlas épigénome 3D du développement de la lignée germinale de la souris, nous avons constaté que les domaines répressifs et l'isolation renforcée maintiennent l'intégrité transcriptionnelle face à la déméthylation globale de l'ADN et à l'omniprésence de signatures de type enhancer pendant la reprogrammation épigénétique dans les cellules germinales primordiales. Par la suite, dans les spermatogonies, ces contraintes isolantes sont supprimées en masse lorsque l'euchromatisation globale et le détachement périphérique de la chromatine ont lieu pour préparer l'entrée méiotique. D'autre part, nous avons exploité un ensemble de profils épigénomiques 3D dans les tumeurs cérébrales pour révéler qu'une mutation spécifique de l'histone, H3K27M, conduit spécifiquement à la formation de structures en boucle répressive via un lecteur de H3K27me3, le PRC1 canonique (cPRC1). Après la validation de l'impact des boucles cPRC1 associées à H3K27M dans les tumeurs primaires des patients, nous avons identifié ce processus comme une vulnérabilité thérapeutique - l'application d'un inhibiteur de cPRC1 démontrant la capacité à atténuer le blocage de la différenciation oncogénique. Cette thèse montre comment l'application systématique de la multi-omique intégrative peut disséquer les déterminants moléculaires de la santé et de la maladie, et fournir des informations utiles pour le développement futur de stratégies thérapeutiques ciblées.

Contents

Abstract i

Résumé ii

Contents iii

List of abbreviations xiii

List of figures xiv

List of tables xvi

Acknowledgments xvii

Contribution to original knowledge xviii

Format of the thesis xix

Contribution of authors xx

Chapter 1: Introduction 1

Chapter 2: Nucleome programming for the foundation of totipotency in mammalian germline development 23

Chapter 3: H3K27me3 spreading organizes canonical PRC1 chromatin architecture to regulate developmental transcriptional program 98

Chapter 4: Discussion 153

Chapter 5: Conclusion 163

Chapter 6: References 165

Appendix Chapter A: Supplementary information for chapter 2 173

Appendix Chapter B: Additional supplementary information for chapter 2 185

Appendix Chapter C: Supplementary information for chapter 3 193

Appendix Chapter D: Copyright permissions 218

List of abbreviations

1D	One-dimensional
3D	Three-dimensional
4C	Circular chromosome conformation capture
AEBP2	AE Binding Protein 2
AM	Allosteric modulator
AP-1	Activator protein 1
AP-2	Activator protein 2
APC	Adenomatous polyposis coli
ARID1A	AT-Rich Interaction Domain 1A
ATAC-seq	Assay for transposase-accessible chromatin with sequencing
ATF2	Activating transcription factor-2
BAP1	BRCA1 Associated Protein 1
BRAF	B-Raf proto-oncogene, serine/threonine kinase
BRCA1	Breast cancer type 1 susceptibility protein
BRD4	Bromodomain-containing protein 4
bZIP	Basic Leucine Zipper
CAGE-seq	Cap analysis gene expression sequencing
Cas9	CRISPR-associated protein 9
CBX2	Chromobox protein homolog 2
CBX4	Chromobox protein homolog 4
CBX6	Chromobox protein homolog 6
CBX7	Chromobox protein homolog 7
CBX8	Chromobox protein homolog 8
CFTR	Cystic fibrosis transmembrane conductance regulator
CGI	CpG island

ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CM	Cardiac mesoderm
CP	Cortical plate
CPC	Cardiac progenitor cells
cPRC1	Canonical PRC1
CR1	Chicken repeat 1
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
DAPI	4′,6-diamidino-2-phenylindole
DAZL	Deleted In Azoospermia Like
DDX4	DEAD-Box Helicase 4
DMC1	DNA Meiotic Recombinase 1
DMRT1	Doublesex And Mab-3 Related Transcription Factor 1
DMRT6	Doublesex And Mab-3 Related Transcription Factor 6
DMRTA1	DMRT Like Family A1
DMRTA2	DMRT Like Family A2
DMRTC2	DMRT Like Family C2
DNA	Deoxyribonucleic acid
DNAme	DNA methylation
DNMT	DNA methyltransferase
DNMT3A	DNA Methyltransferase 3 Alpha
DNMT3B	DNA Methyltransferase 3 Beta
DNMT3C	DNA Methyltransferase 3 C
DNMT3L	DNA Methyltransferase 3 Like
E-P	Enhancer-promoter
EED	Embryonic Ectoderm Development
EHMT1	Euchromatic Histone Lysine Methyltransferase 1
EHMT2	Euchromatic Histone Lysine Methyltransferase 2
ENCODE	Encyclopedia of DNA Elements

EOMES	Eomesodermin
EpiLC	Epiblast-like cell
EPN	Ependymoma
EPOP	Elongin BC And Polycomb Repressive Complex 2 Associated Protein
eQTL	Expression quantitative trait loci
eRNA	Enhancer RNA
ERV1	Endogenous retrovirus-1
ERVK	Endogenous retrovirus-K
ERVL	Endogenous retrovirus-L
ESC	Embryonic stem cells
ESRRA	Estrogen related receptor alpha
ESRRB	Estrogen related receptor beta
ESRRG	Estrogen related receptor gamma
EVX2	Even-Skipped Homeobox 2
EZH2	Enhancer of zeste homolog 2
EZHIP	Enhancer of Zeste Homologs Inhibitory Protein
FIB1	Fibrillarin 1
FISH	Fluorescence in situ hybridization
G1	Gap/growth 1 phase
G2	Gap/growth 2 phase
GAM	Generalized additive model
GAM	Genome architecture mapping
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GATA4	GATA Binding Protein 4
Gb	Gigabasepairs
GBM	Glioblastoma
GFAP	Glial Fibrillary Acidic Protein
GLP	G9a-like protein
GSC	Germline stem cell
GSCLC	Germline stem cell-like cell

GSEA	Gene set enrichment analysis
GWAS	Genome-wide association study
GZ	Germinal zone
H2AK119ub	Histone H2A lysine 119 ubiquitinylation
H3K27ac	Histone H3 lysine 27 acetylation
H3K27M	Histone H3 lysine (27)-to-methionine
H3K27me	Histone H3 lysine 27 methylation
H3K27me3	Histone H3 lysine 27 tri-methylation
H3K36M	Histone H3 lysine (36)-to-methionine
H3K36me	Histone H3 lysine 36 methylation
H3K36me2	Histone H3 lysine 36 di-methylation
H3K36me3	Histone H3 lysine 36 tri-methylation
H3K4me	Histone H3 lysine 4 methylation
H3K4me1	Histone H3 lysine 4 mono-methylation
H3K4me3	Histone H3 lysine 4 tri-methylation
H3K79me2	Histone H3 lysine 79 di-methylation
H3K9me	Histone H3 lysine 9 methylation
H3K9me2	Histone H3 lysine 9 di-methylation
H3K9me3	Histone H3 lysine 9 tri-methylation
HAT	Histone acetyltransferase
hAT	hobo/Activator/Tam3
HDAC	Histone deacetylase
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HGG	High-grade glioma
Hi-C	High-throughput chromosome conformation capture
hiPSC	Human induced pluripotent stem cell
HNSCC	Head and neck squamous cell carcinoma
HOX	Homeobox
HOXD	Homeobox D cluster
HP1	Heterochromatin protein 1

HP1a	Heterochromatin Protein 1 alpha
ICM	Inner cell mass
IF	Immunofluorescence
iPSC	Induced pluripotent stem cell
IRF3	Interferon regulatory factor 3
JARID2	Jumonji And AT-Rich Interaction Domain Containing 2
K9MTase	Histone H3 lysine 9 methyltransferase
Kb	Kilobasepairs
KDM	Histone lysine demethylase
KLF12	Kruppel Like Factor 12
KLF3	Kruppel Like Factor 3
KLF4	Kruppel Like Factor 4
KMT	Histone lysine methyltransferase
KO	Knockout
L1	LINE1
L1Md	L1 in Mus domesticus
L2	LINE2
LAD	Lamina associated domain
LINE	Long interspersed nuclear elements
LINE1	Long interspersed nuclear element 1
lncRNA	Long noncoding RNA
LOCK	Large organized chromatin lysine domains
LTR	Long terminal repeat
MACS	Model-based Analysis of ChIP-seq
MAEL	Maelstrom Spermatogenic Transposon Silencer
MaLR	Mammalian apparent LTR-retrotransposon
Mb	Megabasepairs
MBD	Methyl-CpG-binding domain protein
MBP	Maltose-binding protein
mCG	Methyl Cpg

MCORE	Multiscale correlation evaluation
MED1	Mediator Complex Subunit 1
mESC	Mouse embryonic stem cell
MII	Metaphase II
MIR	Mammalian-wide interspersed repeat
MLL2	myeloid/lymphoid or mixed-lineage leukemia 2
mPGCLC	Mouse primordial germ cell-like cell
mRNA	Messenger RNA
MS	Mass spectrometry
MSCI	Meiotic sex chromosome inactivation
MULE	Mutator–like transposable element
ncPRC1	Non-canonical PRC1
ncRNA	Noncoding RNA
NET-CAGE	Native elongating transcript–cap analysis of gene expression
NFE2	Nuclear Factor, Erythroid 2
NFE2L2	NFE2 Like BZIP Transcription Factor 2
NGS	Next generation sequencing
NIBPL	Nipped-B-Like Protein
NMR	Nuclear magnetic resonance
NPC	Neural progenitor cell
NSD1	Nuclear Receptor Binding SET Domain Protein 1
NSD2	Nuclear Receptor Binding SET Domain Protein 2
OCT4	Octamer-Binding Protein 4
OL	Oligodendrocyte
OPC	Oligodendrocyte progenitor cell
PAD	Polycomb-associating domains
PCA	Principal component analysis
PcG	Polycomb Group
PCH	Pericentromeric heterochromatin
PCL2	Polycomb-like 2

PCM	Primitive cardiomyocytes
PDS5	Precocious dissociation of sisters 5
PDX	Patient-Derived Xenograft
PFA	Posterior fossa group A
PFB	Posterior fossa group B
PGC	Primordial germ cell
PGCLC	Primordial germ cell-like cell
PHC1	Polyhomeotic Homolog 1
PHC2	Polyhomeotic Homolog 2
PHC3	Polyhomeotic Homolog 3
PHD	Plant homeodomain
pHGG	Pediatric high grade glioma
piRNA	Piwi-interacting RNA
PMD	Partially methylated domain
Pol II	RNA polymerase II
Pol2Ser5P	RNA polymerase II phosphorylated at serine 5
PRC1	Polycomb Repressive Complex 1
PRC2	Polycomb Repressive Complex 2
PRDM13	PR/SET Domain 13
PSC	Pluripotent stem cell
PTM	Post-translational modification
RGC	Radial glial cell
RING1B	Really Interesting New Gene 1B
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAPII	RNA polymerase II
RPB1	RNA polymerase II subunit A
RRHO	Rank-rank hypergeometric overlap
rRNA	Ribosomal RNA
RTE	Retrotransposable element

RYBP	RING1 and YY1-binding protein
SA1	Stromal antigen 1
SA2	Stromal antigen 2
SA3	Stromal antigen 3
SCC1	Sister Chromatid Cohesion 1
SCML2	Sex Comb On Midleg Like 2
scRNA	Small cytoplasmic RNA
scRNA-seq	Single cell RNA-seq
SE	Super enhancer
SET	Suppressor Of Variegation 3-9, Enhancer-of-zeste and Trithorax
SETD2	SET domain containing 2
SETDB1	SET Domain Bifurcated Histone Lysine Methyltransferase 1
SHH	Sonic hedgehog
SIM1	Single-minded homolog 1
SINE	Short interspersed nuclear element
SMC1	Structural maintenance of chromosomes protein 1
SMC1A	Structural maintenance of chromosomes protein 1A
SMC3	Structural maintenance of chromosomes protein 3
snRNA	Small nuclear RNA
SOHLH2	Spermatogenesis And Oogenesis Specific Basic Helix-Loop-Helix 2
SOX10	SRY-box transcription factor 10
SOX2	SRY-box transcription factor 2
SPRITE	Split-pool recognition of interactions by tag extension
srpRNA	Signal recognition particle RNA
SRSF2	Serine and arginine rich splicing factor 2
SRY	Sex-determining region Y
SSC	Spermatogonia stem cell
STAG1	Stromal antigen 1
STAG2	Stromal antigen 2
STAG3	Stromal antigen 3

STAT3	Signal Transducer And Activator Of Transcription 3
STE	Supratentorial
SUV39H1	Suppressor Of Variegation 3-9 Homolog 1
SUV39H2	Suppressor Of Variegation 3-9 Homolog 2
SUZ12	Suppressor Of Zeste 12 Homolog
SYCP1	Synaptonemal Complex Protein 1
SYCP2	Synaptonemal Complex Protein 2
SYCP3	Synaptonemal Complex Protein 3
TAD	Topologically-associating domain
TcMar	Tc1/Mariner
TE	Transposable element
TEAD3	TEA domain transcription factor 3
TERF1	Telomeric repeat binding factor 1
TES	Transcription end site
TET	Ten-eleven translocation methylcytosine dioxygenase
TET1	Ten-eleven translocation methylcytosine dioxygenase 1
TET2	Ten-eleven translocation methylcytosine dioxygenase 2
TET3	Ten-eleven translocation methylcytosine dioxygenase 3
TF	Transcription factor
TFAP2C	Transcription Factor AP-2 Gamma
TOP2A	DNA Topoisomerase II Alpha
tRNA	Transfer RNA
TSS	Transcription start site
UHC	Unsupervised hierarchical clustering
UMAP	Uniform Manifold Approximation and Projection
VCM	Ventricular cardiomyocytes
vPRC1	Variant PRC1
WAPL	Wings apart-like protein homolog
WGBS	Whole genome bisulfite sequencing
WT	Wild-type

YY1	Yin Yang 1
ZGA	Zygotic genome activation
ZIC2	Zinc finger of the cerebellum 2
ZIC3	Zinc finger of the cerebellum 3
ZP3	Zona pellucida glycoprotein 3

List of figures

1.1 Central dogma of molecular biology . . . 1
1.2 Mapping disease genes . . . 2
1.3 Interpreting GWAS hits . . . 3
1.4 Epigenetic mechanisms . . . 4
1.5 Gene regulation through chromatin accessibility . . . 4
1.6 Waddington's epigenetic landscape . . . 5
1.7 Tissue-specific cis-regulation of enhancers . . . 6
1.8 Bivalent chromatin dynamics . . . 7
1.9 Epigenetic hallmarks of different genomic regions . . . 7
1.10 Functions of epigenetic modifiers . . . 8
1.11 Multi-scale nucleome organization . . . 9
1.12 Defective higher-order genome architecture . . . 9
1.13 Biomolecular condensates of various scales . . . 10
1.14 Integrative multi-omics . . . 11
1.15 Epigenome mapping . . . 12
1.16 Nucleome mapping . . . 13
1.17 Transcriptome mapping . . . 14
1.18 Mouse germ-cell development . . . 15
1.19 Mammalian embryonic epigenome remodelling . . . 16
1.20 3D epigenome dynamics during early embryogenesis . . . 17
1.21 Cancer-linked epigenetic genes . . . 18
1.22 Oncohistone H3 mutations . . . 18
1.23 PRC2 inhibition by H3K27M/EZHIP . . . 19
1.24 Reciprocity of H3K27 vs H3K36 methylation . . . 20

2.1 3D genome programming . . . 80
2.2 Epigenome profiles and CTCF insulation . . . 83
2.3 Open-site characterizations and CTCF release . . . 85
2.4 Generation of minimal LADs . . . 87
2.5 Heterochromatin re-organization . . . 89
2.6 Mechanism of PMD formation via balancing H3K36me2 vs H3K9me-marked LADs and Y chromosome hypomethylation . . . 91
2.7 Nucleome differences between GSCs and GSCLCs . . . 93
2.8 A model for the nucleome programming during mouse germ-cell development . . . 96

3.1 Restricted H3K27me3 occurs in brain tumor and developmental contexts and associates with stronger distal interactions between CpG islands . . . 138
3.2 Confined H3K27me3 induces cPRC1 concentration leading to compaction of Polycomb bodies in brain tumors . . . 140
3.3 Loss of NSD1 induces H3K27me3 spreading and dissolution of Polycomb bodies in PSCs 142

3.4 Compaction of Polycomb bodies is associated with the repression of target genes 145
3.5 Canonical PRC1 targets H3K27M-repressed genes in primary brain tumors and patient-derived xenografts 147
3.6 Upregulation of cPRC1 target genes upon the removal of cPRC1-mediated chromatin compaction 150
3.7 Model of the link between H3K27me3 spread and cPRC1 chromatin architecture leading to altered developmental programs 152

4.1 Cycle of gametogenesis 154
4.2 Functions of cohesin in 3D genome organization 155
4.3 Germline nucleome dynamics 156
4.4 Polycomb-associated chromatin architecture 160
4.5 Chromatin-targeting drugs and probes 161

A.1 Investigation of global nuclear architecture dynamics through Hi-C and FISH 174
A.2 Quantitative epigenome analysis by mass spectrometry and chromatin accessibility analysis by ATAC-seq 176
A.3 Exploration of cis-regulatory element by NET-CAGE combined with Hi-C and comparison against public Hi-C datasets 178
A.4 Open site chromatin state dynamics and differential CTCF binding throughout germ cell differentiation 180
A.5 Inter-species comparison of germ-cell specific chromatin structure and characterization of H3K9me3-enriched repeats 182

B.1 Hi-C analysis with other lineages 186
B.2 PMD analysis with epigenome and higher-order chromatin structure 188
B.3 Methylome analysis on Y chromosome using alternative mapping method 189
B.4 Nucleome differences between functional GSCs and spermatogenically-impaired GSCLCs 190

C.1 Validation of framework for ChIP-seq signal breadth quantification 194
C.2 Commonalities and differences between K27M and K27M-KO cells 196
C.3 Subgroup-specific 3D genome features 198
C.4 Relationship between restricted H3K27me3 and long-range inter-CGI interaction in additional contexts 200
C.5 Characterizing determinants of long-range polycomb-mediated interaction 202
C.6 Consistent H3K27me3 and cPRC1 confinement across diverse backgrounds 204
C.7 Weakening polycomb-mediated chromatin architecture in pluripotent stem cells through loss of NSD1 206
C.8 Properties cPRC1 sites as compared to other chromatin states in stem cells 209
C.9 Properties cPRC1 sites as compared to other chromatin states in glioma cells 212
C.10 Validation of cPRC1 looping and target repression in primary tumours 214
C.11 Decreased cPRC1 compaction reliably alleviates developmental blockade 216

List of tables

2.1 Reagents and Tools 46

A.1 Sequencing summary 184
A.2 Histone modification abundances 184
A.3 Motif enrichment results 184
A.4 Cluster annotations 184
A.5 Annotation overlap results 184
A.6 Pathway association results 184

Acknowledgments

I would like to first and foremost express my sincere gratitude to Professors Jacek Majewski and Mitinori Saitou for graciously providing me with precious opportunities to work on a variety of exciting projects and ample chance to grow both inside and outside the lab under very favorable circumstances. Starting out with no bioinformatics experience, I would not have been immersed in the domain of computational biology nearly as quickly without Dr. Majewski's tireless guidance and sustained confidence in me. Dr. Saitou's stimulating discussions and infectious enthusiasm were critical bringing me, an outsider to developmental biology, quickly up to speed and soon begin contributing productively. I would also like to thank members of my supervisory committee, Professors Guillaume Bourque and Celia Greenwood, for their consistent and constructive feedback throughout the course of my studies, without which my projects would not have progressed as smoothly. Furthermore, I would like to thank Professors Fumihiko Matsuda and Ryo Yamada for their benevolent academic assistance and generous personal hospitality, making my time in Kyoto a wonderful experience that I will be sure to cherish and recall fondly.

These projects would not have been possible without the tremendous labour of fantastic collaborators with whom I've had the privilege and pleasure of working together. In particular, I'm deeply appreciative of Masahiro Nagano and Brian Krug for their partnership in spearheading these studies from the experimental side and fundamentally enabling my computational efforts; the time they charitably expended to brainstorm together with me and deliberate over the experimental design as well as data interpretation were vital in my swift acquisition of domain knowledge and allowing us to effectively tackle challenging problems. I'm also grateful to other members of the Majewski and Saitou groups as well as our internal and external collaborators who've contributed critical pieces of data and provided decisive advice that led to significant improvements of the manuscripts. Furthermore, I would like to acknowledge that the work described in this thesis is in part supported by financial support from the Canadian Institutes of Health Research, le Fonds de recherche du Québec – Santé, and the Japan Student Services Organization.

Last but not least, I'm thankful towards for family for their unwavering support.

Contribution to original knowledge

2 Nucleome programming for the foundation of totipotency in mammalian germline development

- Identified most variable histone modifications along male murine germline development and clarified regulatory impacts
- Established insulation enhancement as a novel protective mechanism during epigenetic reprogramming
- Uncovered continual maturation and euchromatization of 3D genome organization throughout germ cell development
- Pinpointed functionally important spermatogonial-specific nuclear architectural features that are conserved across species

3 H3K27me3 spreading organizes canonical PRC1 chromatin architecture to regulate developmental transcriptional program

- Detailed the effect of H3K27me3 on long-range interactions across diverse contexts with epigenetic dysregulation
- Revealed association of variable polycomb domain coverage with cPRC1 concentration as a biomarker of specific brain tumor subtypes
- Associated the degree of distal polycomb aggregation with developmental progression via transcriptional regulation
- Unveiled H3K27me3 readers as therapeutic vulnerability to modulate differentiation capacity

Format of the thesis

This thesis is presented in the manuscript-based format for a doctoral thesis according to the guidelines of the Department of Human Genetics at McGill University and is organized into five chapters. CHAPTER 1 introduces the regulatory importance of 3D epigenomic properties to proper cellular function, provides relevant background on the epigenome remodelling events that takes place during germline development and in K27M GBMs, and outlines the overarching goals of this thesis. CHAPTER 2 describes an effort to chart nucleome dynamics during male gametogenesis and highlights novel 3D epigenomic features of key stages, published in the EMBO Journal. CHAPTER 3 is a manuscript in preparation, which outlines the impact of polycomb domain re-organization on 3D genome organization and developmental progression. CHAPTERS 4 AND 5 present an overall discussion as well as general conclusions and future directions.

Contribution of authors

2 Nucleome programming for the foundation of totipotency in mammalian germline development

I conceived the project and designed the experiments together with Masahiro Nagano, Shihori Yokobayashi, and Mitinori Saitou. I performed the vast majority of the data analysis and supervised the remainder. I created all figures, with help from Masahiro Nagano for select schematics and images.

Masahiro Nagano assisted with overall data analysis. Yusuke Imoto, Killian Meehan, and Yasuaki Hiraoka assisted with polymer simulations. Masahiro Nagano performed all cell culture and induction with assistance from Hiroshi Ohta, Yukiko Ishikura and Yoshiaki Nosaka. Masahiro Nagano performed immunofluorescence and its analysis with assistance from Hiroshi Ohta, Naofumi Kawahira and Ken Mizuta. Masahiro Nagano and Fumiya Umemura performed western blot and its analysis with assistance from Yoshiaki Nosaka, Sakura Shimizu, and Yoji Kojima. Masahiro Nagano performed FISH with assistance from Ikuhiro Okamoto. Masahiro Nagano performed histone extraction and Mariel Coradin performed mass spectrometry under the supervision of Benjamin A Garcia. Masahiro Nagano and Akitoshi Yamamura performed ChIP-seq with assistance from Shihori Yokobayashi and Takuya Yamamoto. Masahiro Nagano performed ATAC-seq with assistance from Shihori Yokobayashi, Hiroki Ikeda, and Takuya Yamamoto. Masahiro Nagano performed in situ Hi-C with assistance from Shihori Yokobayashi, Roman Stocsits, Gordana Wutz, Kikue Tachibana, Jan-Michel Peters, and Leonid A Mirny. Masahiro Nagano performed NET-CAGE with assistance from Tomoko Kasahara under the supervision of Yasuhiro Murakawa. Shihori Yokobayashi, Jacek Majewski, and Mitinori Saitou supervised the project.

3 H3K27me3 spreading organizes canonical PRC1 chromatin architecture to regulate developmental transcriptional program

I conceived the project and designed the experiments together with Brian Krug, Jacek Majewski, and Nada Jabado. I performed the vast majority of the data analysis and supervised the remainder. I created all figures, with help from Brian Krug for select schematics and images.

Haifen Chen, Xiao Chen, Michael Johnston, and Marco Gallo assisted the analysis of ChIP-seq and Hi-C datasets. Nisha Kabir and Claudia Kleinman assisted the analysis of single-cell RNA-seq datasets. Brian Krug, Kristjan Gretarsson, Shriya Deshmukh, Elias Jabbour, Maud Hulswit, Ashot Harutyunyan, John Lee, and Michael D Taylor contributed to generating genomics datasets. Brian Krug, Kristjan Gretarsson and Shriya Deshmukh performed functional experiments and generated cell lines models. Damien Faury assisted the generation of single-cell RNA-seq datasets, Caterina Russo assisted performing of xenograft experiments and collection. Chao Lu, Jacek Majewski, and Nada Jabado supervised the project.

Epigenetics imparts a fundamental regulatory system beyond the sequence information of our genetic code and emphasizes that "Mendel's gene is more than just a DNA moiety."

Thomas Jenuwein & C. David Allis

Chapter 1

Introduction

Peering into genome regulation through a 3D epigenomics lens

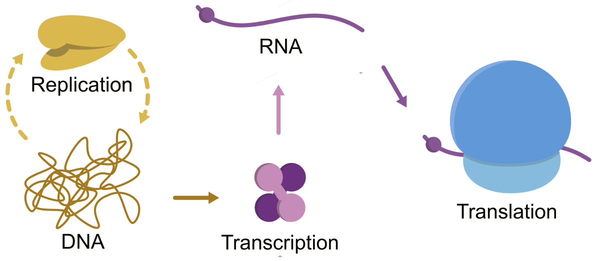


Figure 1.1: Central dogma of molecular biology. Reproduced from Costello & Badran [1]

Deoxyribonucleic acid (DNA), the code for life itself, lies at the foundation of biology through its role as the source of information flow: with the contents of DNA copied via transcription into ribonucleic acid (RNA), which is in turn translated into proteins that ultimately function as critical cogs in the elaborate cellular machineries powering every living organism (Fig. 1.1).[1] This classic view centered around sequences encoding proteins ("genes") has since been formalized as the "central dogma of molecular biology" and has dominated mainstream genetics since the establishment of DNA as the essence of heritability (or "transforming material") in the 1940s,[2] with most efforts focused on understanding how genetic sequence variations contribute to phenotypic differences ranging from subtle traits to complex disorders.[3] In particular, the elegant simplicity of considering genes as discrete units, where variant forms simply lead to a reduction

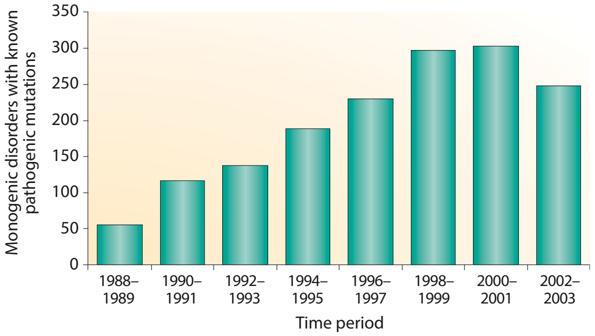


Figure 1.2: Mapping disease genes. Number of pathogenic mutations causing monogenic disorders recorded by OMIM at various times. Reproduced from Antonarakis & Beckmann [4]

in functional protein quantity, have allowed the application of the long-standing concepts such as Mendelian inheritance and pedigree analysis to achieve great successes; using these approaches to pinpoint genes whose loss alone suffices for pathogenesis, hundreds of disease-causing genes were already resolved by the early 2000s (Fig. 1.2).[4] Among one of the most well-known examples, the unambiguous association between mutations in the *CFTR* (cystic fibrosis transmembrane conductance regulator) gene with cystic fibrosis, and the accompanying direct disease mechanism, facilitated concentrated efforts that enabled breakthrough treatments substantially improving the lives of patients suffering from this debilitating affliction.[5]

Alongside the initial assembly of all sequences that constitute human DNA (the "human genome"), genome-wide association studies (GWAS) bursted to the forefront of genetics following the turn of the millennium through demonstrating the capacity to comprehensively and statistically assess the association between traits and variant across the entire genome. From these investigations it then became apparent that more complex models are required to unravel genome regulation. In particular, as protein-coding sequences only constitute less than 2% of the human genome, the vast majority of implicated sequence variants fall outside of protein-coding regions ("intergenic"), further complicating direct interpretation of their downstream impacts. Although traditionally considered "junk DNA", it's become apparent that non-coding intergenic regions harbour indispensable regulatory elements that can fine-tune the activity of nearby target genes through a variety of mechanisms; these elements range from "enhancers" and "silencers" involved in modulating the transcription of DNA into RNA (also known as "gene expression"), to "insulators" demarcating boundaries that segregate the genome into discrete domains subject to co-regulation. While signatures of evolutionary conservation were historically used to gauge functional importance of DNA sequences, detailed genomic element annotations have more recently been empowered by extensive functional profiling initiatives across multiple cellular contexts covering the gamut of health and disease (Fig. 1.3).[6]

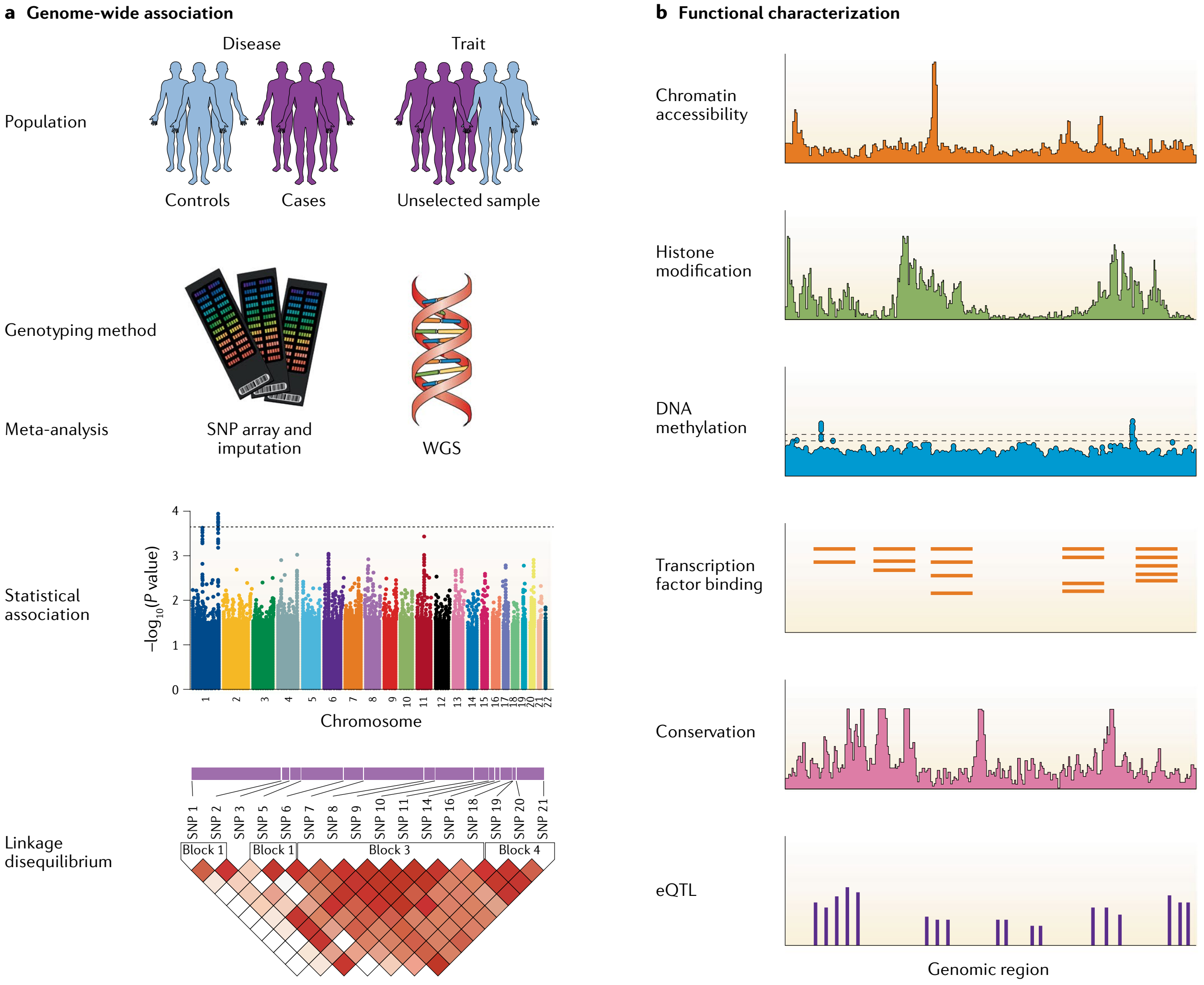


Figure 1.3: Interpreting GWAS hits. a. Set-up of typical GWAS studies; b. Means to interpret non-coding variants. Reproduced from Tam *et al.* [6]

Functional genome profiling often entails investigation into "*epi*genetic" modifications that comprise an information layer "on top" of DNA sequences, directing gene expression programs without genetic alterations – akin to diacritics modulating the tone and meaning of letters.[7] Epigenetic modifications broadly impart their influence by governing the packaging of DNA, as molecules that total meters in length are encased into micrometer-wide cell nuclei (Fig. 1.4).[8] DNA, adenines and cytosines in particular, can be chemically modified through the addition of "methyl" chemical moieties in a process called "methylation", chemically distinguishing modified nucleobases from unmodified ones and providing substrates for methyl-binding domain-containing proteins while sterically hindering the access of other regulators.[9] Stepping up the organizational hierarchy, DNA strands are then wrapped around histone proteins to form fiber-like structures called "chromatin". These histones can also take on post-translational modification at

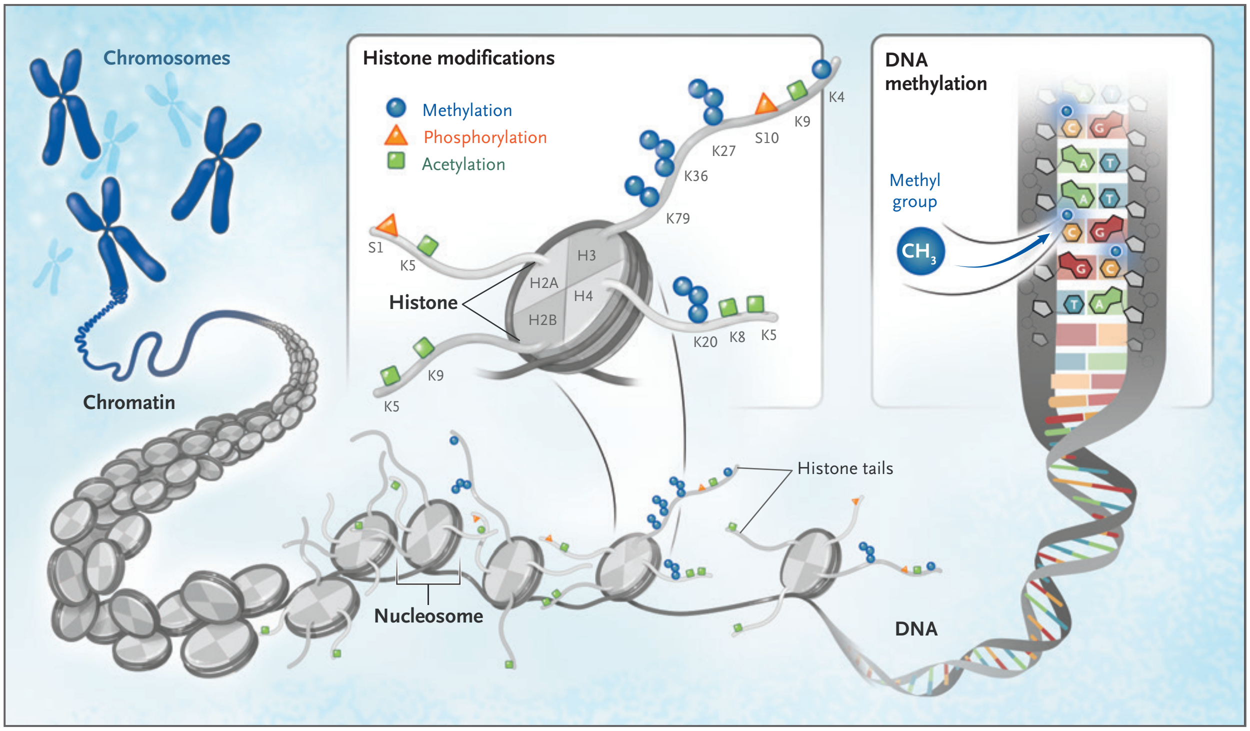


Figure 1.4: Epigenetic mechanisms. Reproduced from Bates [8]

specific residues, usually concentrated on the tail sections that protrude outwards, and acquire contrasting biochemical properties.[10] It's been found that particular combination of histone modifications can be viewed as markers of distinct genome functions, with some marking active regions ("euchromatin") and others silent ("heterochromatin"), corresponding to the degree of DNA compaction. Whereas euchromatic regions of loose DNA packaging are said to be more "accessible" to the binding of DNA-recognizing proteins such as those regulating transcription ("transcription factors"), closed heterochromatin can physically occlude the access of such factors (Fig. 1.5).[11]. It is therefore thought that the incredible diversity of cell types within multi-cellular organisms, despite them all stemming from a

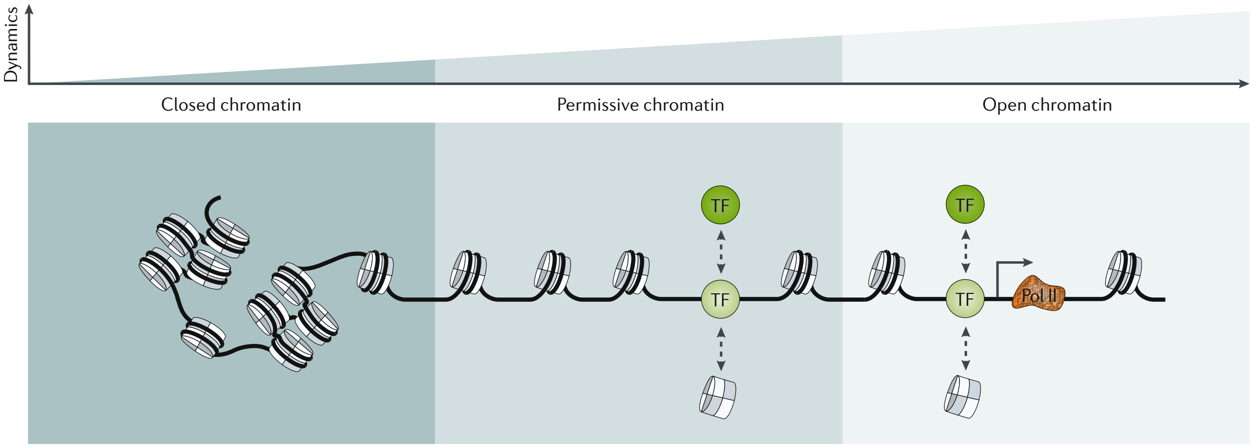


Figure 1.5: Gene regulation through chromatin accessibility. Reproduced from Klemm *et al.* [11]

single zygote and sharing the very same genetic blueprint, necessarily involves epigenetic mechanisms.[12] Indeed, the very concept of an "epigenetic landscape" stems from the works of Conrad Waddington in the mid 20th century, analogizing development and cell fate decisions as traversal between discrete cellular states.[13] With the advent of functional genome profiling, the contemporary model of the Waddington landscape is thus one where cell type-specific gene regulatory programs, precisely facilitating the expression of genes while keeping others silent, derives from the fine-tuning of epigenetic processes. And it is through the inherent characteristics of various epigenetic modifications as well as their writers and readers that complex interaction networks arise: involving crosstalks, cascades, and feedback loops that presents a daunting yet bounteous opportunity to scrupulously disentangle these interconnections that make up the "epigenome".

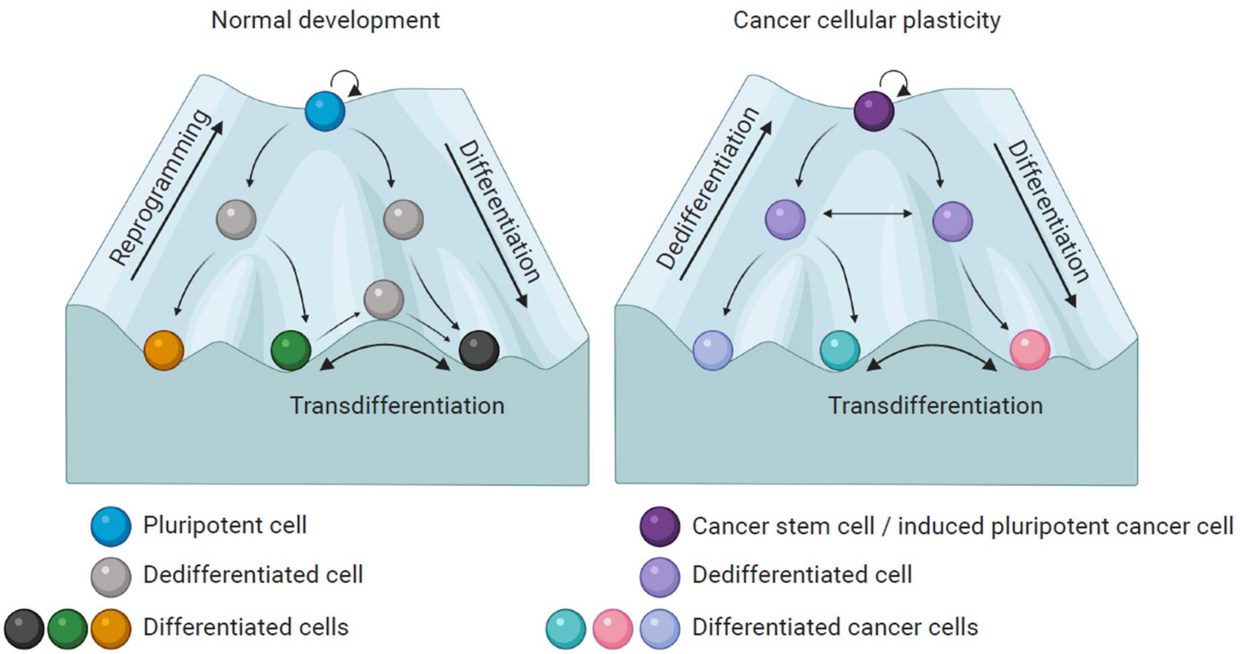


Figure 1.6: Waddington's epigenetic landscape depicting normal development and carcinogenesis. Reproduced from Granados *et al.* [14]

Beyond differentiation in development, the epigenome's malleability has also been noted as a useful model for disease: while malignant cells may co-opt epigenetic plasticity for transformation and escape physiological surveillance, this property can be likewise exploited clinically to alter and sensitize cells for otherwise untenable therapeutic strategies (Fig. 1.6).[14] Therefore, the initial step towards uncovering such epigenomic vulnerabilities will necessitate first and foremost the acquisition of comprehensive molecular portraits across both physiological and pathological settings. As an example, tissue-specific regulatory elements identified via epigenetic profiles were reported to not only fall near genes belonging to tissue-specific expression programs, but the very same regions have also exhibited strong enrichment for variants previously implicated with phenotypes known to affect the associated tissue (e.g., GWAS hits for psychiatric traits in regions specifically accessible in neurons alone).[15] On the other hand, it's also been shown in assorted cancers that epigenetic information such as DNA methylation can prove invaluable in uncovering distinct molecular subtypes – substantially accelerating the development of targeted strategies by enabling finer patient stratification.[16]

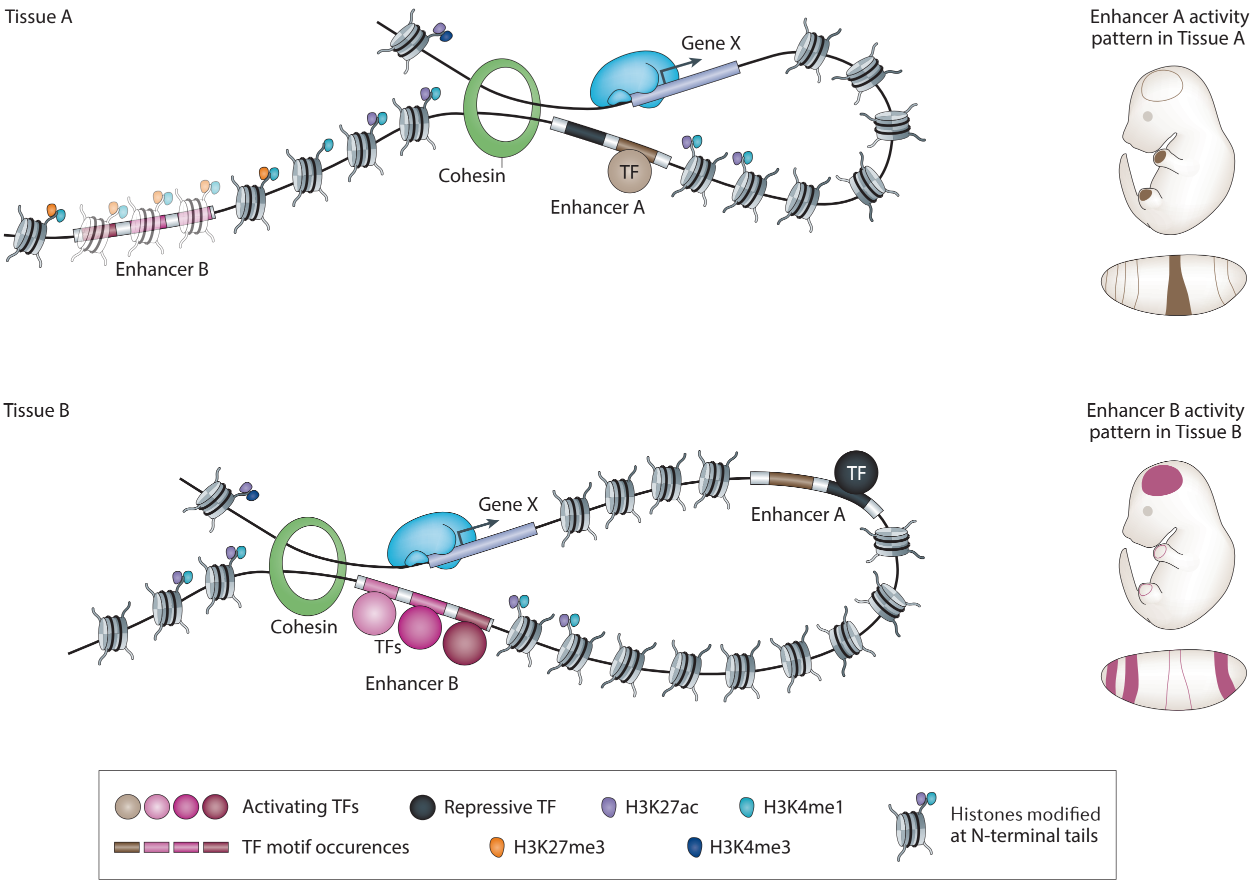


Figure 1.7: Tissue-specific cis-regulation of enhancers. Reproduced from Shlyueva *et al.* [17]

The expansive repertoire through which epigenetic mechanisms affect cellular processes range from controlling the binding of vital transcription factors that may modulate the expression of nearby genes to facilitating distal interaction between chromatin segments, and the specific identity of these processes have been linked with a combination of several telltale epigenetic markers (Fig. 1.7).[17] For instance, active elements are typically surrounded by specific modified forms of histone H3 – acetylated at 27th lysine residue (H3K27ac) and/or methylated at lysine 4 of H3 (H3K4me); it has also been observed that whereas

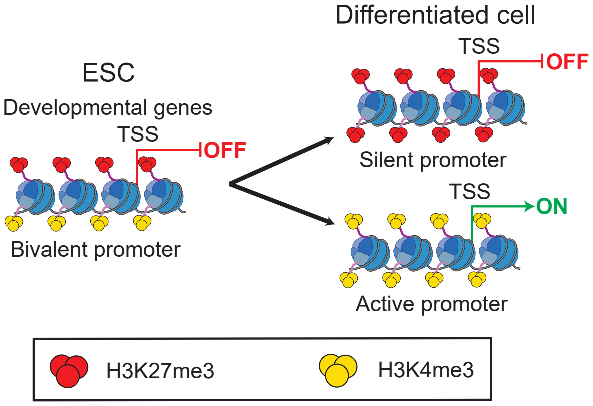


Figure 1.8: Bivalent chromatin dynamics. Reproduced from Blanco *et al.* [18]

tri-methylation of H3K4 (H3K4me3) is enriched at "promoter" elements upstream of expressed genes' transcription start sites, mono-methylated H3K4 (H3K4me1) instead marks "enhancer" elements that loop around to engage active promoters and boost transcription despite their potentially large separation on the linear genome spanning tens to hundreds of kilobases (kb). While the exact mechanisms remain under debate, it is thought that enhancers facilitate enhanced transcription of nearby genes ("in *cis*") through increasing concentration of transcription factors that are conducive to higher expression in the immediate vicinity of promoters.[19] In contrast, H3K27me3 is generally associated with repressed cis-regulatory elements (enhancers and promoters) through the inducing heterochromatinization. Yet H3K27me3 is not always distributed in a mutually exclusive manner with H3K4me, with their co-occurrence often designating regulatory elements of developmental genes, corresponding to a "poised"/"bivalent" state that may only resolve to fully active and repressed depending on subsequent cues such as developmental signals (Fig. 1.8).[18] At grander scales, other modifications can also broadly pattern heterochromatic (e.g., H3K9me) and euchromatic (e.g., H3K36me) domains that are up to hundreds of kbs or even megabases (mb) in widths. Therefore, deconvolving highly combinatorial chromatin states in a genome-wide manner generally demands the simultaneous survey of a varied set of modifications. (Fig. 1.9).[10] In addition to the

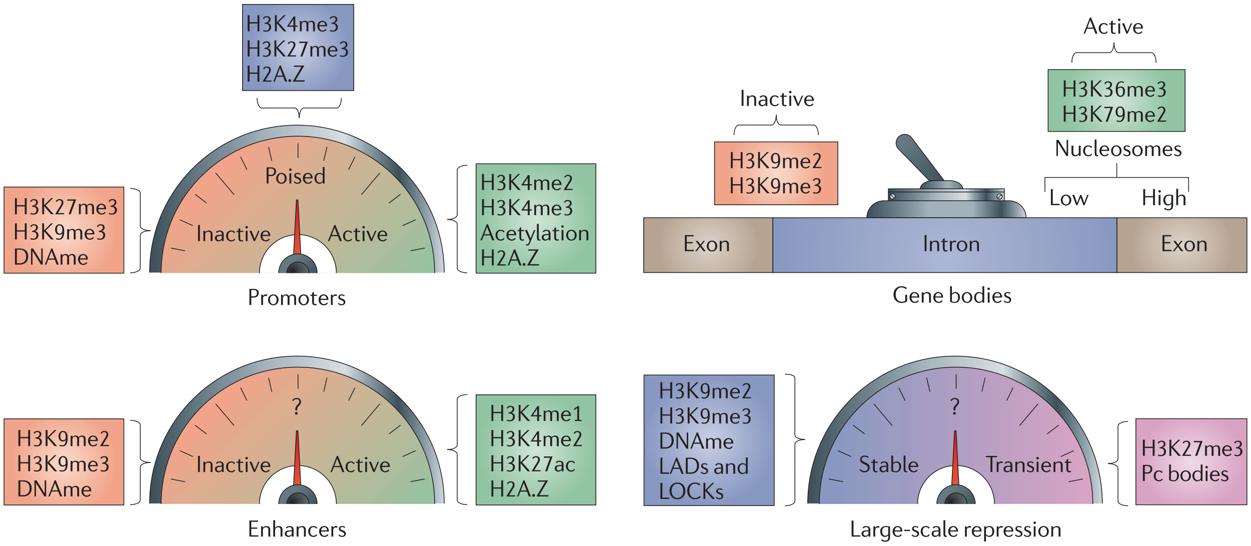


Figure 1.9: Epigenetic hallmarks of different genomic regions. Reproduced from Zhou *et al.* [10]

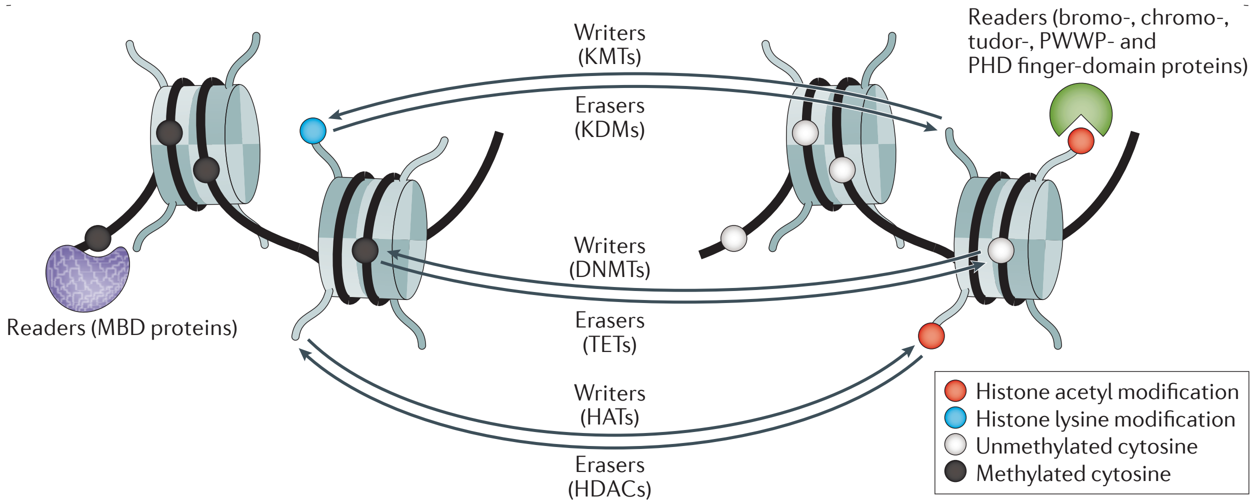


Figure 1.10: Functions of epigenetic modifiers. Reproduced from Jones *et al.* [20]

modifications themselves, further complications arise from the plethora of chromatin modifiers that read, write, and erase epigenetic modifications, as the same complex can often be endowed with multiple functions through its various subunits (Fig. 1.10).[20] For example, Polycomb Repressive Complex 2 (PRC2), a H3K27 methyltransferase (i.e., writer), recognizes H3K27me3 through the Embryonic Ectoderm Development (EED) subunit for allosteric activation, while the Enhancer of zeste homolog 2 (EZH2) subunit possessing methyltransferase catalytic activity simultaneously recognizes unmethylated H3K36 through a separate domain – partially explaining the observed general exclusion of H3K27me3 from H3K36me2/3-decorated chromatin.[21] Taken together, the collection of epigenetic markers must be considered in conjunction with the accompanying set of modifiers to decipher how proper epigenome dynamics foster normal physiological functioning as well as how epigenomic perturbations engender dysfunction.

But a conventional view of the epigenome as an 1D scaffold cannot account for how chromatin fiber is further arranged in 3D space to ultimately fit within the tiny nuclei of cells that are magnitudes smaller in width than the length of chromatin fiber simply stretched in full; and rather than behaving as an ideal chain in the context of polymer physics, it's been shown that chromatin adopts non-random 3D conformations in reality, subject to regulation at multiple scales.[22] At the highest level, the architecture of chromatin within the nucleus ("nucleome") is dictated by landmarks such the nucleolus and nuclear lamina that serve as scaffolds around which particular genomic regions attach and take on specific roles including transcriptional factories and repressive hubs (Fig. 1.11).[23] Chromatin itself, with myriad epigenetic modifications, likewise influences genome folding patterns through passive forces

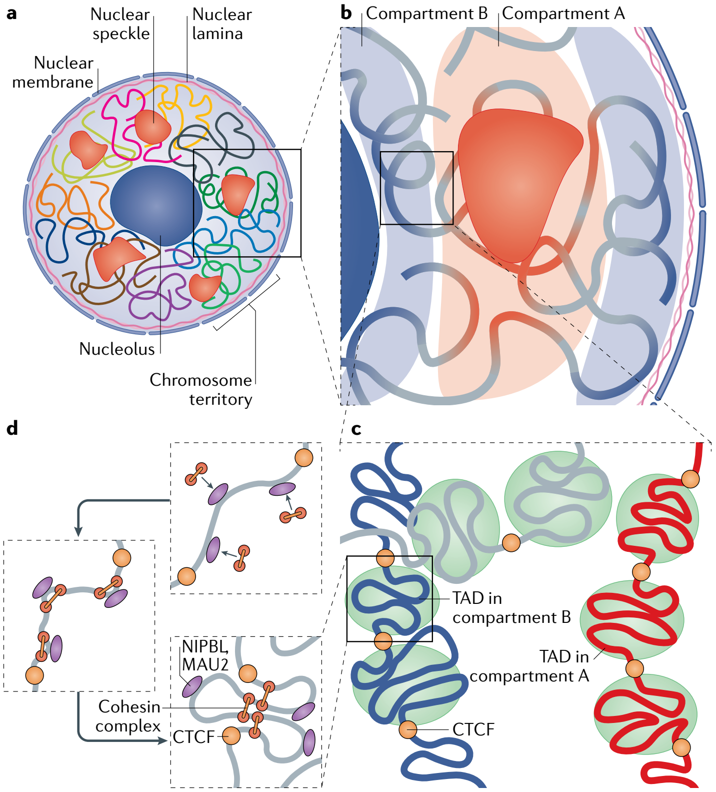


Figure 1.11: Multi-scale nucleome organization. a. Chromosomal territories organize around nuclear landmarks during interphase; b. Euchromatin and heterochromatin separate into distinct compartments; c. Preferentially self-interacting domains exist within compartments; d. Active loop extrusion by cohesin in concert with boundary CTCF elements contribute to domain formation and sub-domain looping. Reproduced from Zheng & Xie [23]

such as liquid-liquid phase separation. For instance, H3K9me-associated constitutive heterochromatin was mechanistically linked with a reader protein of H3K9me, HP1, organizing marked sections of chromatin into liquid-like droplets.[24]. Active processes involving energy consumption also exists as a major force, with a prime example being loop extrusion, where ring-like cohesin complexes act as motors to extrude chromatin loops until their encounter with boundaries bound by proteins CCCTC-binding factor (CTCF).[25] As they organize the very same chromatin fiber, the active and passive forces of disparate origins seldom act in isolation, leading to complex interactions. For example, active cohesin-mediated loop extrusion is known to antagonize a host of passive processes, including compartmentalization of the genome into euchromatin and heterochromatin as well as other condensates such as H3K27me3-enriched polycomb bodies.[27,28] Proper 3D genome organization and nuclear architecture, a healthy nucleome, has also been shown to serve essential physiological functions in view of the wide

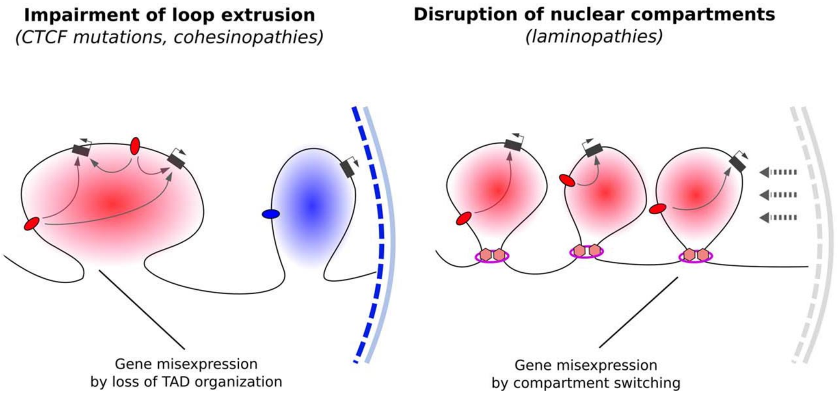

Impairment of loop extrusion
(CTCF mutations, cohesinopathies)
Disruption of nuclear compartments
(laminopathies)
Gene misexpression by loss of TAD organization
Gene misexpression by compartment switching

Figure 1.12: Defective higher-order genome architecture. Reproduced from Anania & Lupiáñez [26]

range of diseases associated with defective architectural proteins such as laminopathies and cohesinopathies, often linked to developmental and cognitive anomalies (Fig. 1.12).[26]. Therefore, a simultaneous appreciation of both local one-dimensional (1D) compaction and three-dimensional (3D) global organization, at multiple scales (Fig. 1.13),[29] is necessary to grasp chromatin dynamics and establish a strong grasp of their influence in sickness and health alike.[30]

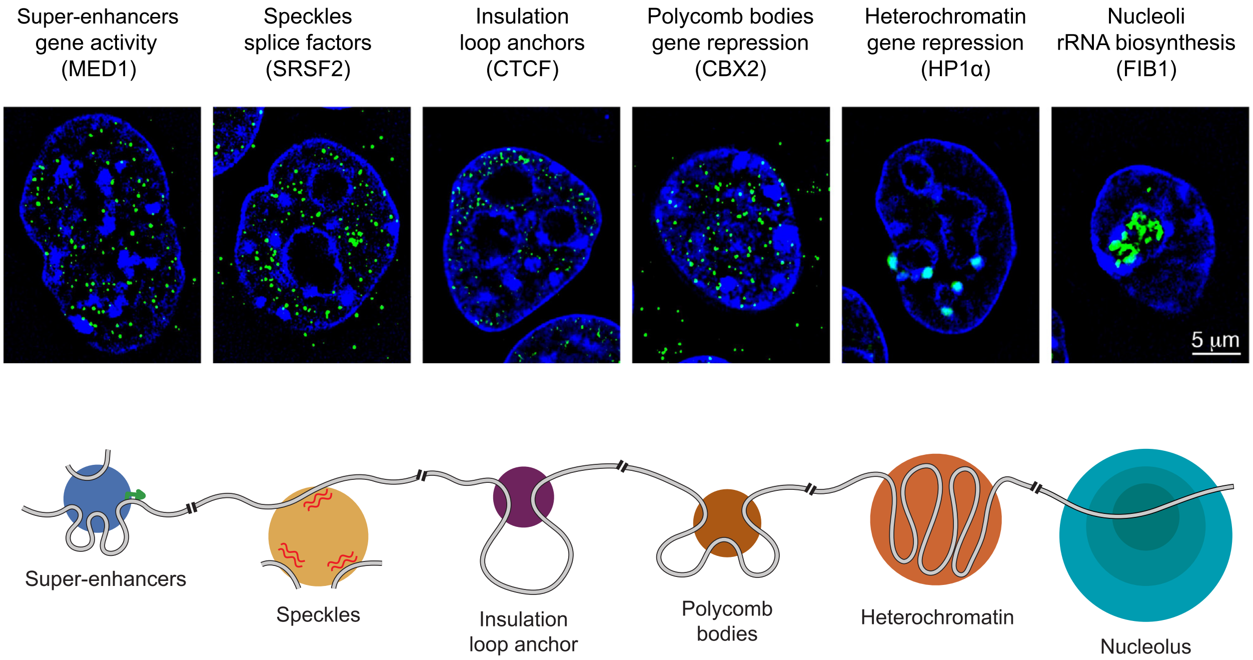


Figure 1.13: Biomolecular condensates of various scales. Reproduced from Sabari *et al.* [29]

Though rapid expansion of the modern genomics toolbox has enabled snapshots of life at unprecedented granularity and depth, technological and analytical barriers siloing discrete data modalities continue to obstruct a complete picture for cellular activity – akin to the parable of the elephant in the dark. Considering the inherent complexity of biological processes based on current knowledge of protein interaction networks and gene regulatory circuits, it's imperative now more than ever to capitalize on the existing and accumulating wealth of multi-omics datasets with integrative approaches to distill holistic insights into specifically the interplay among omics layers (Fig. 1.14).[31] Apart from sharpening our understanding of fundamental principles, an end-to-end comprehension of how changes in a particular aspect propagates through other modalities will additionally supplement novel targets for therapeutic development and thus accelerate translational efforts in precision medicine.

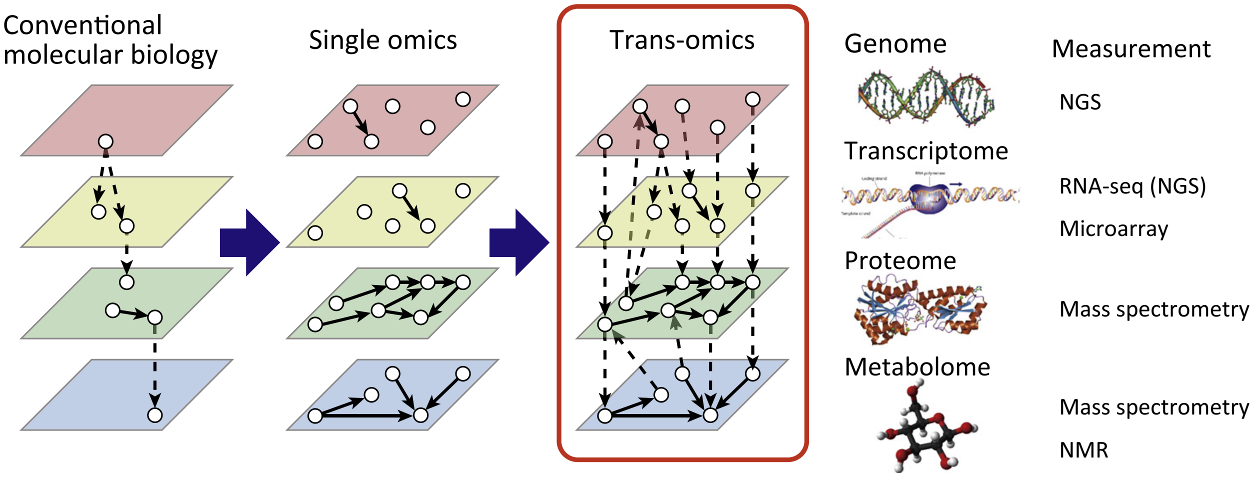


Figure 1.14: Integrative multi-omics. Consideration of information flow within and across omics layers necessary for comprehensive understanding of biological system. Reproduced from Yugi *et al.* [31]

Key multi-omics platforms enabling 3D epigenomics

DNA sequencing technologies, especially next-generation sequencing with their incredible throughput and efficiency, were indispensable in fuelling the explosive growth of early undertakings dissecting genetic variation.[32] Subsequently, ingenious means of encapsulating information from other processes into DNA molecules has been one of the major drivers enabling the detailed characterization and precise quantification of myriad cellular properties and activities. Microscopy and cell imaging has, in parallel, benefitted from advances in optics and fluorophore chemistry and now routinely used to not only provide orthogonal validation of sequencing-based datasets, but also fill in salient gaps such as the spatial and temporal distributions of diverse biomolecules including DNA, RNA, and proteins.[33] Likewise, methodological and computational improvements have also enabled mass spectrometry-based methods to investigate the assortment of proteins present in various samples ("proteome") at increased throughputs without significantly hampering sensitivity and accuracy.[34] Modern studies aiming to molecularly profile specific cellular states thus frequently adopt a multi-pronged approach leveraging many of these technologies in concert, with particularly strong emphasis on three fronts: epigenome, nucleome, and transcriptome (the set of RNA transcripts expressed in a cell).

The epigenome is amenable to a number of methods for dissecting its intricacies from multiple angles (Fig. 1.15).[35] Epigenetic control of chromatin compaction can be evaluated using methods such as assay for transposase-accessible chromatin with sequencing (ATAC-seq) that preferentially enriches for accessible DNA for subsequent sequencing, in the process pinpointing sites of DNA unwound from histones usually driven by transcription factor binding and associated activities. Modifications to DNA itself, specifically methylation of cytosines, can be measured via whole genome bisulfite sequencing (WGBS), where only unmethylated cytosines are converted to thymines while methylated cytosines remain intact, allowing their disambiguation downstream computationally. In contrast, to evaluate the distribution of modifications decorating chromatin across the genome, chromatin immunoprecipitation followed by sequencing (ChIP-seq) draws on the affinity of antibodies for specific targets (e.g., transcription factors or histone modifications) to pull down associated DNA for further analysis. Similarly, western blotting uses antibodies to bind specific targets for quantification. As opposed to focusing on specific targets individually, high-throughput proteomics can also be employed to assay histones and accurately measuring the global abundance of various histone post-translational modifications – therefore complementing sequencing dataset's marker distributional information with precise quantification.[36]

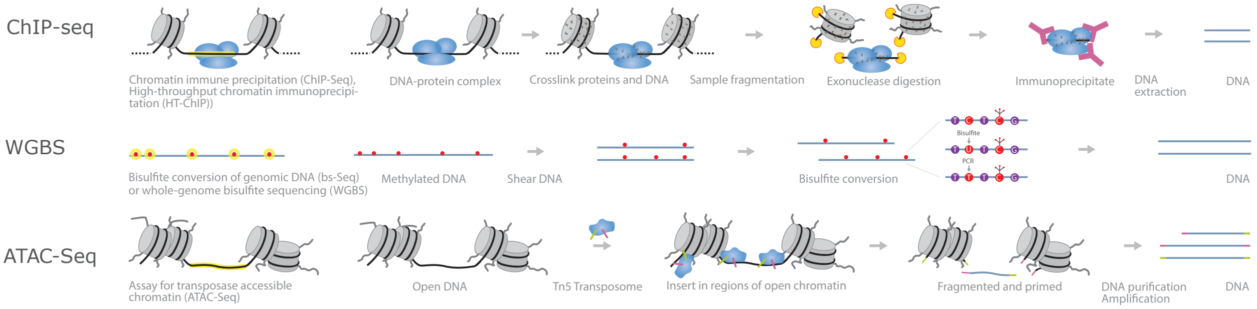

ChIP-seq
Chromatin immune precipitation (ChIP-Seq), High-throughput chromatin immunoprecipitation (HT-ChIP))
DNA-protein complex
Crosslink proteins and DNA
Sample fragmentation
Exonuclease digestion
Immunoprecipitate
DNA extraction
DNA
WGBS
Bisulfite conversion of genomic DNA (bs-Seq) or whole-genome bisulfite sequencing (WGBS)
Methylated DNA
Shear DNA
Bisulfite conversion
T C T C G
Bisulfite
T U T C G
PCR
T T T C G
DNA
ATAC-Seq
Assay for transposase accessible chromatin (ATAC-Seq)
Open DNA
Tn5 Transposome
Insert in regions of open chromatin
Fragmented and primed
DNA purification Amplification
DNA

Figure 1.15: Epigenome mapping. Common techniques to investigate different facets of the epigenome. Reproduced from Illumina [37]

Higher-order organization of the DNA can, too, be assessed at multiple resolutions through several orthogonal means (Fig. 1.16).[38] High-throughput chromosome conformation capture (Hi-C) has emerged as a powerful platform by which the contact frequency between all pairwise combination of DNA segments can be deduced, as spatially nearby pieces of chromatin are subjected to proximity ligation and subsequently profiled using paired-end sequencing to enumerate the number of pairs originating from various genomic regions. Although contact probabilities as measured by Hi-C strongly correlate with physical separations in 3D space, there can exist disparities between "contact frequency" versus "average spatial separation" for pairs of genomic loci across a cell population due to non-random folding of chromatin *in vivo* (e.g., as mediated by loop extrusion),[39] proximity ligation is thus frequently

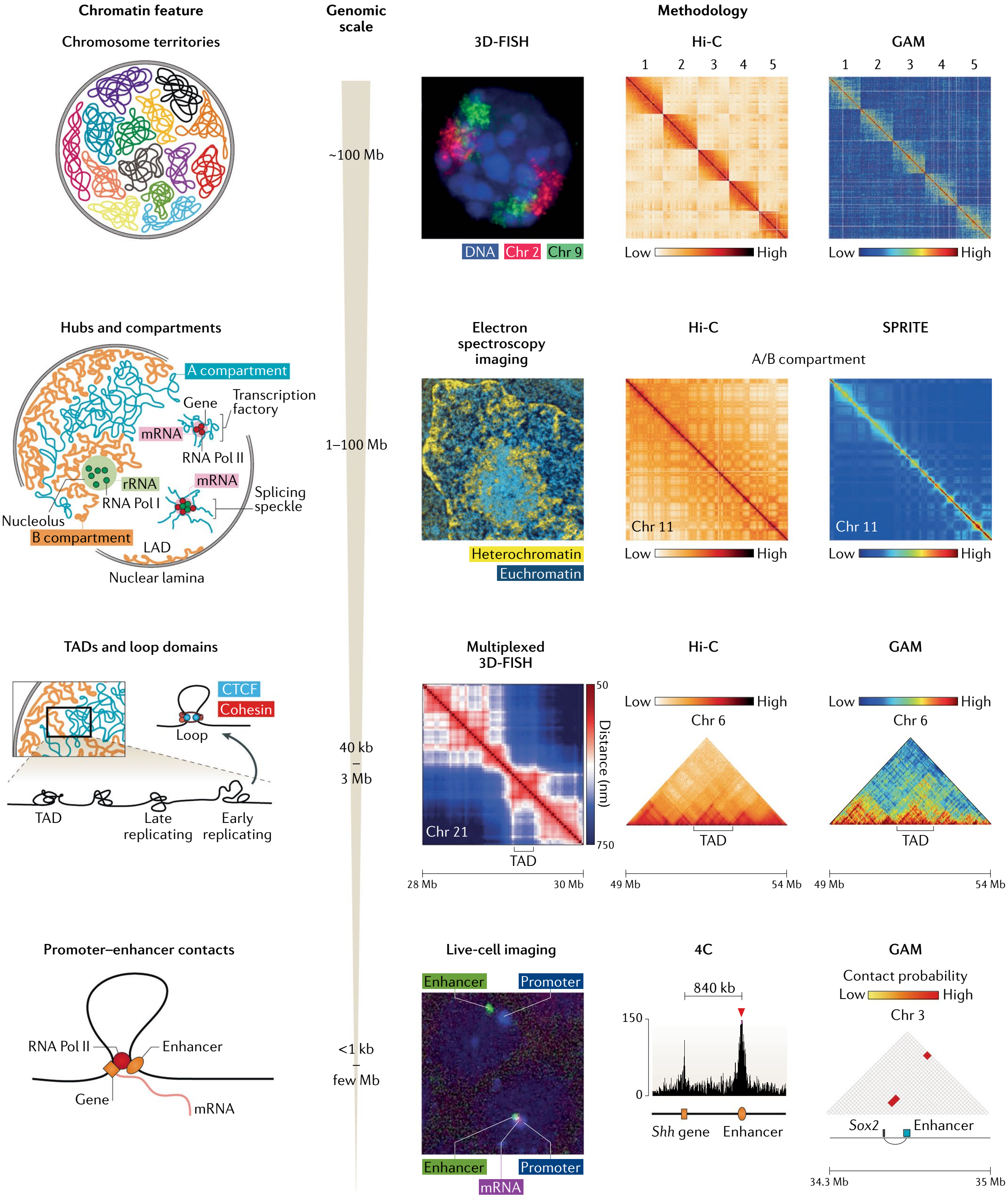
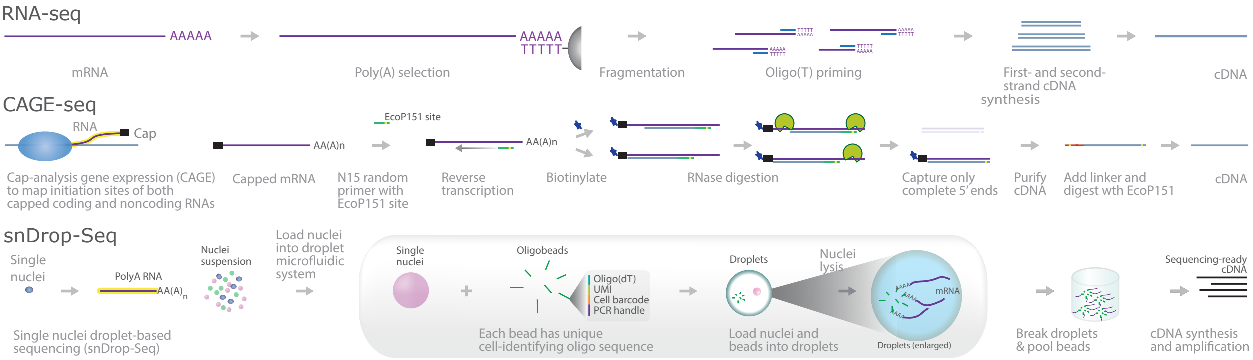


Figure 1.16: Nucleome mapping. Common techniques to investigate different facets of the transcriptome. Reproduced from Kempfer & Pombo [38]

supplemented by fluorescence in situ hybridization (FISH) to label different sequences with targeted probes, after which their true separation can be quantified using microscopy at a single cell resolution. Through recent approaches multiplexing and applying FISH in serial (e.g., in seqFISH+), it's even possible to evaluate the spatial localization for a large number of targets in the same cells, ranging from histone modifications and proteins to RNAs.[40] Instead of DNA sequences, fluorophore-labelled antibodies are also frequently used in immunofluorescence imaging applications to determine the spatial distribution of nuclear structures such as lamina-associated heterochromatin or nucleolus-associated euchromatin.



RNA-seq
AAAAA
AAAAA
TTTTT
mRNA
Poly(A) selection
Fragmentation
Oligo(T) priming
First- and second-strand cDNA synthesis
cDNA
CAGE-seq
RNA
Cap
AA(A)n
EcoP151 site
AA(A)n
Cap-analysis gene expression (CAGE) to map initiation sites of both capped coding and noncoding RNAs
Capped mRNA
N15 random primer with EcoP151 site
Reverse transcription
Biotinylate
RNase digestion
Capture only complete 5′ends
Purify cDNA
Add linker and digest wth EcoP151
cDNA
snDrop-Seq
Single nuclei
PolyA RNA
AA(A)n
Nuclei suspension
Load nuclei into droplet microfluidic system
Single nuclei
Oligobeads
Oligo(dT)
UMI
Cell barcode
PCR handle
Each bead has unique cell-identifying oligo sequence
Droplets
Nuclei lysis
mRNA
Load nuclei and beads into droplets
Droplets (enlarged)
Break droplets & pool beads
Sequencing-ready cDNA
cDNA synthesis and amplification
Single nuclei droplet-based sequencing (snDrop-Seq)

Figure 1.17: Transcriptome mapping. Reproduced from Illumina [41]

Linking chromatin dynamics to downstream outcomes is no less significant than understanding the upstream processes themselves, with one of the most frequently adopted read-outs for phenotypic consequences being gene expression (Fig. 1.17). Although reverse transcription can readily convert RNA into DNA that is ready to be sequenced, the diversity of RNA species introduces additional complexities. Apart from protein-coding messenger RNA (mRNA), there exists a variety of non-coding classes of RNA (ncRNA) molecules such as ribosomal RNA (rRNA) and long noncoding RNA (lncRNA) that can fulfill various biological functions without necessarily being translated.[42] As a result, typical RNA sequencing (RNA-seq) methods usually explore only a subset of total RNA via strategies such as rRNA depletion and poly(A) selection to enrich for mRNA, tagging the 5' cap of mRNAs to map transcription initiation sites, or pulling down RNA polymerase II (RNAPII)-associated RNA to examine nascent transcripts.[43] Many techniques can be further augmented with single cell barcoding to tag molecules from the same cell with unique sequences that can be analytically decoded and enable transcriptomic characterization of individual cells to disentangle the inter-cellular heterogeneity of complex samples.

Despite the richness of information afforded by individual modalities, studies usually stand to gain from a more holistic approach combining signals across omics layers. Besides helping overcome the inherently noisy nature of biological data by means of pinpointing consistent trends across orthogonal methods, multi-omics datasets importantly facilitate the use of data-hungry unsupervised data mining methods to reveal unexpected patterns – critical to elucidating elaborate processes such as development and tumorigenesis.

Charting 3D epigenome dynamics towards totipotency

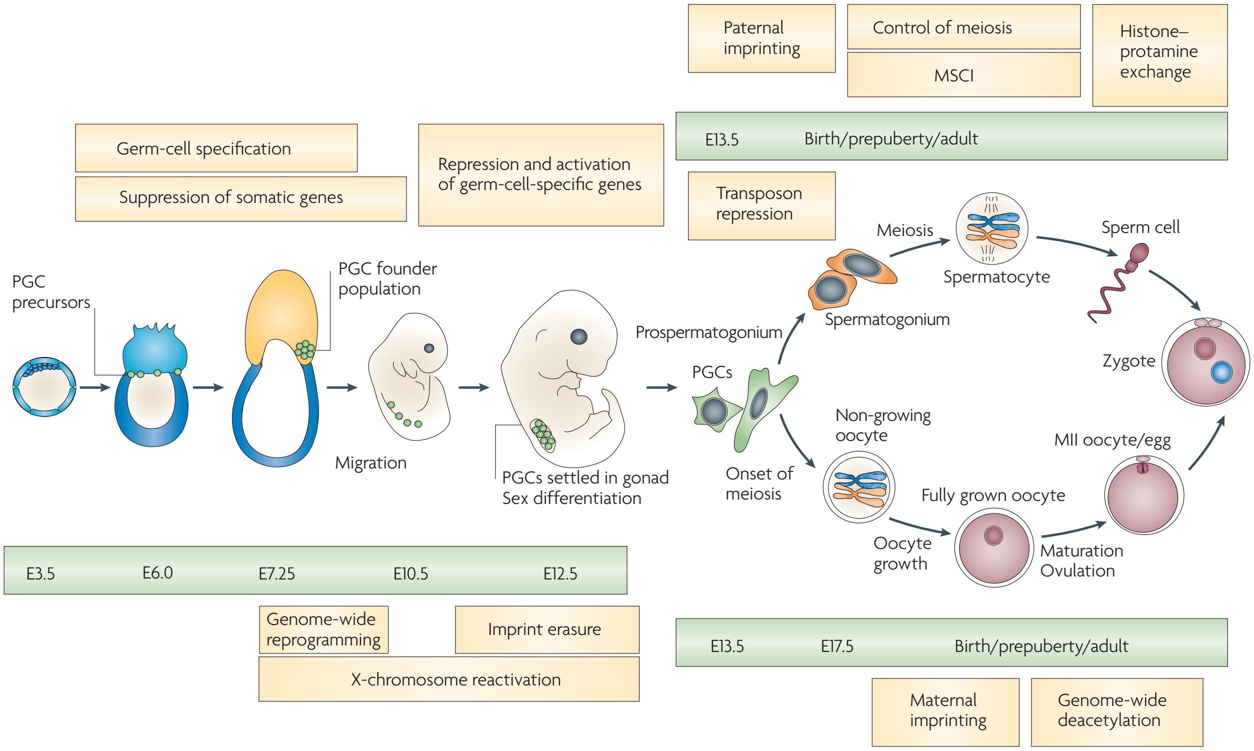


Figure 1.18: Mouse germ-cell development. Reproduced from Sasaki & Matsui [44]

Genetic information is specifically transmitted across generations through the germline, making gametogenesis not only a fascinating system to unravel from a basic biology perspective of understanding the path towards totipotency, but also bear relevance for deciphering candidate mechanisms underlying reproductive disorders. (Fig. 1.18).[44] Apart from the three primary germ layers arising from the epiblast that eventually differentiates into various somatic tissue, primordial germ cell (PGC) specification also occurs in the epiblast and marks one of the earliest conclusive cell fate decisions.[45] A hallmark event that soon ensues is epigenetic reprogramming, a process by which global DNA de-methylation occurs before becoming re-established in a sexual dimorphic manner: largely prenatal for prospermatogonia

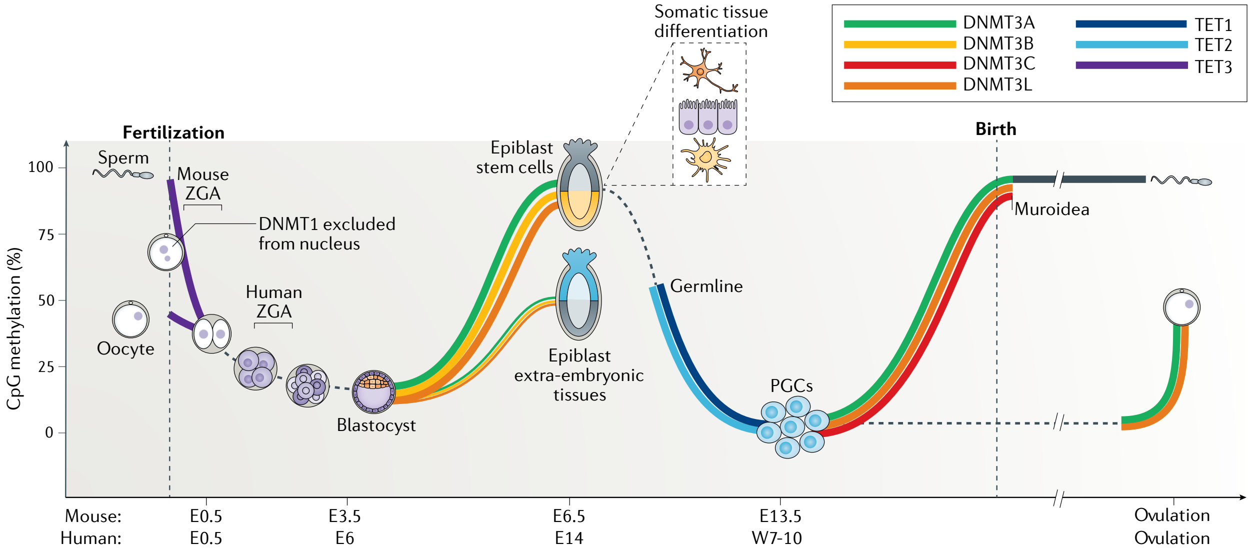


Figure 1.19: Mammalian embryonic epigenome remodelling. Dynamic DNA methylation landscape during early development. Reproduced from Greenberg & Bourc'his [9]

and gradually after birth for growing oocytes (Fig. 1.19).[9] The dramatic chromatin remodeling that takes place during this unique lineage has thus garnered substantial interest from developmental and chromatin biologists alike. However, the meagre population of germ cells *in vivo* has been a roadblock for conventional assays, as it imposes a difficult burden on obtaining sufficient input material. Towards circumventing this obstacle, *in vitro* reconstitution systems have been developed that can faithfully produce functional oocytes and spermatozoa entirely from embryonic stem cells in a dish, empowering a thorough examination of chromatin dynamics during the course of germ cell development.

Meiotic development up until fertilization and early zygotic stages have specifically been intensively investigated to begin unravelling the foundation of totipotency (Fig. 1.20).[46] For instance, oocytes were noted for their 3D genome's lack of compartmentalization and insulation, with gradual emergence of domain and loop structures only taking place once zygotic genome activation begins in early embryos.[47] Additionally, the polycomb system is also known to exhibit unique dynamics including the allele-specific distributions of H3K27me3 and H2AK119ub immediately before and following fertilization,[48] as well as the transient strengthening of interactions between polycomb domains in growing oocytes.[49] Yet mitotic germline development has remained less understood and needs to be explored in greater detail.

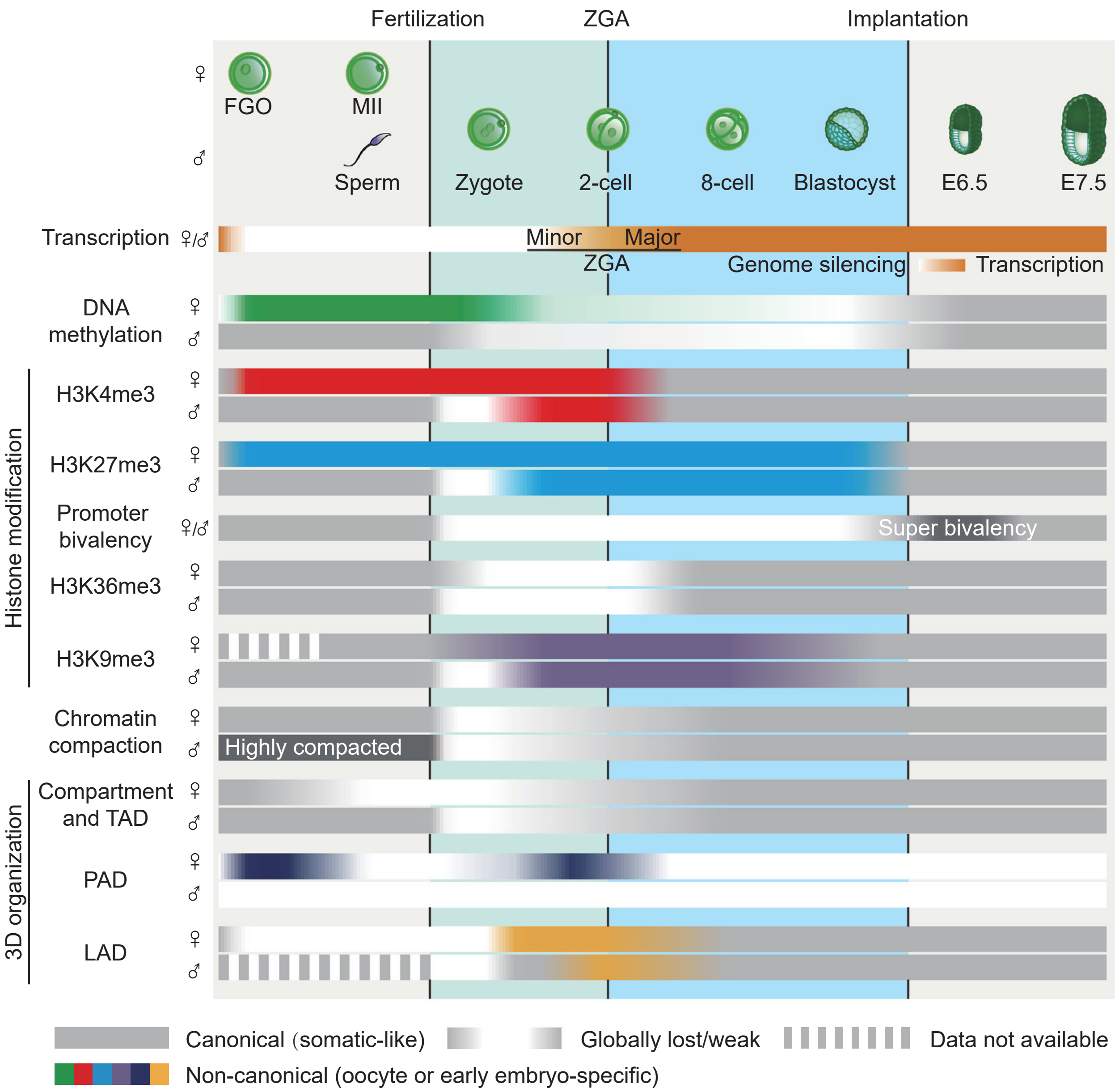


Figure 1.20: 3D epigenome dynamics during early embryogenesis. Reproduced from Xia & Xie [46]

Effects of polycomb re-organization on the 3D epigenome and beyond

While the epigenome undergoes extensive remodelling extensively during development, similarly dramatic transformations can likewise take place during pathogenesis, especially for cancer. Accompanying the growth in sequencing technologies for basic research, clinical applications of genetic and epigenetic profiling to cancers have revealed the existence of molecularly distinct subtypes within the more traditional tumor categories broadly based on anatomical location. These classifications are

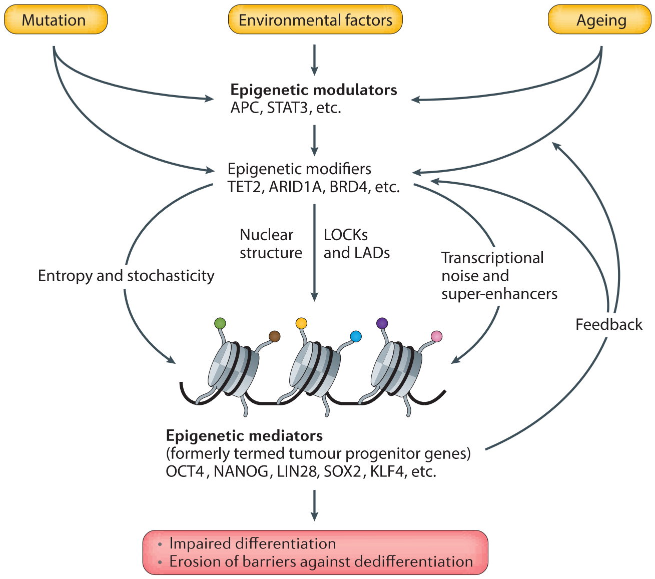

Mutation
Environmental factors
Ageing
Epigenetic modulators
APC, STAT3, etc.
Epigenetic modifiers
TET2, ARID1A, BRD4, etc.
Nuclear structure
LOCKs and LADs
Entropy and stochasticity
Transcriptional noise and super-enhancers
Feedback
Epigenetic mediators
(formerly termed tumour progenitor genes)
OCT4, NANOG, LIN28, SOX2, KLF4, etc.
• Impaired differentiation
• Erosion of barriers against dedifferentiation

Figure 1.21: Cancer-linked epigenetic genes. Reproduced from Feinberg *et al.* [50]

particularly important due to high variability in treatment responses, posing an urgent need for finer patient stratification and more tailored selection of therapeutic strategies – the aim of precision medicine. Although the irony of personalized medicine lies in the necessity of massive datasets to detect exquisite differentiators separating individuals, global efforts fortunately heeded these calls with the requisite scale.[51] Consequently, it became apparent that driver mutations in chromatin modifiers or histone themselves constitute a formidable subgroup of their own, often leaving notable epigenomic as well as transcriptomic footprints in addition to characteristic genetic alterations.[52] Given the established role of epigenetic regulation in cellular differentiation, the link between mutations in chromatin regulators with impaired differentiation or aberrant de-differentiation in the course of tumorigenesis came as less than surprising (Fig. 1.21).[50] But beyond mutations in epigenetic modifiers, the canvas itself – chromatin – has also been found to possess characteristic mutations, with the most prominent ones entirely localized on the tail of histone H3 (Fig. 1.22).[53] Such oncogenic histone mutants ("oncohistones"), including histone H3 lysine (27)-to-methionine (H3K27M) and H3K36M, are believed to dominantly exert an inhibitory influence on the corresponding methyltransferases (e.g., PRC2, NSD1/2) in *trans*, causing the genome-wide depletion of methylation on the mutated residue, even for wild-type histones.[52]

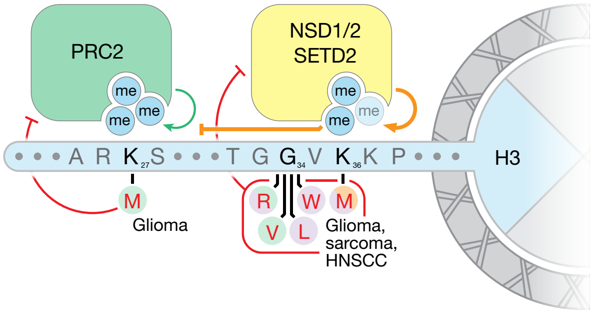


Figure 1.22: Oncohistone H3 mutations. Reproduced from Nacev *et al.* [53]

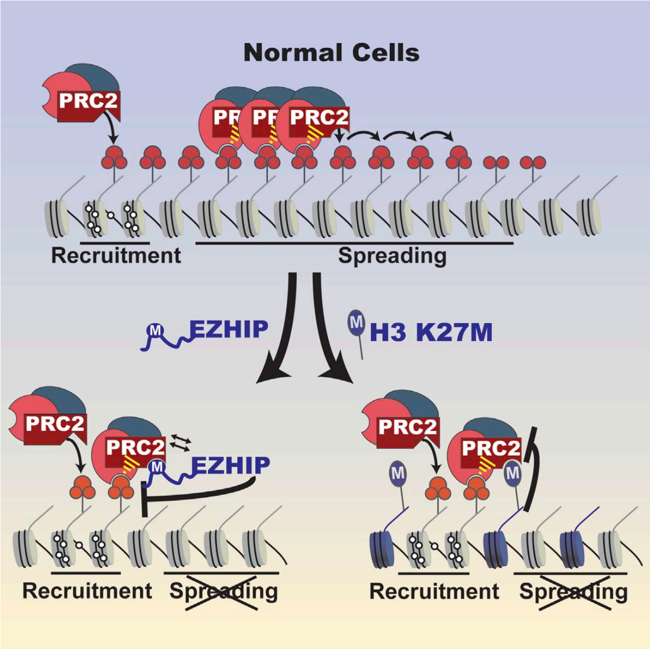


Figure 1.23: PRC2 inhibition by H3K27M/EZHIP. H3K27M oncohistones and EZHIP specifically impede spreading of PRC2. Reproduced from Jain *et al.* [54]

Whether it be through cell-intrinsic or -extrinsic mechanisms, tumorigenic mutations, especially those in chromatin modifiers, can show high levels of specificity to particular cancer types, and H3K27M is no exception. As one of the most common pediatric malignancies, childhood brain tumors have remained deeply troubling in view of potential long-term complications associated with surgically operating on the brain;[55] and among the diverse patient population, a epigenetically distinct subset of pediatric high-grade gliomas (pHGGs) was identified to specifically possess H3K27M as a driver mutation.[56] Following the initial discovery of H3K27M-associated pHGGs, a previously uncharacterized protein that is over-expressed in posterior fossa type A ependymomas (PFA-EPNs) was revealed to contain a H3K27M-like peptide subsequence, and was therefore named Enhancer of Zeste Homologs Inhibitory Protein (EZHIP). This name stemmed from the discovery that while H3K27M/EZHIP by and large does not affect the recruitment of PRC2 to chromatin, they significantly impair the spreading PRC2 through binding to its catalytic subunit EZH2; the result of H3K27M mutation and EZHIP over-expression is thus the contraction of repressive broad H3K27 methylation domains and consequent up-regulation of opposing euchromatic modifications such as H3K27 acetylation (Fig. 1.23).[54] While cancer stem cells are thought to be involved across both adult and pediatric cancers, the path through which this state is achieved diverges: adult cancers are generally considered to revert to a stem-like state via de-differentiation, whereas pediatric cancers fail to progress past a certain development stage. In line with this school of thought, it was found that H3K27M glioma stem cells are unresponsive to differentiation stimuli *in vitro*, unlike isogenic comparisons where the mutation was removed via CRISPR-Cas9 that readily proceeds to become mature astrocytes/oligodendrocytes.[57] Indeed, single-cell analysis of both H3K27M pHGGs and EZHIP-overexpressing posterior PFA-EPNs uncovered substantial undifferentiated progenitor cells – implicating differentiation blockade as the most likely tumorigenic mechanism.[58]

H3K27me3 can also undergo global re-distribution without any direct defects in the Polycomb system itself, as it's been reported that the higher methylation states of H3K27 and H3K36 residues frequently engage in a tug-of-war across a variety of biological contexts ranging from healthy stem cells to malignant cancer cell lines, with their domain edges frequently acting as reciprocal boundaries (Fig. 1.24).[59] As a result, mutations affecting different chromatin modifiers have frequently been linked back to the dysregulation of precisely this balance between H3K36 and H3K27 methylation.[60]. With the genome serving as a consistent backdrop, the expansion of one domain thus compels the encroachment upon another in a zero-sum game,[61] with the competition in force throughout development as well as pathogenesis to orchestrate the coordinated activation and repression of various transcriptional programs.

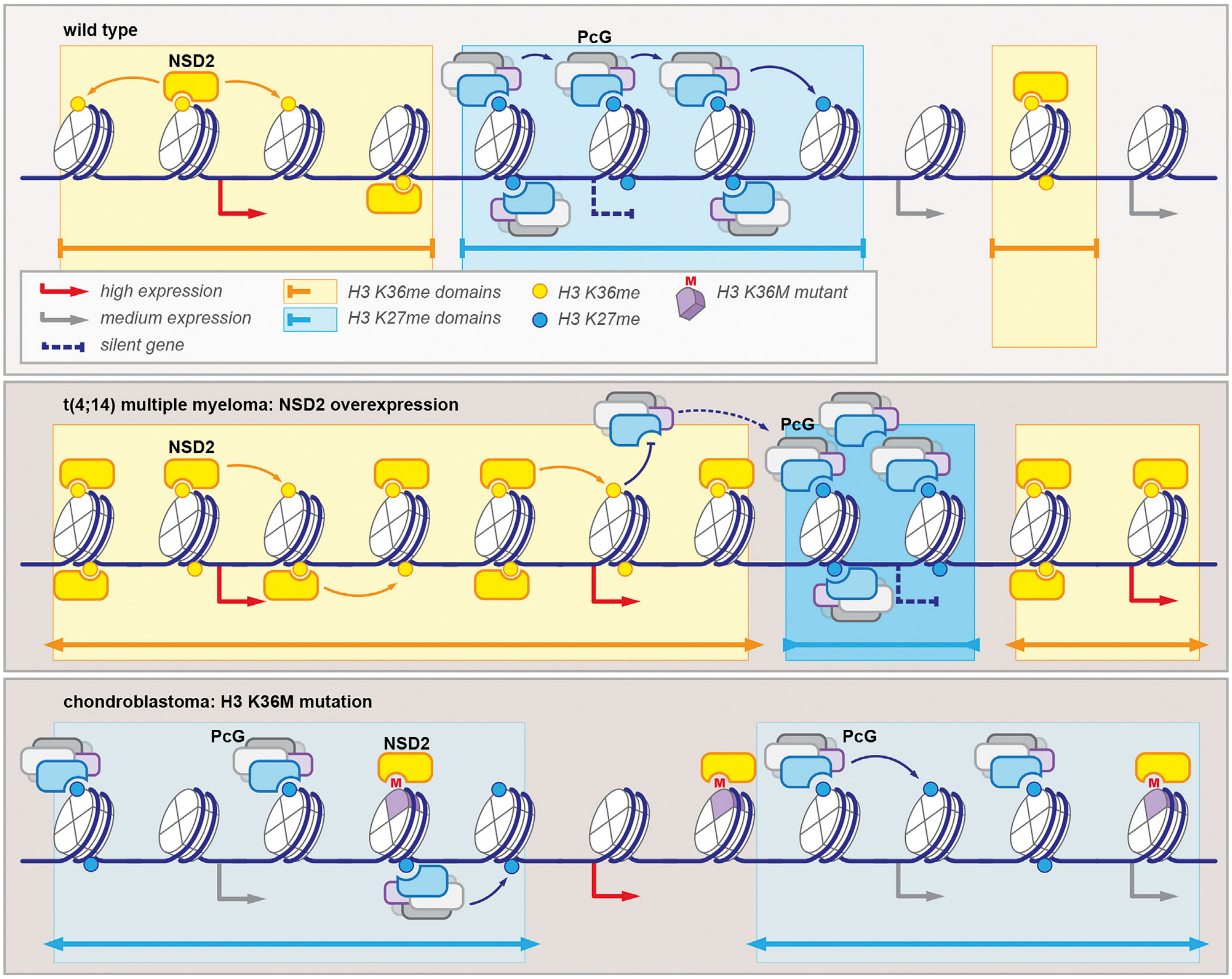


Figure 1.24: Reciprocity of H3K27 vs H3K36 methylation. Reproduced from Soshnev *et al.* [59]

H3K36 methylation, in particular, have been deemed a major force in shaping the epigenome, with mutations to H3K36 methyltransferases such as NSD1 and NSD2 or in histone H3 itself (e.g., H3 lysine(36)-to-methionine, or H3K36M, mutations) leading to not only the dramatic depletion of H3K36 methylation, but also a simultaneous gain of H3K27 methylation and loss of H3K27 acetylation as well as DNA methylation (or vice versa in the case of elevated H3K36me).[62–64] It's therefore evident that a careful maintenance of the balance between H3K27 and H3K36 methylation is necessary for healthy cells, whereas tipping the scales can wreck havoc.

Pinpointing the consequences of 3D epigenomic alterations

In view of the 3D epigenome's emergent importance in mammalian health and disease, we resolved to apply multi-omics profiling to normal germ cell differentiation in mice as well as developmentally stalled brain tumors in humans, demonstrating how integrative analytical approaches can step up to the challenge and fulfill unmet gaps in the molecular understanding of chromatin-centric genome regulation.

To tackle the deficit in our understanding of germline mitotic development, we used our *in vitro* differentiation system to closely characterize vital events such as naïve-to-primed transition, primordial germ cell specification, epigenetic reprogramming, and sexual dimorphic germ cell maturation. We hypothesized that a comprehensive 3D epigenomic atlas of male germ cell development will furnish insights into both the physiological idiosyncrasies of the germline and potential points of failure where defects may lead to dire consequences. We thus aimed to build such a compendium through the collection of a time-series multi-omics dataset spanning the nucleome, epigenome, and transcriptome with a range of complementary approaches. Subsequently, we sought to systematically apply genome-wide analytical strategies to quantitatively summarize temporal 3D epigenome variations at multiple scales in an unsupervised fashion, from chromosomal territories and compartmentalization down to domain structures and regulatory elements as well as chromatin loops bridging such regions. Finally, we methodically documented the differences between fully functional and aberrant spermatogonia with impaired spermatogenic potential, paving the path towards understanding the chromatin determinants of reproductive capacity.

On the other hand, based on existing results implicating the dysregulation of H3K27 and H3K36 methylation as driver pathogenic events, we hypothesized that comprehensively profiling the 3D epigenome should provide additional insights into the mechanistic link between primary chromatin alterations targetting H3K27 and H3K36 methylation with downstream phenotypic consequences such as faulty developmental progression. To this end, we set out to assemble a collection of 3D epigenomic datasets from different cell types and biological systems where the balance of H3K27 vs H3K36 is tipped to varying degrees in either direction, including both published resources and newly generated in-house data. We next pursued the specific intermediaries ultimately responsible for alterations in chromatin dynamics through the integration of 1D epigenetic and 3D chromatin conformation datasets. Building on top of the chromatin-based findings, we resolved to identify whether coordinated local and higher-order chromatin changes can be eventually traced to differential gene expression. Following our initial discoveries in cell line models with abundant data, we then validated whether the results are consistent with observations from patient-derived xenografts and primary tumors. Taken the exploratory results altogether, we finally homed in to the most promising co-factors linking upstream chromatin dynamics with downstream transcriptomic alterations for pharmacological perturbation, setting the stage for promising targeted therapeutic strategies.

Most cells contain the same set of genes and yet they are extremely diverse in appearance and functions. Germ cells, stem cells and early embryos all exhibit pluripotency, but each cell type also displays certain unique properties. Mechanisms that regulate this exceptional genomic plasticity and the state of totipotency are being unravelled, and will enhance our ability to manipulate stem cells for therapeutic purposes.

M. Azim Surani

Chapter 2

Nucleome programming for the foundation of totipotency in mammalian germline development

Germ cells are known to undergo epigenetic reprogramming – a highly unique process involving global DNA demethylation as well as other dramatic epigenomic alterations; it is believed that this reset is critical in facilitating the acquisition of totipotency in the next generation. Despite previous microscopic observations of notable nuclear architecture changes accompanying this drastic remodelling event, the mechanisms governing such multi-scale transformations and the functional implications of their interplay with other regulatory modalities remain largely unknown – as the limited number of gonadal germ cells *in vivo* has hindered the application of conventional genome-wide assays. Capitalizing on our murine *in vitro* differentiation system, we here investigated the determinants of nuclear totipotency underlying male germline development. In particular, we assembled a time-series multi-omics compendium of 6 cell types using Hi-C, histone mass spectrometry, ChIP-seq of histone modifications/transcription factors/architectural proteins, ATAC-seq, RNA-seq, NET-CAGE, and WGBS.

Nucleome programming for the foundation of totipotency in mammalian germline development

Masahiro Nagano[1,2,18], Bo Hu[2,3,18], Shihori Yokobayashi[1,2,4], Akitoshi Yamamura[1,2], Fumiya Umemura[1,2], Mariel Coradin[5,6,7], Hiroshi Ohta[1,2], Yukihiro Yabuta[1,2], Yukiko Ishikura[1,2], Ikuhiro Okamoto[1,2], Hiroki Ikeda[4,8,], Naofumi Kawahira[9,10], Yoshiaki Nosaka[1,2], Sakura Shimizu[1,2], Yoji Kojima[1,2,4], Ken Mizuta[1,2], Tomoko Kasahara[1,11], Yusuke Imoto[1], Killian Meehan[1], Roman Stocsits[12], Gordana Wutz[12], Yasuaki Hiraoka[1], Yasuhiro Murakawa[1,11], Takuya Yamamoto[1,4,13], Kikue Tachibana[14,15], Jan-Michel Peters[12], Leonid A Mirny[16], Benjamin A. Garcia[5,6,17], Jacek Majewski[3], Mitinori Saitou[1,2,4]

[1]Institute for the Advanced Study of Human Biology (WPI-ASHBi), [2]Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan.

[3]Department of Human Genetics, McGill University, Montreal, Quebec, Canada.

[4]Center for iPS Cell Research and Application (CiRA), Kyoto University, 53 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan.

[5]Department of Biochemistry and Biophysics, [6]Penn Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[7]Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder, Boulder, CO 80309, USA.

[8]Department of Embryology, Nara Medical University, Nara, Japan.

[9]Department of Molecular Cell Developmental Biology, School of Life Science, University of California, Los Angeles, CA, USA.

[10]Laboratory for Developmental Morphogeometry, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan.

[11]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan.

[12]Research Institute of Molecular Pathology, Vienna BioCenter, Vienna, Austria.

[13]Medical-risk Avoidance based on iPS Cells Team, RIKEN Center for Advanced Intelligence Project, Kyoto, Japan.

[14]Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna BioCenter, Vienna, Austria.

[15]Department of Totipotency, Max Planck Institute of Biochemistry, Martinsried, Germany.

[16]Institute for Medical Engineering and Science, and Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA.

[17]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110-1010, USA.

[18]These authors contributed equally.

Running title: Nucleome Programming for Germ Cells

**Correspondence:* Mitinori Saitou, M.D., Ph.D.

E-mail: saitou@anat2.med.kyoto-u.ac.jp; Tel: +81-75-753-4335; Fax: +81-75-751-7286

ABSTRACT

Germ cells are unique in engendering totipotency, yet the mechanisms underlying this capacity remain elusive. Here, we perform comprehensive and in-depth nucleome analysis of mouse germ-cell development *in vitro*, encompassing pluripotent precursors, primordial germ cells (PGCs) before and after epigenetic reprogramming, and spermatogonia/spermatogonial stem cells (SSCs). Although epigenetic reprogramming, including genome-wide DNA de-methylation, creates broadly open chromatin with abundant enhancer-like signatures, the augmented chromatin insulation safeguards transcriptional fidelity. These insulatory constraints are then erased en masse for spermatogonial development. Notably, despite distinguishing epigenetic programming, including global DNA re-methylation, the PGCs-to-spermatogonia/SSCs development entails further euchromatization. This accompanies substantial erasure of lamina-associated domains (LADs), generating spermatogonia/SSCs with minimal peripheral attachment of chromatin except for pericentromeres—an architecture conserved in primates. Accordingly, faulty nucleome maturation, including persistent insulation and improper euchromatization, leads to impaired spermatogenic potential. Given that PGCs after epigenetic reprogramming serve as oogenic progenitors as well, our findings elucidate a principle for the nucleome programming that creates gametogenic progenitors in both sexes, defining a basis for nuclear totipotency.

INTRODUCTION

Germ cells are the origin of totipotency, which in turn is the foundation for individual development. Mechanisms underlying totipotency have been a focus of intensive investigations, ranging from studies involving somatic-cell nuclear transfer [1] to recent efforts exploring the three-dimensional (3D) chromatin organization in zygotes and early embryos [2]. The latter works involving chromatin conformation capture have revealed a relaxed chromatin configuration in zygotes in part resulting from unique meiotic intermediates, and the progressive maturation of this configuration in early embryos [3-10]. On the other hand, the manner by which germ cells elaborate the higher-order chromatin organization during their mitotic development, and the founding states for gametogenesis and totipotency, remain poorly understood. In-depth understanding of genome functions requires investigations of the 3D genome organization complemented by thorough epigenome and transcriptome profiling, an approach known as "nucleome" profiling [11]. While nucleome profiling has been performed in a few somatic lineages [12-14], studies applying this approach to germ-cell development are lacking.

In mammals, germ cells arise as primordial germ cells (PGCs) during early embryonic development [15]. PGCs undergo migration and colonize the embryonic gonads, where they differentiate either into spermatogonia/spermatogonial stem cells (SSCs), the source for spermatogenesis, or oocytes with an immediate entry into the first prophase of meiosis [16-18]. A key event that characterizes PGCs is epigenetic reprogramming, including de-methylation of genome-wide DNA to the point that it contains almost no DNA methylation, as well as histone-modification remodeling, which creates a facultative "naïve" epigenome [19,20]. In males, epigenetic reprogramming is followed by the provision of a distinct spermatogenic epigenome, including global DNA re-methylation, for spermatogonia/SSC development, whereas in females, the naïve epigenome serves as a direct precursor for the oogenic meiotic entry [19]. Thus, male germ-cell development requires at least one additional epigenetic programming step to create spermatogenic progenitors. Here, to explore the principles that create a basis for gametogenic potential, we performed nucleome profiling of an *in vitro* system that faithfully reconstitutes mouse germ-cell development from pluripotent precursors to PGCs before and after epigenetic reprogramming and then to spermatogonia/SSCs [21-24]. We show that the *in vitro* system recapitulates not only gene-expression and epigenetic properties, but also 3D genome-organization dynamics during germ-cell development *in vivo*, lending credence to our analyses using scalable materials to provide a more complete picture of nucleome dynamics with high resolution during germ-cell development. In addition, to delineate the functional significance of appropriate nucleome programming, we analyzed the nucleome of an *in vitro* counterpart of spermatogonia/SSCs with an impaired spermatogenic potential [25].

RESULTS

Mouse germ-cell development *in vitro*

We analyzed the following male cell types (Fig 2.1A): mouse embryonic stem cells (mESCs) derived from blastocysts [26], epiblast-like cells (EpiLCs) [21], mouse PGC-like cells at day 2 of induction (d2 mPGCLCs) [21], d4 mPGCLCs expanded *in vitro* for 7 days for epigenetic reprogramming (d4c7 mPGCLCs) [22,23], and germline stem cells (GSCs) derived from neonatal spermatogonia [24]. These cells show gene-expression, epigenetic, and functional properties equivalent to those of their *in vivo* counterparts, i.e., mESCs to epiblast at embryonic day (E) 4.5 with naïve pluripotency [27,28], EpiLCs to epiblast at ~E6.0 with formative pluripotency [21], d2 mPGCLCs to mPGCs during their specification at ~E7.0 and before epigenetic reprogramming [21,29], d4c7 mPGCLCs to PGCs at E11.5 after epigenetic reprogramming [22,23], and GSCs to spermatogonia/SSCs [25]. Note that PGCs before E11.5 do not show overt sexual differences in gene-expression and epigenetic properties, except X-chromosome reactivation in females [22,30]. Accordingly, male PGCs bear a capacity to form functional oocytes [31], and male mPGCLCs take on the oogenic fate and enter into the meiotic prophase in response to appropriate signals at an efficiency comparable to that of female mPGCLCs [32,33]. Thus, while our present analysis focuses on male germ-cell development, male d4c7 mPGCLCs can be considered to bear an oogenic potential as well. In addition, to evaluate the functional relevance of proper nucleome programming, we analyzed GSC-like cells (GSCLCs) that were derived from d4 mPGCLCs *in vitro* and had an impaired spermatogenic potential [25] (see the "**Nucleome programming engenders gametogenic potential**" section).

Higher-order genome organization: maturation towards a highly euchromatized state

We first examined the nuclear morphology of the five cell types (mESCs, EpiLCs, d2 mPGCLCs, d4c7 mPGCLCs, and GSCs) stained with DAPI (4',6-diamidino-2-phenylindole) using high-resolution confocal microscopy. Counterintuitive to GSCs' acquisition of a distinct spermatogenic epigenome, including global DNA re-methylation, on the epigenome of naïve PGCs, the areas of high DAPI density (peri-centromeric heterochromatin) [34], the variances of DAPI density (chromatin condensation heterogeneity), and the distances of the DAPI-dense areas from the nuclear periphery (chromosome radial positioning), all exhibited a monotonically decreasing transformation towards GSCs (Fig 2.1B-C). This indicates that chromatin de-condensation (i.e., euchromatization), as well as peripheral tethering of centromeres, proceeds progressively beyond the canonical epigenetic reprogramming period. Notably, formative EpiLCs showed more discrete chromatin condensation than naïve mESCs, while mESCs and d4c7 mPGCLCs (latent pluripotency) [35] exhibited significant differences in chromosome radial positioning (Fig 2.1B-C). Fluorescence in situ hybridization (FISH) confirmed that, in line with chromatin de-condensation, GSCs bore larger chromosome volumes than mESCs and EpiLCs (Fig 2.1D, A.1A).

We next analyzed the five cell types by *in situ* Hi-C (~5 kb resolution) with reproducible biological replicates (Fig A.1B, Table A.1). Consistent with the morphological observations, 3D genome organization was transformed in an unidirectional manner during germ-cell development: the chromosomal contact profile shifted progressively from the conventional proximal contact-enriched state to a more uniform profile with heightened distal interactions (Fig 2.1E, A.1C, B.1A), and the compartment score distributions and euchromatin-to-heterochromatin balance exhibited a monotonical increase (Fig 2.1G, A.1D). Notably, while the vast majority (~33.3% genome-wide) of the A compartment in mESCs remained an A compartment, more than one third (~38.9% genome-wide) of the B compartment in mESCs progressively turned into A, with the largest B-compartment fraction (~7.5% genome-wide) turning into A upon the d4c7 mPGCLC-to-GSC transition. In stark contrast, the compartment scores exhibited a gradual decrease during somatic development, including neuronal, B-cell, and cardiomyocyte differentiation (Fig 2.1G, B.1B) [12-14]. The brief decrease in the compartment score upon EpiLCs-to-d2 mPGCLCs differentiation (Fig 2.1G) is consistent with the transient activation of a somatic program during mPGCLC specification [29]. Accordingly, principal component analysis (PCA) of the compartment scores segregated the germline from somatic development (Fig B.1C). Along with the expansion of the A compartment (Fig 2.1G, Fig A.1D), euchromatic A-A interactions became less intense, while the reduced B compartment exhibited stronger B-B interactions both within (cis) and between (trans) chromosomes, implying the formation of repressive condensates (Fig 2.1F).

On a smaller scale, topologically associating domain (TAD) boundaries exhibited a substantial overlap during germ-cell development, with the degree of their conservation being similar to that of somatic lineages (Fig A.1E-F, B.1D). However, inter-TAD interactions involving the simultaneous aggregations of multiple non-neighboring TADs, referred to as "TAD-cliques" [36], became dramatically less prevalent in the A compartments, while they were over-represented in the B compartments in both d4c7 mPGCLCs and GSCs, which was in stark contrast to their opposite/relatively stable behaviors in somatic lineages (Fig 2.1H-I, B.1E). Through polymer simulations, we generated representative 3D structures of whole chromosomes [37], which similarly demonstrated the progressive expansion of chromosome volume during germ-cell development (Fig A.1G, B.1F, Movie A.1).

To examine whether the five cell types recapitulate their *in vivo* counterparts at the 3D genome organization level, we retrieved published Hi-C data of the inner cell mass at ~E4.0, epiblast at E6.5, PGCs at E11.5, and spermatogonia in adults, which were generated from small numbers of samples [4,38,39]. Remarkably, not only at the transcriptomic and epigenomic level that we reported previously [21,22,25], the *in vitro* cell types exhibited a strong concordance with their *in vivo* counterparts at the 3D genome organization level (Fig A.1C&E) (despite the elevated noise of contact matrices from *in vivo* samples), with unsupervised hierarchical clustering (UHC) and PCA using compartment scores consistently placing corresponding cell types next to one another (Fig A.1H-I). Thus, the *in vitro* system faithfully captures the nucleome dynamics of *in vivo* germ-cell development, further empowering our strategy for using scalable *in vitro* materials to delineate a more complete picture of nucleome dynamics during germ-cell development. We conclude that, beyond the canonical epigenetic reprogramming period, higher-order genome organization undergoes a continuous maturation and culminates in a largely euchromatic genome and peripherally positioned centromeres in spermatogonia/SSCs (GSCs). Thus, global DNA methylation and euchromatization are separable events. Moreover, our findings revealed that, despite their profound epigenomic differences, PGCs (d4c7 mPGCLCs) with both oogenic and spermatogenic potential and spermatogonia/SSCs (GSCs) show relatively similar higher-order genome organization.

Epigenome profiling: epigenetic reprogramming for highly open chromatin with enhanced insulation

To explore the mechanism underlying the higher-order genome organization unique to the germ line, we conducted comprehensive epigenome profiling of the five cell types. We performed mass spectrometry (MS) of histones; chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) of 13 different targets, including 9 histone modifications; assay for transposase-accessible chromatin with deep sequencing (ATAC-seq) for open chromatin; and native elongating transcript–cap analysis of gene expression (NET-CAGE) for transcribed *cis* regulatory elements (Dataset EV1). For some assays, we analyzed d4 mPGCLCs, which are in the middle of epigenetic reprogramming, as an intermediate between d2 and d4c7 mPGCLCs and mouse embryonic fibroblasts (MEFs) as a somatic control.

MS revealed dynamic changes in histone-modification levels with high reproducibility (Fig 2.2A, Table A.2). Consistent with previous observations [22,29], histone H3 lysine 9 di-methylation (H3K9me2) was substantially reduced and H3K27 tri-methylation (H3K27me3) was strongly up-regulated in d4c7 mPGCLCs (Fig 2.2A, A.2A-B). With respect to active modifications, H3K27 acetylation (H3K27ac: active cis-regulatory elements) and H3K18ac were the most abundant in EpiLCs, whereas H3K4 mono-methylation (H3K4me1: poised enhancers), H3K14ac, and H3K23ac were the most adundant in d4c7 mPGCLCs, and, interestingly, H3K4me3 (promoters) was the least prevalent in d4c7 mPGCLCs (Fig 2.2A). UHC based on H3-modification abundance segregated each cell type with their unique sets of associated H3 modifications (Fig 2.2B), and PCA demonstrated characteristic transitions of epigenetic properties, with the transition from d2 to d4c7 mPGCLCs representing the epigenetic reprogramming to a latent pluripotency and the transition from d4c7 mPGCLCs to GSCs signifying the acquisition of a spermatogenetic epigenome (Fig 2.2C). We proceeded to normalize all histone modification ChIP-seq signals with MS-based scaling factors for subsequent analyses (Fig A.2C-D) [40].

We first scrutinized the open-chromatin landscape. Consistent with d4c7 mPGCLCs being globally DNA demethylated (~5%) (Fig 2.1A) [22,23], they exhibited a pervasively open chromatin with coincident up-regulation of H3K4me1, bearing large open domains in a genome-wide manner (Fig 2.2D-E). Indeed, among a diverse panel of mouse fetal tissues [41], d4c7 mPGCLCs showed the highest degree of openness (Fig A.2E). Consistent with the analysis of the abundance of H3 modifications (Fig 2.2C), PCA with the most variable open sites (Fig A.2F) and UHC revealed that d4c7 mPGCLCs share open sites for pluripotency with mESCs and those for germ-cell identity with GSCs: the former (clusters 1, 2, 4) being enriched in transcription-factor (TF)-binding sites for POU5F1, NANOG, SOX2, ZIC2/3, and KLF3/12, and the latter (clusters 3, 7) in those for DMRTs (Fig A.2G, Table A.3).

Despite their genome-wide DNA demethylation, PGCs and d4c7 mPGCLCs do not exhibit transcriptional hyperactivity or promiscuousness [22,23,42]. To explore higher-order regulatory mechanisms, we identified enhancer-promoter (E-P) pairs using the activity-by-contact model by integrating ATAC-seq, H3K27ac, and Hi-C data (Fig A.3A) [43]. Notably, d4c7 mPGCLCs showed a reduced number and range of active E-P pairs as compared to the other cell types (Fig 2.2F, A.3B). Furthermore, NET-CAGE revealed an under-representation of E-P co-transcription in d4c7 mPGCLCs (Fig A.3C). d4c7 mPGCLCs were also predicted to bear the largest numbers of insulating TAD boundaries (Fig 2.2G, A.3D), showed the smallest genomic separation (Fig 2.2H) and exhibited the broadest compartment profile (Fig A.3E), in agreement with the notion that heightened insulation can mask smaller compartments [44]. While CTCF and RAD21, a key component of cohesin, exhibited comparable enrichment at TAD boundaries across the five cell types (Fig A.3F) (we discuss the CTCF depletion in GSCs below), ATAC-seq revealed that d4c7 mPGCLCs uniquely exhibited lower chromatin information content around regions with co-localized CTCF/RAD21 bindings (Fig A.3G), suggesting that d4c7 mPGCLCs bore a shorter CTCF/RAD21 residence time [45]. Taken together, these findings support the idea that, due to a reduced residence time of the loop extrusion machinery with no major changes in global binding sites, d4c7 mPGCLCs bear shorter chromatin loops and enhanced insulation (Fig A.3H-I). Additionally, E13.5 male PGCs *in vivo* also demonstrate similarly enhanced insulation (Fig A.3J-K). We conclude that PGCs with a naïve epigenome bear a highly open chromatin, but undergo enhanced insulation to ensure their transcriptional integrity.

Insulation erasure for spermatogonia development and oogenesis

We next classified ATAC-seq peaks (open sites) based on their combinatorial epigenetic states. Building on the Ensembl Regulatory Build and ENCODE's registry of candidate Cis-Regulatory Elements (cCREs), we applied uniform manifold approximation and projection (UMAP) in combination with hierarchical density-based spatial clustering of applications with noise (HDBSCAN) in a semi-supervised manner through iterative sub-clustering (Tables A.4 and A5). This framework classified the open sites into 19 distinct sets (Fig 2.3A), which we grouped into 6 broader categories (Fig 2.3B, Fig A.4A). While d4c7 mPGCLCs showed the largest number of enhancer elements (clusters 5, 6, 15, 18) (Fig 2.3B-C), GSCs exhibited a relatively large number (~ >10,000) of non-promoter bivalent open sites (clusters 8, 9, 10, 13). Additionally, we uncovered a set of open sites with unique trivalency of H3K4me3, H3K27ac and H3K9me3 that were enriched in EpiLCs (cluster 19) (Fig 2.3B) and overlapped not only with the promoter of long interspersed nuclear elements 1 (LINE1) but also with the binding site of YY1 (Fig A.4B, Table A.5), underscoring the capacity of our epigenetic compendium for uncovering biologically distinct regulatory regions. A vast majority of enhancers were cell-type specific, whereas most CTCF bindings were conserved upon each cell-fate transition until d4c7 mPGCLCs; strikingly, however, a majority of CTCF-bound sites in d4c7 mPGCLCs were lost in GSCs (Fig A.3C) (see below).

We performed the same analyses for promoters (Fig A.4D-F, Table A.4). In accord with our previous finding [29], EpiLCs bore the largest number of bivalent promoters (Fig A.4F). Evaluation of the promoter-promoter (P-P) interactions revealed that active as well as bivalent promoters exhibited significantly enriched interactions in all cell types, but to lesser extents in d4c7 mPGCLCs bearing elevated insulation (Fig 2.2G, Fig A.4G).

We next explored the depletion of CTCF binding upon d4c7 mPGCLCs-to-GSCs transition (Fig 2.3C). In GSCs, decreased CTCF protein expression accompanied a dramatic reduction in the number of CTCF peaks (Fig 2.3D-F). In particular, CTCF was depleted from relatively weak binding sites (Fig 2.3E-F). These CTCF-depleted sites exhibited elevated DNA methylation as well as enrichment of H3K9me2/me3 and H3K36me2/me3, whereas CTCF peaks enriched in GSCs showed divergent patterns (Fig 2.3G, Fig A.4H). Importantly, despite relatively weak bindings, CTCF depletion from such sites resulted in a reduction in insulation (Fig 2.3G), leading to a rewiring of neighboring cis-regulatory interactions as exemplified for *Ddx4*, a key gene up-regulated upon d4c7 mPGCLCs-to-GSCs transition, whose promoter strengthened its long-range interaction with a distal enhancer (Fig 2.3H, Fig A.4I). We then systematically identified E-P pairs straddling CTCF sites depleted in GSCs and ranked the target genes according to coordinated expression up-regulation and increased E-P interactions (Fig A.4J). Genes with coordinated activation were enriched in gene ontology (GO) functional terms such as "homologous chromosome pairing at meiosis," and "piRNA metabolic process," and included *Ddx4*, *Mael*, *Piwil2*, *Piwil4*, *Zbtb16*, *Sycp1*, *Syce3*, *Mei4*, and *Prdm9* (Fig 2.3I, Table A.6) [these genes are referred to as "germline genes" [46] ; also, see below], indicating a critical role of the insulation erasure in spermatogonia development and the acquisition of meiotic competence.

To explore whether insulation erasure may also occur upon oogenesis, we re-analyzed published Hi-C data for E11.5 PGCs (d4c7 mPGCLC counterparts) and E13.5 germ cells initiating their male or female differentiation [4,38]. We found that a majority of E13.5 male germ cells were still in the mitotic phase and bear similar properties to E11.5 PGCs, whereas most E13.5 female germ cells were in the leptotene stage of the meiotic prophase [33,47]. Consistent with our comprehensive analyses (Fig 2.2 and 2.3), the point of fastest decline in the chromosomal cis-contact decay rate, an index for TAD width [48], occurred at the smallest genomic separation in E11.5 PGCs and d4c7 mPGCLCs (Fig 2.2H, A.3J-K), suggesting that, similar to d4c7 mPGCLCs, E11.5 PGCs bear enhanced insulation. Notably, while the fastest point of decline of E13.5 male germ cells was in a range comparable to E11.5 PGCs and d4c7 mPGCLCs, that in E13.5 female germ cells occurred at a much longer distance, suggesting a rapid weakening of insulation upon the initiation of oogenesis. We conclude that insulation erasure occurs both for spermatogonia development and oogenesis, with the latter having an earlier onset.

Mechanism for euchromatization: dynamics of LADs, pericentromeric heterochromatin, and H3K9 methylation

We next explored potential mechanisms for the progressive euchromatization unique to germ-cell development (Fig 2.1G). While the five cell types exhibited relatively conserved correlations between their compartment scores and epigenetic modification profiles, there nevertheless existed cell-type specific variations (Fig 2.4A). We noted that the binding profiles of lamin B1, which forms the nuclear lamina and tethers chromosomes to create lamina-associated domains (LADs) [49], were the strongest predictor for compartment-score differences between mESCs and GSCs (Fig 2.4B), and the LADs changed dramatically with a sweeping reduction across regions that undergo euchromatinization in GSCs (Fig 2.4C). Consequently, among a number of other cell types [50-53], GSCs bore the smallest genomic coverage of LADs (~10%) (Fig 2.4D), a vast majority of which were a subset of constitutive LADs found across all other cell types (Fig 2.4E-F). Indeed, GSCs exhibited low lamin B1 levels and enrichments (Fig 2.4G-H). Thus, GSCs constitute a cell type with minimal LADs.

While LADs were prominent toward the distal ends of long arms in mESCs and EpiLCs, they became more uniformly distributed in d2/d4c7 mPGCLCs with a reduction in their coverage in d4c7 mPGCLCs, and they eventually become depleted around the distal ends of long (q) arms in GSCs, where they were only retained towards the opposing (p/short) end, i.e., around centromeres of the telocentric mouse chromosomes (Fig 2.4C, I, and J). This is consistent with the progression of nuclear peripheral association of DAPI-dense areas along germ-cell development (Fig 2.1B-C). Accordingly, DNA FISH for major satellite repeats, a pericentromere marker, revealed that while such regions were localized mainly within the nuclear interior in EpiLCs, they were predominantly positioned around the nuclear periphery in GSCs (Fig 2.4L).

To explore whether the peripherally positioned centromeres and extensive euchromatization in other chromosomal regions in GSCs are a conserved feature in mouse spermatogonia *in vivo* and in other mammals such as primates, we re-analyzed relevant published datasets [7,38]. The distributions of chromosome-wide compartment-score differences between GSCs and EpiLCs were very similar to those between spermatogonia and fibroblasts in both mice and rhesus monkeys, with spermatogonia showing the lowest compartment score around centromeres and widespread euchromatization across other regions (note that rhesus monkeys bear metacentric chromosomes) (Fig A.5A-B). We conclude that higher-order genome organization in GSCs is conserved in spermatogonia *in vivo* and, through evolution, in monkeys.

As a mechanism that gives rise to the minimal LADs, we noted significant changes in the abundance and distributions of H3K9me2/me3, hallmarks of chromatin anchored to the nuclear lamina [54-56]. The abundance of both H3K9me2/me3 increased progressively from mESCs to d2 mPGCLCs, and then decreased dramatically in d4c7 mPGCLCs (Fig 2.2A). While the low abundance of H3K9me2 persisted in GSCs, the abundance of H3K9me3 increased in GSCs to the highest level among the five cell types (Fig 2.2A). The distributions of H3K9me2 were widespread across the chromosomes and well conserved among the five cell types except in d4c7 mPGCLCs, which, unlike the other cell types, retained H3K9me2 at a relatively high level around the pericentromeres (Fig 2.4K). On the other hand, in all cell types, H3K9me3 showed a unique and conserved distribution with a characteristic enrichment around the pericentromeres, with GSCs bearing broader/expanded H3K9me3 domains that bridge several peaks present in other cell types (Fig 2.4K, 2.5A-B). Notably, consistent with the increased B-B interactions, the broad H3K9me3 domains in GSCs exhibited elevated intra- as well as inter-domain aggregations (Fig 2.5C).

LADs consistently showed positive correlations with both H3K9me2/me3, except in GSCs, which had minimal LADs showing a positive correlation only with H3K9me3 (Fig 2.5D). IF analysis verified that GSCs showed a nuclear peripheral enrichment of H3K9me3 but not me2, while EpiLCs bore peripheral H3K9me2 but not me3 enrichment (Fig 2.5E). Interestingly, regions constitutively enriched with H3K9me3 across all five cell types, i.e., putative nucleation sites for H3K9me3 expansion in GSCs, were enriched with evolutionarily young transposable elements (TEs) including ERVK, ERV1 and LINE1 (Fig 2.5F, A.5C, Table A.5). Accordingly, the densities of these TEs were highly predictive of the minimal LADs in GSCs (Fig 2.5G, A.5D). Thus, minimal LADs in GSCs are the regions that show consistent attachment to the nuclear lamina across all cell types, likely contributing to the continued repression of evolutionarily young TEs and the maintenance of genome fidelity. Collectively, these results indicate that, during germ-cell development, LADs progressively remodel toward a minimal state, positionally shifting from the distal ends of long arms predominantly associated with H3K9me2 to the opposite ends of the chromosomes, the centromeres. These pericentromeric regions, with newfound peripheral attachment in GSCs, are predominantly associated with H3K9me3 and are populated with evolutionarily young TEs, enabling extensive euchromatization on the opposing chromosome arm (long/q arm).

Next, to gain insights into the mechanisms underlying H3K9 methylome dynamics, we examined the expression of major H3K9 methyltransferases (K9MTases). At the transcriptional level, *Suv39h1* and *h2*, which are responsible for the H3K9 methylation in the peri-centromeric heterochromatin and other B compartment regions [57], showed progressive up-regulation, whereas *Setdb1*, *Ehmt1* (*Glp1*), and *Ehmt2* (*G9a*), which are involved in the H3K9 methylation in both A and B compartments [57], were gradually repressed until d4c7 mPGCLCs and then up-regulated in GSCs (Fig 2.5H). At the protein level, SETDB1, EHMT1 and EHMT2 were repressed until d4c7 mPGCLCs and remained at a low level in GSCs as well (we were not able to determine the SUV39H1/H2 levels due to the lack of appropriate antibodies) (Fig 2.5I). These findings are consistent with the dynamics of the H3K9me2/me3 levels and distributions, suggesting that the H3K9 methylome is regulated at least in part by the differential expression of K9MTases.

Additionally, we explored the impact of the global remodeling of H3K9me3 on gene expression. In particular, we noted that during the d2-to-d4c7 mPGCLC transition, 728 promoters showed H3K9me3 down-regulation (Fig 2.5J), and they were enriched with GO terms such as "multi-organism reproductive process," "sexual reproduction," and "gamete generation," and included *Dazl*, *Ddx4*, *Sycp1*, *Sohlh2*, and *Mael* (Fig 2.5J, Table A.6). These genes, which included many subject to insulation erasure upon spermatogonia development (Fig 2.3I-J), are referred to as "germline genes" [46], and are known to be repressed by DNA methylation in somatic cells and by H3K27me3 and H3K9me2 in mPGCLCs [29,46]. Furthermore, a recent report has shown that the germline genes were repressed in EpiLCs with H3K9me3 imposed by *Setdb1* [58]. In good agreement, the transcriptional start sites (TSSs) of germline genes repressed by *Setdb1* up-regulated H3K9me3 in EpiLCs and, more prominently, in d2 mPGCLCs, and lost it in d4c7 mPGCLCs (Fig 2.5K-L). The TSSs of germline genes defined in another study [29] exhibited a comparable reduction of H3K9me3 during d2-to-d4c7 mPGCLC transition (Fig 2.5M). Thus, the germline genes are endowed with multiple layers of mechanisms, including higher-order genome organization involving the insulation by CTCF and compound repressive epigenetic modifications, to prevent their activation in somatic cells, and such mechanisms are exempted in a stepwise manner—i.e., erasure of DNA and H3K9 methylation occurs first and then release from

H3K27me3/H2AK119u1 and CTCF insulation ensues—during germ-cell development.

Heterochromatin compaction excludes H3K36me2 to create PMDs and Y-chromosome hypomethylation

A unique epigenetic characteristic of male germ cells (pro-spermatogonia, spermatogonia and spermatozoa) is the presence of large partially methylated domains (PMDs) in intergenic regions [59]. PMDs can be defined as broad genomic domains with a comparatively lower methylation level than the rest of the genome and typically cover a substantial fraction of the genome [60]. They were first identified in a human cultured cell line [60] and subsequently found to be prevalent in cancers, aged cells, and tissues such as placenta [61-63]. While evidence suggests that PMDs arise from an imperfect maintenance of methylation during mitosis [64], the mechanism that engenders PMDs in mitotically arrested pro-spermatogonia and their subsequent maintenance in male germ cells remains unclear.

We found that GSCs bore PMDs larger than 140 Mb in total, a majority (~86%) of which were overlapped with those in spermatogonia (Fig 2.6A) [59]. The PMDs in GSCs consisted almost entirely of B compartments and were enriched with heterochromatic modifications such as H3K9me3, while depleted of active modifications including H3K36me2, H3K27ac and H3K4me1/3 (Fig B.2). The epigenomic profiles revealed that the epigenome of d4c7 mPGCLCs exhibited the greatest predictive power for PMDs in GSCs (greater than that of the epigenome of GSCs themselves) (Fig 2.6B), and among individual epigenetic markers, H3K9me2/me3 and lamin B1 in d4c7 mPGCLCs were the strongest negative predictors (Fig 2.6C), suggesting that the constitutive heterochromatin in d4c7 mPGCLCs contributes to the subsequent formation of PMDs. Accordingly, we found that H3K36me2, which is catalyzed by NSD1 and serves as a recruiter of the androgenetic DNA methylome [65], showed a specific depletion in the B compartments and the regions retaining H3K9me3, but not H3K27me3, in d4c7 mPGCLCs (Fig 2.6D-F), resulting in an exquisite concordance of H3K36me2 with the A compartments and a near-complete exclusion from LADs in d4c7 mPGCLCs (Fig 2.6G). We found that the TADs involved in larger-sized TAD cliques in d4c7 mPGCLCs exhibited the greatest H3K9me3 enrichment (Fig 2.6H). Given that the heterochromatic TAD-cliques become dominant in d4c7 mPGCLCs and GSCs (Fig 2.1H-I), these findings suggest that an increased aggregation of constitutive heterochromatin in d4c7 mPGCLCs may exclude the recruitment of NSD1 and hence the deposition of H3K36me2, leading to the formation of PMDs in GSCs.

In this regard, we noted that, as compared to the autosomes and the X chromosomes, the Y chromosomes, which bear a highly repetitive structure [66], were the most enriched with H3K9me3 in all five cell types, and interestingly, exhibited a progressive enrichment of lamin B1 during germ-cell development, with the Y chromosomes in GSCs showing the highest lamin B1 enrichment level (Fig 2.6I). In addition, we found that the Y chromosome in GSCs was hypo-methylated across almost its entire length, with ~75% of it identified as falling within PMDs—a much greater proportion than in autosomes (4%) or the X chromosome (21%) (Fig 2.6J, L, and M). Indeed, by alternatively mapping directly to the consensus repeat sequences of the Y chromosome, we found that all repetitive units demonstrate reduced methylation levels in GSCs as compared to EpiLCs (Fig B.3A-B). Consistent with the de-condensation of chromatin in GSCs (Fig 2.1B-D), the Y chromosomes in GSCs exhibited loose structures and were associated with the nuclear periphery with a lower sphericity (Fig 2.6K), indicating greater surface contact with the nuclear lamina through chromosome elongation. Thus, the Y chromosome in GSCs achieves chromosome-wide hypomethylation likely via a convergent mechanism with PMDs in autosomes. Together, these results lead us to conclude that the unique 3D epigenomic character of the progenitors (d4c7 mPGCLCs) serves as a blueprint for the formation of PMDs in male germ cells.

Nucleome programming engenders gametogenic potential

To delineate the functional significance of a proper nucleome for gametogenesis, we performed nucleome analyses (morphology; *in situ* Hi-C; MS; ChIP-seq for 13 targets; ATAC-seq; and NET-CAGE) of GSC-like cells (GSCLCs), which were derived from d4 mPGCLCs with their differentiation into spermatogonia-like cells in reconstituted testes followed by expansion under a GSC derivation condition [25] (Fig 2.7A). GSCLCs derived under this condition bore a morphology, transcriptome, and DNA methylome similar to those of GSCs, but showed a severely impaired capacity for spermatogenesis for unclear reasons [25] (Fig B.4A). We hypothesized that aberrant nucleome programming during the derivation of GSCLCs might underlie their impaired function.

GSCLCs were similar to GSCs in terms of the areas of high DAPI density and the distances of the DAPI-dense areas from the nuclear periphery, but showed greater variances of DAPI density than GSCs (Fig 2.7B-C), indicating that GSCLCs bear a more heterogeneous chromatin de-condensation. *In situ* Hi-C revealed that, compared to GSCs, GSCLCs exhibited a depletion in long-range interactions, indicative of incomplete chromatin uniformalization (Fig 2.7D, B.4B), and notably, failed to acquire the positively skewed compartment score distribution characteristic of GSCs (Fig 2.7E). A multi-scale model dividing the genome into the eight subcompartments with distinct epigenetic properties [67] revealed that major difference between GSCLCs and GSCs were localized to intermediate compartments, with GSCLCs bearing fewer and more intermediate A and B sub-compartments, respectively (Fig 2.7F-G, B.4C).

Accordingly, MS revealed that GSCLCs bore an elevated level of H3K27me3 and H3K9me2, which are associated with a state intermediate between compartments A and B [68] (Fig 2.7H). The regions with higher H3K27me3 in GSCLCs were enriched in promoters and CpG islands (CGIs) (Fig B.4D, Table A.6), which were, importantly, associated with pathways such as "male meiotic nuclear division," and "recombinatorial repair," and included *Ddx4*, *Dmrt1*, *Dmc1*, *Stag3*, and *Spo11* (Fig 2.7I and J, Table A.6). These genes bore higher levels of H3K27me3 on their gene bodies as well (Fig 2.7I, B.4E). In contrast, the regions with higher levels of H3K9me2 in GSCs were enriched in enhancers and distal active regulatory elements (Fig B.4F-G), and were associated with pathways such as "response to ciliary neurotrophic factor," "rod bipolar cell differentiation," and "adrenal cortex formation" (Fig B.4H, Table A.6).

Moreover, GSCLCs bore a larger number of the CTCF-binding peaks coinciding with insufficient accumulation of H3K9me3 (Fig 2.7K, B.4I-J), and indeed GSCLCs developed higher intra-TAD interaction strength compared to GSCs (Fig 2.7L), indicating that the chromatin of GSCLCs is more insulated than that of GSCs. In a megabase-scale domain encompassing *Ddx4*, the insulating CTCF peak separating the *Ddx4* promoter from one of its potential enhancers was removed only partially in GSCLCs, resulting in a reduced activation as evidenced by the comparatively lower H3K36me3 levels on *Ddx4* (Fig 2.7M). Collectively, these results lead us to conclude that GSCLCs exhibit aberrant nucleome programming, including insulation erasure and epigenome programming, with partial retention of the properties of d4c7 mPGCLCs, resulting in their impaired spermatogenic potential.

DISCUSSION

Germ-cell development lays the groundwork for nuclear totipotency, creating sexually dimorphic haploid gametes, the oocytes and the spermatozoa, which unite to form totipotent zygotes. PGCs bear naïve epigenome after epigenetic reprogramming and can serve as a direct precursor for oocyte differentiation; they can also acquire a distinct spermatogenic epigenome, including global DNA re-methylation, to differentiate into spermatogonia/SSCs, a direct precursor for spermatozoa differentiation [19]. PGCs and spermatogonia/SSCs therefore exhibit dimorphic epigenomic properties and have been thought to represent highly distinct cellular states. Contrary to this notion, our nucleome analyses have uncovered a smooth and unidirectional maturation of higher-order genome organization from pluripotent precursors (mESCs/EpiLCs) to PGCs (d2/d4/d4c7 mPGCLCs) and then to spermatogonia/SSCs (GSCs), involving progressive euchromatization and radial chromosomal re-positioning (Fig 2.1 and 2.8). This finding delineates a common nuclear-architectural foundation towards gamete generation in both sexes, a coordination not found in somatic lineages. This widespread euchromatization might underlie the potential of GSCs to de-differentiate into pluripotent stem cells, albeit at a low frequency [69]. Thus, germ-cell development entails mechanisms that create and preserve a broadly euchromatic genome, while simultaneously accommodating essential epigenetic orchestrations. Our findings also demonstrate that global DNA methylation and euchromatization are dissociable events.

As a key mechanism for global euchromatization, we have shown that germ-cell development distinctly down-regulates H3K9me2, an aggregative force for heterochromatin formation [70], and progressively restricts LADs to around centromeres (Fig 2.2 and 2.4). These events would be mediated at least in part through the repression of SETDB1 and EHMT1, K9MTases acting in both the A and B compartments [57], as well as lamin B1 itself. On the other hand, germ cells up-regulate *Suv39h1* and *h2*, K9MTases specific to the B compartment and particularly for pericentromeric regions. This results in an expansion of H3K9me3 into broad domains in GSCs with an appreciable increase in both local and distal compaction among such domains (Fig 2.5), consistent with the notion of a critical threshold of H3K9me3 domain width for phase separation to take place via HP1 [71]. This compaction would also contribute to the formation of PMDs, and most remarkably, those on the Y chromosome, likely by physically excluding spermatogenesis-associated NSD1 and preventing H3K36me2 depositions (Fig 2.6). Thus, typical LADs mediated by H3K9me2, which are seen in pluripotent precursors as well as in most somatic lineages, are progressively re-organized into a minimal state marked by H3K9me3 during germ-cell development. Importantly, the positional preference of H3K9me3-associated minimal LADs is in part attributable to the density of evolutionarily young TEs that are enriched near centromeres (Fig 2.5, A.5), indicating a critical role of inherent genomic properties in shaping the fundamental nuclear architecture. In good agreement with this concept, cell-type specific LADs have been reported to be enriched in such TEs [72]. The involvement of H3K9 demethylases and the interplay among associated machineries for LAD formation warrant further investigation.

Despite adopting a highly permissive epigenome with abundant enhancer-like open sites, d4c7 mPGCLCs strengthened their chromatin insulation to thwart spurious distal activation, which, combined with a mechanism to ensure low H3K4me3 levels, would prevent the pervasive poised enhancers from realizing their potential (Fig 2.2 and 2.3). Thus, epigenetic reprogramming creates PGCs that have almost no DNA methylation and a highly open epigenome, but that are protected by elevated H3K27me3 [22] and CTCF insulation against hyper-transcription. As to a possible mechanism for the enhanced insulation, we revealed a reduced residence time of the loop extrusion machinery at TAD boundaries in d4c7 mPGCLCs (Fig A.3H-J). Such a reduction in residence time could be achieved through multiple mechanisms, including the use of variant cohesin complexes and modulating the balance between cohesin loading/release factors [73,74]. Clarification of these potential mechanisms warrants future investigation.

On the other hand, such protective mechanisms must be at least partly disentangled upon male and female germ-cell specification to eventually achieve full activation of the gametogenic program. Accordingly, a failure of such unraveling and a partial retention/aberrant development of the PGC-like nucleome together contributed to the limited spermatogenic capacities of GSCLCs (Fig 2.6, B.4). In the original GSCLC induction strategy, d4 mPGCLCs, which are in the middle of epigenetic reprogramming and bear ~50% genome-wide DNA methylation, were aggregated with embryonic testicular somatic cells for differentiation into spermatogonia-like cells [25]. We speculate that precocious testicular sex-determining signals on mPGCLCs might be a reason for mis-organized nucleome in the originally reported GSCLCs. In good agreement with this speculation, we have recently succeeded in deriving fully functional GSCLCs using d4c5 mPGCLCs, which have an almost fully complete epigenetic reprogramming, as starting materials for aggregation culture with embryonic testicular somatic cells [75]. The nucleome analysis of these newly established GSCLCs would be important to confirm this hypothesis.

The nucleome programming for germ-cell development that we have delineated herein, which involves progressive euchromatization with peripheral centromere positioning, is reminiscent of climbing up the Waddington's landscape of epigenesis (Fig 2.8), and we propose that it constitutes at least part of the mechanism for creating nuclear totipotency, including meiotic potential. Elucidation of the nucleome programming during germ-cell development in other mammals, including humans, will be crucial for a more comprehensive understanding of nuclear totipotency and its evolutionary divergence. The rich datasets we have assembled would be invaluable as a benchmark for mammalian *in vitro* gametogenesis studies [15] and for future studies aiming to identify unifying principles for the acquisition of unique cellular identities across lineages. Further, they could contribute to the development of powerful computational frameworks, which in turn could help integrate time-series multi-omics datasets and unveil hidden insights.

ACKNOWLEDGMENTS

We thank the members of our laboratory for their helpful input on this study. We are grateful to Y. Nagai, N. Konishi, E. Tsutsumi, and M. Kawasaki of the Saitou Laboratory, to the DNAFORM genetic analysis department for NET-CAGE library preparation and sequencing, to the Single-Cell Genome Information Analysis Core (SignAC) in ASHBi for their technical assistance and help with all the other sequencing experiments, to R. Maeda of the Tachibana Laboratory for helpful suggestions for ChIP-seq, to S. Nagaoka and K. Kurimoto of the Kurimoto Laboratory and G. Bourque of the Bourque Laboratory for thoughtful discussions on the data analysis, and to C. Horth from the Majewski Laboratory for her assistance with histone extraction. This work was supported in part by a Grant-in-Aid for Specially Promoted Research from JSPS (17H06098), a JST-ERATO Grant (JPMJER1104), a Grant from HFSP (RGP0057/2018), Grants from the Pythias Fund and Open Philanthropy Project to M.S., JSPS KAKENHI Grants (JP18H02613, JP20H05387) to S.Y., and NIH grants (CA196539, NS111997) to B.A.G. M.N. is a fellow of the Takeda Science Foundation. B.H. is supported by studentship awards from the Canadian Institutes of Health Research and the Fonds de recherche du Québec – Santé.

AUTHOR CONTRIBUTIONS

M.N., B.H., S.Y., and M.S. conceived the project and designed experiments. M.N. performed all cell cultures and inductions with assistance from H.O., Y.Ishikura. and Y.N. M.N. performed immunofluorescence and its analysis with assistance from H.O., N.K. and K.M. M.N. and F.U. performed western blot and its analysis with assistance from Y.N., S.S., and Y.K. M.N. performed FISH with assistance from I.O. M.N. performed histone extraction and M.C. performed mass spectrometry under the supervision of B.A.G. M.N. and A.Y. performed ChIP-seq with assistance from S.Y. and T.Y. M.N. performed ATAC-seq with assistance from S.Y., H.I., and T.Y. M.N. performed in situ Hi-C with assistance from S.Y., R.S., G.W., K.T., J-M.P., and L.A.M. M.N. performed NET-CAGE with assistance from T.K. under the supervision of Y.M. M.N. and B.H. performed all data analysis with assistance from Y.Y. and J.M. B.H. performed polymer simulation and analysis with assistance from Y.Imoto., K.M., and Y.H. M.N., B.H., and M.S. wrote the manuscript with input from all co-authors. S.Y., J.M., and M.S. supervised the project.

DISCLOSURE AND COMPETING INTERESTS STATEMENT

M.S. is an associate EMBO member.

MATERIALS AND METHODS

Table 2.1 Reagents and Tools

Reagent/Resource	**Reference or Source**	**Identifier or Catalog Number**
Experimental Models		
AAG 129/B6 GSC2 (Acrosin-EGFP; beta-Actin-EGFP, 129Sv×C57BL/6, P7 spermatogonia, Germline stem cell line)	[25]	https://doi.org/10.1016/j.celrep.2016.11.026
AAG 129/B6 GSCLC16_1 (Acrosin-EGFP; beta-Actin-EGFP, 129SvJ×C57BL6, Germline stem cell-like line, derived from mESCs)	[25]	https://doi.org/10.1016/j.celrep.2016.11.026
BVSC BDF1-2-1 mESCs (Blimp1-mVenus; Stella-ECFP , DBA/2×C57BL/6, embryonic stem cell line)	[23]	https://doi.org/10.1093/biolre/ioaa195
m220-5 (sub-cloned from Sl/Sl4-m220, resistant to mitomycin C, expressing membrane-bound SCF, stromal cell)	[22]	https://doi.org/10.15252/embj.201695862
MEF (ICR, mitomycinC-treated mouse embryonic fibroblasts prepared from E12.5 fetuses)	N/A	N/A
Antibodies		
Anti-CTCF	CST	#3418
Anti-G9a	R&D Systems	PP-A8620A-00
Anti-GFP(Rat IgG2a), Monoclonal(GF090R), CC	Nacalai Tesque	04404-84
Anti-GLP	R&D Systems	PP-B0422-00
Anti-H2Aub	CST	#8240
Anti-H3	CST	#9715
Anti-H3K27ac	MBL	MABI0309
Anti-H3K27me3	MBL	MABI0323
Anti-H3K27me3	Merk	07-449
Anti-H3K36me2	CST	#2901
Anti-H3K36me3	Active Motif	61101
Anti-H3K4me1	CST	#5326
Anti-H3K4me3	MBL	MABI0304
Anti-H3K9me2	MBL	MABI0317
Anti-H3K9me3	MBL	MABI0318
Anti-Laminb1	Proteintech	12987-1-AP
Anti-Laminb1	Abcam	ab16048
Anti-mouse IgG (whole molecule)–peroxidase antibody produced in sheep affinity isolated	Sigma	A5906-1ML

antibody, buffered aqueous solution		
Anti-normal mouse IgG	Santa Cruz	sc-2025
Anti-normal rabbit IgG	Santa Cruz	sc-2027
Anti-rabbit IgG (whole molecule)–peroxidase antibody produced in goat affinity isolated antibody, buffered aqueous solution	Sigma	A6154-1ML
Anti-Rad21	Abcam	ab992
Anti-Ring1b	CST	#5694
Anti-Setdb1	Proteintech	11231-1-AP
Anti-α-tubulin	Sigma	T9026
Anti-β-actin	MBL	M177-3
Goat anti-mouse IgG (H+L) highly cross-adsorbed secondary antibody, Alexa Fluor 568	Invitrogen	A-11031
Goat anti-rabbit IgG (H+L) cross-adsorbed secondary antibody, Alexa Fluor 568	Invitrogen	A-11011
Goat anti-rat IgG (H+L) cross-adsorbed secondary antibody, Alexa Fluor 488	Invitrogen	A-11006
Oligonucleotides and sequence-based reagents		
XMP 1 orange	MetaSystems	D-1401-050-OR
XMP 16 orange	MetaSystems	D-1416-050-OR
XMP Y orange	MetaSystems	D-1421-050-OR
Chemicals, enzymes, and other reagents		
16% Formaldehyde solution	Thermo Fisher Scientific	28906
2-Mercaptoethanol	Nacalai Tesque	21438-82
20xSCC	Sigma	S6639
37% Formaldehyde(FA)	Sigma	252549
4% Paraformaldehyde	Nacalai Tesque	26126-25
4X Laemmli sample buffer	Bio-Rad	#1610747
Activin A (human/mouse/rat)	Peprotech	120-14
AlbuMaxI	Gibco	11020062
Amanitin 1mg	Wako	1022961
Apo transferrin	Sigma	T1147
Axygen® AxyPrep MAG PCR Clean-Up Kit	Corning	MAG-PCR-CL-250
B27	Thermo Fisher Scientific	12587010
bFGF	Invitrogen	13256029
Biotin-14-dATP	Thermo Fisher Scientific	19524-016
Bovine serum albumin cold ethanol fraction, pH 5.2, ≥96%	Sigma	A4503-10G
BSA fraction V	Gibco	15260-037
CHIR99021	Bio Vision	4423

cOmplete™, protease inhibitor cocktail	Roche	4693116001
cOmplete™, EDTA-free protease inhibitor cocktail	Roche	11873580001
cOmplete™, mini, EDTA-free	Roche	4693159001
Cyclosporin A	Sigma	30024
DAPI	Wako	342-07431
Difco™ skim milk	BD Biosciences	232100
Digitonin	Promega	G9441
DMEM/F12	Gibco	11330-057
DMEM/F12 (phenol red free)	Thermo Fisher Scientific	21041025
DNA polymerase I, large (Klenow) fragment	NEB	M0210S
DNaseI 1 unit/ul, RNase-free	Thermo Fisher Scientific	89836
DpnII	NEB	R0543L
DTT 100 ul	Promega	P1171
Dynabeads M-280 sheep anti-mouse IgG	Thermo Fisher Scientific	DB11201
Dynabeads protein A	Thermo Fisher Scientific	DB10001
Dynabeads® MyOne™ Streptavidin C1	Thermo Fisher Scientific	65001
EGF, mouse, recombinant, carrier-free	RSD	2028EG
Fetal bovine serum (FBS)	Hyclone	SH30910.03
Fibronectin (human)	Merck Millipore	FC010
Formamide	Nacalai Tesque	16228-05
Forskolin	Sigma	F3917
GDNF, rat, recombinant	RSD	512GF
Glasgow's MEM (GMEM)	Thermo Fisher Scientific	11710035
GlutaMAX supplement	Life Technologies	35050061
Immobilon-P PVDF membrane	Merck Millipore	IPVH00010
Insulin	Sigma	#I-1882
Insulin-transferrin-selenium (ITS)-G	Gibco	41400045
KnockOut™ serum replacement	Gibco	10828028
L-Glutamine	Thermo Fisher Scientific	25030149
Laminin	BD Bioscience	354232
LIF(ESGRO®)	Merck Millipore	ESG1107
MEM non-essential amino acids solution	Thermo Fisher Scientific	11140-050
Minimum Essential Medium (MEM) Vitamin Solution	Thermo Fisher Scientific	11120052
Mitomycin C	kyowakirin	KWN-057039107
NEBNext High-Fidelity 2x PCR Master Mix	NEB	M0541S
Neurolbasal™ medium	Invitrogen	2113-049
Nuclei EZ Prep	Sigma	NUC101
Orange-dUTP	Abbott	02N33-050

PD325901	Stemgent	04-2006
Penicillin-Streptomycin (10,000 units/mL, 10,000 µg/mL)	Thermo Fisher Scientific	15140148
PhosSTOP™	Roche	4906837001
Pierce™ Protease Inhibitor Mini Tablets, EDTA-free	Thermo Fisher Scientific	A32955
Poly-L-ornithine	Sigma	P3655
Progesterone	Sigma	P8783
Proteinase K solution	Thermo Fisher Scientific	AM2546
Putrescine	Sigma	P5780
Recombinant Human BMP-4	RSD	314BP01M
Recombinant Mouse SCF	RSD	455MC
RIPA lysis buffer system 50ml	Santa Cruz	SC-24948
RNase A	Thermo Fisher Scientific	EN0531
Rolipram	Abcam	AB120029
Sodium pyruvate	Thermo Fisher Scientific	11360-070
Sodium selenate	Sigma	S5261
StemPro™-34 SFM (1X)	Gibco	10639011
SUPERase	Thermo Fisher Scientific	AM2694
T4 DNA ligase 1U/ µl	Thermo Fisher Scientific	15224090
T4 DNA polymerase	NEB	M0203L
Tks Gflex™ DNA Polymerase	Takara	R060A
TryPLE-Express	Thermo Fisher Scientific	12604-021
VECTASHIELD® Antifade Mounting Medium	Vector Laboratories	H-1000-10
β-Mercaptoethanol	Thermo Fisher Scientific	21985023
Software		
ABC commit 7fd69b0	[43]	https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction
BEDTools v2.29.2	[76]	https://github.com/arq5x/bedtools2
Bismark v0.22.1	[77]	https://www.bioinformatics.babraham.ac.uk/projects/bismark/
Bowtie2 v2.3.4.1	[78]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
CAGEfightR v1.7.6	[79]	https://bioconductor.org/packages/release/bioc/html/CAGEfightR.html
CAGEr v1.32.0	[80]	https://www.bioconductor.org/packages/release/bioc/html/CAGEr.html
CALDER commit 32220e8	[67]	https://github.com/CSOgroup/CALDER
Chrom3D	[81]	https://github.com/Chrom3D/pipeline
ChromA v2.1.1	[82]	https://github.com/marianogabitto/ChromA
Chromosight v1.5.1	[83]	https://github.com/koszullab/chromosight
cooler v0.8.10	[84]	https://github.com/open2c/cooler
coolpup.py v0.9.7	[85]	https://github.com/open2c/coolpuppy
cooltools v0.4.0	[86]	https://github.com/open2c/cooltools
CSynth commit 26e21fb	[37]	https://github.com/csynth/csynth
Cutadapt v1.9.1	[87]	https://cutadapt.readthedocs.io/en/stable/

dcHiC commit 7b1727f	[88]	https://github.com/ay-lab/dcHiC
deepTools v3.5.0	[89]	https://github.com/deeptools/deepTools
DESeq2 v1.28.1	[90]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
DiffBind v3.0.13	[91]	https://bioconductor.org/packages/release/bioc/html/DiffBind.html
EDD v1.1.19	[92]	https://github.com/CollasLab/edd
epic2 v0.0.41	[93]	https://github.com/biocore-ntnu/epic2
EpiProfile v2.0	[94]	https://github.com/zfyuan/EpiProfile2.0_Family
FACSDiva Software	BD Biosciences	N/A
FAN-C v0.9.13	[95]	https://github.com/vaquerizaslab/fanc
fastp v0.21.0	[96]	https://github.com/OpenGene/fastp
GimmeMotifs v0.15.3	[97]	https://github.com/vanheeringen-lab/gimmemotifs
HDBSCAN v0.8.27	[98]	https://github.com/scikit-learn-contrib/hdbscan
HiCKey commit 6e282b9	[99]	https://github.com/YingruWuGit/HiCKey
HiCRep.py v0.2.3	[100]	https://github.com/Noble-Lab/hicrep
HiCRes v1.1	[101]	https://github.com/ClaireMarchal/HiCRes
HiCSeg v1.1	[102]	https://cran.r-project.org/web/packages/HiCseg/index.html
HiCUP v0.8.0	[103]	https://github.com/StevenWingett/HiCUP
IDR2D v1.4.0	[104]	https://github.com/kkrismer/idr2d
Imaris v9.1.2	N/A	https://imaris.oxinst.com/
Juicer tools v1.22.01	[105]	https://github.com/aidenlab/juicer
MACS v2.1.1	[106]	https://github.com/macs3-project/MACS
OnTAD v1.2	[107]	https://github.com/anlin00007/OnTAD
Picard Tools v2.18.23	N/A	https://broadinstitute.github.io/picard/
Python v3.8.8	N/A	https://www.python.org/
R (v4.0.3)	https://www.r-project.org/	https://www.r-project.org/
RobusTAD	[108]	https://github.com/rdali/RobusTAD
S3V2-IDEAS commit b7cc2d5	[109]	https://github.com/guanjue/S3V2_IDEAS_ESMP
Salmon v1.4.0	[110]	https://github.com/COMBINE-lab/salmon
SAMtools v1.7	[111]	https://github.com/samtools/samtools
SpectralTAD v1.4.0	[112]	https://github.com/dozmorovlab/SpectralTAD
TADpole 0.0.0.9000	[113]	https://github.com/3DGenomes/TADpole
TopDom v0.10.1	[114]	https://github.com/HenrikBengtsson/TopDom
Trim-Galore! v0.6.3	[115]	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
tximport v1.16.1	[116]	https://github.com/mikelove/tximport
UMAP v0.5.1	[117]	https://github.com/lmcinnes/umap
Other		
Chemi-Lumi One Super	Nacalai Tesque	02230-14
FastGene Adapter Kit	FastGene	FG-NGSAD24
Illumina Tagment DNA Enzyme and Buffer, Small Kit	Illumina	20034197
KAPA Hyper Prep Kit	KAPA	KK8504

KAPA Library Quantification Kit	KAPA	KK4824
MinElute PCR purification Kit (50)	QIAGEN	28004
miRNeasy Mini Kit 50	QIAGEN	217004
NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1)	NEB	E7335S
NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®	NEB	E7645S
NextSeq 500/550 High Output Kit v2.5 （150 Cycles）	Illumina	20024907
NextSeq 500/550 High Output Kit v2.5 （75 Cycles）	Illumina	20024906
NextSeq 500/550 High-Output v2 Kit (150 cycles)	Illumina	FC-404-2002
NextSeq 500/550 High-Output v2 Kit (75 cycles)	Illumina	FC-404-2005
NextSeq 500/550 Mid Output Kit v2.5 （150 Cycles）	Illumina	20024904
NextSeq 500/550 Mid-Output v2 Kit (150 cycles)	Illumina	FC-404-2001
NovaSeq 6000 S1 Reagent Kit (200 cycles)	Illumina	20012864
NovaSeq 6000 SP Reagent Kit (200 cycles)	Illumina	20040326
QIAquick PCR Purification Kit (50)	QIAGEN	28104
Qubit RNA HS Assay Kit	Thermo Fisher Scientific	Q32855
RNase-Free DNase Set	QIAGEN	79254
Film-bottom dish	Matsunami Glass	FD10300
MAS-GP type A	Matsunami Glass	S9901-9905

Methods and Protocols

Culture of mESCs

The BDF1-2-1 mouse mESCs bearing *Blimp1-mVenus* and *Stella-ECFP* (BVSC) transgenes [23] were cultured as described previously [21]. Briefly, mESCs were maintained in N2B27 medium supplemented with PD0325901 (0.4 uM) (Stemgent, 04-2006), CHIR99021 (3 uM) (Bio Vision, 4423), and leukemia inhibitory factor (LIF) (1000 U/ml) (Merck Millipore, ESG1107) on a 12-well plate coated with poly-L-ornithine (0.01%) (Sigma, P3655) and laminin (10 ng/ml) (BD Biosciences, 354232). In this study, all cells were cultured at 37°C under an atmosphere of 5% CO2 in air.

Induction of EpiLCs and mPGCLCs

Induction of EpiLCs and PGCLCs was performed as described previously [21] with minor modifications. Briefly, the EpiLCs were induced by plating 8×10^4 mESCs on a well of a 12-well plate coated with human plasma fibronectin (16.7 mg/ml) (Merck Millipore, FC010) in N2B27 medium containing activin A (20 ng/ml) (Peprotech, 120-14), bFGF (12 ng/ml, 13256029) (Invitrogen), and KSR (1%) (Gibco, 10828028). mPGCLCs were induced from d2 EpiLCs (2 days after induction) under a floating condition in wells of a low-cell-binding U-bottom 96-well plate in GMEM medium (Thermo Fisher Scientific, 11710035) containing 15% KSR (Gibco, 10828028), 0.1 mM NEAA (Thermo Fisher Scientific, 11140-050), 1 mM sodium pyruvate (Thermo Fisher Scientific, 11360-070), 0.1 mM β-mercaptoethanol (Thermo Fisher Scientific, 21985023), 100 U/ml penicillin, 0.1 mg/ml streptomycin (Thermo Fisher Scientific, 15140148) and 2 mM L-glutamin (Thermo Fisher Scientific, 25030149) supplemented with BMP4 (500 ng/ml) (RSD, 314BP01M), LIF (1000 U/ml) (Merck Millipore, ESG1107), SCF (100 ng/ml) (RSD, 455MC), and EGF (50 ng/ml) (RSD, 2028EG).

Expansion culture of mPGCLCs

The expansion culture of mPGCLCs was performed as previously described [23]. Briefly, following incubation in TrypLE™ Express (Gibco, 12604-021) for 10 min, the aggregates of d4 mPGCLCs (PGCLCs induced for 4 days) were dissociated into single cells by rigorous pipetting. Subsequently, BV-positive cells were sorted with a FACSAria III cell sorter. Purified d4 mPGCLCs were cultured on m220-5 cells as the feeder cells in GMEM (Gibco, 11710035) containing 10% KSR (Gibco, 10828028), 0.1 mM NEAA (Thermo Fisher Scientific, 11140-050), 1 mM sodium pyruvate (Thermo Fisher Scientific, 11360-070), 0.1 mM β-mercaptoethanol (Thermo Fisher Scientific, 21985023), 100 U/ml penicillin, 0.1 mg/ml streptomycin (Thermo Fisher Scientific, 15140148), 2 mM L-glutamin (Thermo Fisher Scientific, 25030149), 2.5% FBS (Hyclone, SH30910.03), SCF (100 ng/ml) (RSD, 455MC), 10 mM forskolin (Sigma, F3917), 10 uM rolipram (Abcam, AB120029), and 5 uM CsA (Sigma, 30024). Half of the culture medium was changed every two days.

Culture of GSCs and GSCLCs

GSCs and GSCLCs bearing *Acrosin-EGFP* and beta-*Actin-EGFP* (AAG) transgenes [118] were cultured as described previously [25]. Briefly, cells were cultured in Stempro-34 SFM supplemented with Stempro Supplement (Gibco, 10639011), with 0.1 mM β-mercaptoethanol (Thermo Fisher Scientific, 21985023), 1% FBS (Hyclone, SH30910.03), 1×MEM vitamin solution (Thermo Fisher Scientific, 11120052), 5.0 mg/ml AlbMAXI (Gibco, 11020062), 0.1 mM NEAA (Thermo Fisher Scientific, 11140-050), 1 mM sodium pyruvate (Thermo Fisher Scientific, 11360-070), 0.1 mM β-mercaptoethanol (Thermo Fisher Scientific, 21985023), 100 U/ml penicillin, 0.1 mg/ml streptomycin (Thermo Fisher Scientific, 15140148), 2 mM L-glutamin (Thermo Fisher Scientific, 25030149), 1×Insulin-Transferrin-Selenium (ITS-G) (Gibco, 41400045), 10 ng/ml bFGF (Invitrogen, 13256029), 20 ng/ml GDNF rat recombinant (RSD, 512GF), 20 ng/ml EGF (RSD, 2028EG), and 1000 U/ml LIF (Merck Millipore, ESG1107) in a well of a 6-well plate on mouse embryonic fibroblast cells (MEFs) as feeder cells. Half of the medium was replaced every two or three days.

Immunofluorescence staining

The following primary antibodies were used at the indicated dilutions: rabbit anti-Laminb1 (1/1000; Abcam ab16048); mouse anti-H3K9me2 (1/500; MBL, MABI0317); mouse anti-H3K9me3 (1/500; MBL, MABI0318); and mouse anti-H3K27me3 (1/500; Merk, 07-449).

The following secondary antibodies from Thermo Fisher Scientific were used at a 1/500 dilution: Alexa Fluor 568 goat anti-rabbit IgG; Alexa Fluor 488 goat anti-rabbit IgG; and Alexa Fluor 568 goat anti-mouse IgG.

Immunofluorescence (IF) staining was performed as previously described [22] with minor modifications. Briefly, cells were fixed in 4% PFA (paraformaldehyde) (Nacalai Tesque, 26126-25) for 30 min at RT. After fixation, cells were washed in PBS three times and then permeabilized in 1% Triton-X100/PBS for 5 min on ice. Then, they were washed in PBS three times and incubated in 1% BSA (Sigma, A4503-10G)/PBS for 1 h. The cells were incubated with primary antibodies in 1% BSA/PBS overnight. After incubation with primary antibodies, the cells were washed in PBS three times and then incubated for 2 h with secondary antibodies and DAPI (1 mg/ml) (Wako, 342-07431) at RT. Then, they were washed three times in PBS and mounted in VECTOR SHIELD (Vector Laboratories, H-1000-10). Images were captured with a confocal microscope (LSM780 or LSM980 with Airyscan2; Zeiss).

Probe preparation for DNA-FISH against major satellite repeats

The probe against major satellite repeats was generated as previously described [119] with some modifications. DNA fragments were amplified with forward (5’-GCGAGAAAACTGAAAATCAC-3’) and reverse (5’-TCAAGTCGTCAAGTGGATG-3’) primers using mouse genomic DNA as a template, and purified using a

QIA quick PCR purification kit (QIAGEN, 28104). 500 ng of the PCR product was labeled with Orange-dUTP (Abbott, 02N33-050) using a Nick translation kit (Roche, 10976776001).

DNA-FISH

DNA-FISH was performed as described previously [120]. Briefly, cells were cultured in a film-bottom dish (Matsunami Glass, FD10300) and fixed in 3% PFA/PBS (Nacalai Tesque, 26126-25) for 10 min at RT. After a brief wash in PBS, cells were permeabilized in 0.5% Triton-X100 in PBS for 5 min on ice and stored in 70% ethanol at −30°C by the day of use. Then, the DNA was denatured in 50% FA (formamide) (Nacalai Tesque, 16228-05)/2×SSC pH 7.4 (Sigma, S6639) for 40 min at 80°C and dehydrated through an ice-cold ethanol series. Hybridization with probes was performed at 37°C overnight. After incubation, the samples were washed in 50% FA/2×SSC followed by 2×SSC. The samples were counterstained with DAPI (1 mg/ml) (Wako, 342-07431), and mounted and viewed under a confocal microscope (Zeiss LSM980 with Airyscan2). Images were analyzed using Imaris 9.1.2 software (Bitplane).

Western blot analysis

The following primary antibodies were used at the indicated dilutions: rabbit anti-Lamin b1 (1/1000; Abcam ab16048); mouse anti-H3K9me2 (1/500; MBL, MABI0317); mouse anti-H3K9me3 (1/500; MBL, MABI0318); and mouse anti-H3K27me3 (1/500; MBL, MABI0323); rabbit anti-H3 (1/10000; CST, #9715); rabbit anti-CTCF (1/500; CST, #3418); mouse anti-G9a (1/500; R&D, PP-A8620A-00); mouse anti-GLP (1/500; R&D, PP-B0422-00); rabbit anti-Setdb1 (1/1000; Proteintech, 11231-1-AP); mouse anti-α-tubulin (1/5000; Merk, T9026); and mouse anti-β-actin (1/5000; MBL, M177-3).

The following secondary antibodies from Merk were used at the indicated dilutions: goat anti-rabbit IgG conjugated with peroxidase (1/8000); and sheep anti-mouse IgG conjugated with peroxidase (1/10000).

Western blot was performed as previously described [21] with slight modifications. Briefly, cells were lysed by RIPA buffers (Santa Cruz, SC-24948). After incubation for 30 min at 4°C with rotation, the lysates were sonicated by Bioruptor using 10 cycles of 30 s on/30 s off. Then, the lysates were spun down at 14000 rpm for 15 min at 4°C and the supernatant was collected. A BCA assay was performed using a Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, 23227) to measure the protein concentration. For western blot, 4.5 mg of whole cell lysate or 2.25 mg of chromatin fraction was loaded onto each lane. After addition of 4×Laemmli buffer (Bio-Rad, #1610747), the sample was run by SDS-PAGE, followed by blotting to PVDF membrane (pore size: 0.45 mm) (Millipore, IPVH00010) in CAPS buffer (10 mM CAPS-NaOH pH 11, 5% methanol). After blotting, the membrane was incubated for 1 h in 0.1% Tween-20/PBS (PBST) with 1% skim milk (BD Bioscience, 232100). After blocking, the membrane was incubated overnight with the primary antibodies in 0.1% PBST with 1% skim milk. The membrane was washed in 0.1% PBST, followed by incubation for 2 h with the secondary antibodies in the 0.1% PBST with 1% skim milk. After washing in 0.1% PBST three times, secondary antibodies were detected by Chemilumi One Super (Nacalai Tesque, 02230-14) using Fusion solo 4S (Vilber). Quantification analysis of the signal intensity was performed in ImageJ v2.1.0 (NIH). Target protein signals were normalized by the loading control.

Chromatin fraction isolation

Chromatin fractionation was performed as previously described [73]. In brief, cells were resuspended in extraction buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 2 mM NaF, 10% glycerol, 0.2% NP-40, 20 mM β-glycerophosphate, 0.5 mM DTT, and protease inhibitor cocktail (Roche, 11873580001)). The chromatin pellet was fractionated by centrifugation at 2000 g for 5 min and washed in the same buffer three times. Then, the chromatin pellet was resuspended in RIPA buffer (Santa Cruz, SC-24948) and processed along with the whole cell lysate by a downstream BCA assay (Thermo Fisher Scientific, 23227) followed by western blot.

Visualization and analysis of nuclei by DAPI staining

All cells except d2 mPGCLCs were cultured in a film-bottom dish (Matsunami Glass, FD10300). d2 mPGCLCs were attached on a slide glass (MATSUNAMI, S9901-9905) using Cyto Spin 4 (Thermo Fisher Scientific) as previously described [22]. Cells were fixed in 4% PFA (Nacalai Tesque, 26126-25) at RT for 30 min and washed in PBS three times. For permeabilization, cells were incubated on ice in 0.5% TritonX-100/PBS for 5 min. Then, cells were incubated in DAPI solution (1 mg/ml) (Wako, 342-07431) for 8 min, mounted and viewed under a fluorescence microscope. Confocal z-series images with an interval of 0.14 μm were captured by Zeiss LSM980 with Airyscan2 using a 405 nm wavelength and a 63×objective oil-immersion lens. For DAPI-staining analysis, cells were attached to slides using Cyto Spin 4 (Thermo Fisher Scientific) as previously described [22] in order to avoid the effect of differences in their colony shapes. DAPI-staining and image acquisition were performed as described above. Acquired images were processed as follows. The nuclear mask, nuclear rim, and DAPI dense regions were defined in each z-slice using ImageJ custom script as previously described [121]. Then, the slice showing the maximum diameter was decided for each cell as a representative slice, and the representative slice ±5 slices for each cell (i.e., 11 slices/cell) were used in the downstream analysis. Approximately 20–30 cells were analyzed in each cell type. The parameters presented in the Figures were calculated using R custom script.

Histone extraction for mass spectrometry

Frozen cell pellets containing 3 million cells were lysed in nuclear isolation buffer (15 mM Tris pH 7.5, 60 mM KCl, 15 mM NaCl, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 250 mM sucrose, 10 mM sodium butyrate, 0.1% v/v b-mercaptoethanol (Nacalai Tesque, 21438-82), commercial phosphatase and protease inhibitor cocktail tablets (Roche, 4906837001; Thermo Fisher Scientific, A32955)) containing 0.3% NP-40 alternative on ice for 5 min. Nuclei were washed in the same solution without NP-40 twice and the pellet was slowly resuspended while vortexing in chilled 0.4 N H_2SO_4, followed by 3 h of rotation at 4°C. After centrifugation, the supernatants were collected and proteins were precipitated in 20% TCA overnight at 4°C, washed once with 0.1% HCl (v/v) acetone and then twice with acetone only, and resuspended in deionized water. Acid-extracted histones (20–50 μg) were resuspended in 100 mM ammonium bicarbonate pH 8, derivatized using propionic anhydride and digested with trypsin as previously described [122]. After the second round of propionylation, the resulting histone peptides were desalted using C18 Stage Tips, dried using a centrifugal evaporator and reconstituted using 0.1% formic acid in preparation for liquid chromatography-mass spectrometry (LC–MS) analysis.

LC/LC-MS

Nanoflow liquid chromatography was performed using a Thermo Fisher Scientific Dionex UltiMate 3000 LC system equipped with a 300mm ID x 0.5-cm trap column (Thermo) and a 75 mm ID x 20-cm analytical column packed in-house using Reprosil-Pur C18-AQ (3 mm; Dr. Maisch). Buffer A was 0.1% formic acid and Buffer B was 0.1% formic acid in 80% acetonitrile. Peptides were resolved using a two-step linear gradient from 5% B to 33% B over 45 min, then from 33% B to 90% B over 10 min at a flow rate of 300 nL min^{-1}. The HPLC was coupled online to an Orbitrap QE-HF mass spectrometer operating in the positive mode using a Nanospray Flex Ion Source (Thermo Fisher Scientific) at 2.3 kV. Two full mass spectrometry scans (m/z 300–1,100) were acquired in the Orbitrap Fusion mass analyzer with a resolution of 120,000 (at 200 m/z) every 8 data-independent acquisition tandem mass spectrometry (MS/MS) events, using isolation windows of 50 m/z each (for example, 300–350, 350–400, 650–700). MS/MS spectra were acquired in the ion trap operating in normal mode. Fragmentation was performed using collision-induced dissociation in the ion trap mass analyzer with a normalized collision energy of 35. The automatic gain control target and maximum injection time were 5×10^5 and 50 ms for the full mass spectrometry scan, and 3×10^4 and 50 ms for the MS/MS scan, respectively. Raw files were analyzed using EpiProfile 2.0 [94]. The area for each modification state of a peptide was normalized against the total signal for that peptide to give the relative abundance of the histone modification.

ChIP-seq library preparation and sequencing

The ChIP-seq library preparation was performed as previously described [123] with minor modifications. We used harvested mESCs and EpiLCs, and FACS-sorted BV-positive cells for d2 mPGCLCs and d4c7 mPGCLCs samples, and FACS-sorted AAG-positive cells for GSCs and GSCLCs samples. Briefly, the harvested cells were crosslinked with 1% formaldehyde (Thermo Fisher Scientific, 28906)/PBS for 10 min at RT and quenched with 125 mM glycine. Crosslinked cells were lysed consecutively using LB1 (50 mM HEPES-KOH pH 7.5, 1 mM EDTA, 140 mM NaCl, 10% glycerol, 0.5% NP-40, 0.25% Triton-100, protease inhibitors (Roche, 11873580001)), LB2 (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 200 mM NaCl, protease inhibitors), and LB3 (50 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine, protease inhibitors) and then sonicated by a picoruptor to achieve a mean DNA fragment size of around 200–400 bp. Sonicated chromatin was incubated with Dynabeads M-280 Sheep anti-Mouse IgG beads (Thermo Fisher Scientific, DB11201) or Dynabeads ProteinA beads (Thermo Fisher Scientific, DB10001) for 35 min at 4°C for preclear. Precleared chromatin was then incubated with antibodies that were preincubated with the appropriate Dynabeads in 0.5% BSA (Gibco, 15260-037) in PBS as follows: a chromatin equivalent of 5×10^5 cells with anti-H3K4me1 (rabbit monoclonal, CST, #5326, 5 μl), anti-H3K9me2 (mouse monoclonal, MBL, MABI0317, 5 μl), anti-H3K27me3 (mouse monoclonal, MBL, MABI0323, 5 μl); 1×10^6 cells with anti-H3K4me3 (mouse monoclonal, MBL, MABI0304, 5 μl), anti-H3K9me3 (mouse monoclonal, MBL, MABI0318, 5 μl), anti-H3K36me2 (rabbit monoclonal, CST, #2901, 5 μl), anti-H2AK119ub1 (rabbit monoclonal, #8240, 10 μl), anti-H3K36me3 (rabbit polyclonal, Active Motif, 61101, 2 μl); 1.5×10^6 cells with anti-H3K27ac (mouse monoclonal, MBL, MABI0309, 5 μl); 2×10^6 cells with anti-CTCF (rabbit monoclonal, CST, #3418, 5 μl), anti-Laminb1 (rabbit polyclonal, Proteintech, 12987-1-AP, 10 μl); 4×10^6 cells with anti-Ring1b (rabbit monoclonal, CST, #5694, 10 μl); and 4.5×10^6 cells with anti-Rad21 (rabbit monoclonal, ab992, 5 μl).

After incubation for 6 h at 4°C, the beads were washed 4 times in wash buffer 1 (20 mM Tris-HCl pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% TritonX-100, 0.1% SDS), 2 times in wash buffer 2 (20 mM Tris-HCl pH 8.0, 2 mM EDTA, 500 mM NaCl, 1% TritonX-100, 0.1% SDS), and 2 times in wash buffer 3 (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 250 mM LiCl, 1% Na-Deoxycolate, 1% NP-40). Then, the washed beads were eluted in 10 mM Tris-HCl pH 8.0, 5 mM EDTA, 300 mM NaCl, and 1% SDS, and crosslinks were reversed overnight at 65°C. Input samples were treated in a similar manner. The following day, the IP and Input samples were incubated with RNaseA (Thermo Fisher Scientific, EN0531) and proteinase K (Thermo Fisher Scientific, AM2546). IP or Input DNA was purified using a QIA quick PCR purification kit (QIAGEN, 28104).

ChIP-seq libraries were prepared using a KAPA Hyper Prep Kit (KAPA, KK8504) following the manufacturer's guidelines. An adaptor kit (Fastgene, FG-NGSAD24) was used for the sample indexes. The average size and concentration of libraries were analyzed using LabChIP GX (PerkinElmer) and a KAPA library Quantification kit (KAPA, KK4824), respectively. Libraries were sequenced as 75 bp single-end reads on an Illumina NextSeq 500/550 platform with a NextSeq 500/550 High Output kit (75 cycles) (Illumina, 20024906).

ATAC-seq library preparation and sequencing

The ATAC-seq experiment was performed as described previously [124,125] with minor modifications. We used FACs-sorted viable cells for mESCs and EpiLCs; FACS-sorted BV-positive cells for d2 mPGCLCs, d4 mPGCLCs, and d4c7 mPGCLCs; and FACS-sorted AAG-positive cells for GSCs and GSCLCs. 50,000 cells were permeabilized in cold lysis buffer 1 (10 mM Tris-HCl pH8.0, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 0.1% Tween20, 0.1% Digitonin (Promega, G9441)) for 3 min followed by addition of 1 ml of cold lysis buffer 2 (10 mM Tris-HCl pH8.0, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween20). Nuclei were centrifuged and resuspended with 50 ml of transposase reaction mixture (25 ul of 2×TD buffer (Illumina, 20034197), 2.5 ml of Transposase (Illumina, 20034197), 16.5 ml of PBS, 0.5 ml of Digitonin, and 0.5 ml of Tween-20, 5 ul of DDW). After incubation at 37°C for 30 min, the tagged DNA was purified using a Minelute PCR purification kit (QIAGEN, 28004). The purified DNA was amplified for 8 cycles by a PCR reaction (NEB, M0541S) followed by size selection using AMPure XP beads (Corning, MAG-PCR-CL-250) to remove primer dimers. Libraries were sequenced as 2×75bp paired-end reads on an Illumina NextSeq 500/550 platform with a NextSeq 500/550 Mid Output Kit (150 cycles) (Illumina, 20024904) or NextSeq 500/550 High Output Kit (150 cycles, 20024907) (Illumina).

In situ Hi-C library preparation and sequencing

In situ Hi-C library preparation was performed as described previously [126,127] with minor modifications. We used the whole harvested cells for mESCs and EpiLCs; FACS-sorted BV-positive cells for d2 mPGCLCs and d4c7 mPGCLCs; and FACS-sorted AAG-positive cells for GSCs and GSCLCs. 2.5×10^6 cells were used for one replicate. The cells were fixed by 1% formaldehyde (Sigma, 252549)/HBSS and lysed in lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% NP-40) for 30 min on ice with frequent inversion. The cells were digested by 500 U of DpnII (NEB, R0543L) overnight at 37°C. Following biotin filling (Thermo Fisher Scientific, 19524-016; NEB, M0210S), proximity ligation (Thermo Fisher Scientific, 15224090) and reverse crosslinking, DNA was purified by ethanol precipitation and sheared to 200-400 bp fragments using a Covaris E220 sonicator (Covaris) at 4°C (10% Duty Factor, 200 cycles/burst, 175 W Peak Incident Power, 110 s). Ligation fragments containing biotin were immobilized on MyOne Streptavidin T1 beads (Thermo Fisher Scientific, 65001) followed by library preparation using a NEB library preparation kit (NEB, E7645S; NEB, E7335S) according to the manufacturer's guidelines. The libraries were amplified in 8 cycles and DNA fragments of 300–800 bp were selected using AMPure XP beads (Corning, MAG-PCR-CL-250). Libraries were sequenced as 2×100bp paired-end reads on an Illumina NovaSeq 6000 platform with a NovaSeq 6000 S1 Reagent Kit (200 cycles) (Illumina, 20012864).

NET-CAGE library preparation and sequencing

NET-CAGE library preparation was performed as described previously [128] with minor modifications. For extraction of nascent RNA, cells were first lysed with 1400 µl of Buffer A, which is Nuclei EZ Lysis Buffer (Sigma, NUC101-1KT) supplemented with 25 µM α-amanitin (Wako, 1022961), 1×cOmplete Protease Inhibitor Cocktail (Roche, 4693116001) and SUPERase•IN RNase Inhibitor (20 units; Thermo Fisher Scientific, AM2694), and then incubated on ice for 10 min and centrifuged at 800 g for 5 min at 4°C followed by washing once with the same buffer. Washed pellets were resuspended in 200 µl of Buffer B, containing 1% NP-40, 20 mM HEPES pH 7.5, 300 mM NaCl, 2 M urea, 0.2 mM EDTA, 1 mM dithiothreitol (DTT)

(Promega, P1171), 25 μM α-amanitin, 1×cOmplete Protease Inhibitor Cocktail and SUPERase•IN RNase Inhibitor (20 units), and incubated for 10 min on ice. The suspension was centrifuged at 3,000g for 2 min at 4°C. After removing the supernatant, the nuclear insoluble fraction was washed once with 100 μl of Buffer B. DNase I solution (50 μl) containing DNase I (10 units; Thermo Fisher Scientific, 89836), 1×DNase I Buffer (Thermo Fisher Scientific) and SUPERase•IN RNase Inhibitor (20 units) was added to the pellets. The samples were incubated for 30 min at 37°C while being pipetted up and down several times at 10-min intervals. QIAzol (700 μl) was then added and the solution was thoroughly mixed. RNA was extracted with an miRNeasy Mini kit (QIAGEN, 217004) according to the manufacturer's instructions. On-column DNase I digestion was carried out with an RNase-free DNase set (QIAGEN, 79254). RNA was eluted in 30 μl RNase-free water, and its quality and quantity were measured with a Qubit RNA HS assay kit (Thermo Fisher Scientific, Q32855) and 2100 BioAnalyzer (Agilent). cDNA was synthesized from 200 ng of nascent RNA. CAGE libraries were generated according to the no amplification non-tagging CAGE libraries for Illumina next-generation sequencers (nAnT-iCAGE) protocol [129] with PCR amplifications (Takara, R060A). All CAGE libraries were sequenced in 75 bp single-end reads on an Illumina NextSeq 500 platform.

ChIP-seq data processing

Single-end reads were processed using Trim-Galore! v0.4.1/cutadapt v1.9.1 [87,115] to remove adaptor sequences. The truncated reads were then aligned to (GRCm38p3) using Bowtie2 v2.3.4.1 [78] with the "-very-sensitive" option. Reads aligned to chromosomes 1 to 19, X, and Y were converted to the BAM format by SAMtools v1.7 [111]. BED files were obtained from the BAM files using the bamtobed command of BEDTools v2.29.2 [76]. BigWig files were generated from the BAM files using bamcoverage for raw count with the "--normalizeUsing CPM -bs 25" or bamcompare for IP/Input command with the "--pseudocount 1 -bs 1000" option of deepTools v3.5.0 [89] In both cases, the blacklist regions [130] were excluded.

The regions enriched by epigenetic marks were identified using peak calling tools. For CTCF peaks, MACS v2.1.1 [106] was used with the "-q 0.01 --nomodel --keep-dup all --extsize 200" option. For H3K9me3 domains, epic2 v0.0.41 [93] was used with "-kd -fdr 0.01" option. The number of IP or Input reads in 10/25/50/100 kb genomic windows were counted by the intersect command of BEDTools v2.29.2, and normalized by total million mapped reads (FPM) and transformed to Log2(IP/Input) for the downstream analysis. The bins in which no reads were detected in the Input samples were excluded.

ATAC-seq data processing

ATAC-seq data processing including public data was performed as previously described [125] with minor modifications. First, adaptor sequences were trimmed from the reads using TrimGalore! v0.4.1/cutadapt v1.9.1. These reads were aligned using Bowtie2 v2.3.4.1 to GRCm38p3 with the "--very-sensitive -X 2000" option. The properly mapped reads with the flag (99, 147, 83 or 163) were extracted by awk, and mitochondrial reads were excluded. Duplicated reads were removed using the MarkDuplicates command of Picard Tools v2.18.23 (https://broadinstitute.github.io/picard/). These de-duplicated reads were then filtered for high quality (MAPQ≧30). The reads with an insert size of less than 100 bp were extracted as nucleosome free region (NFR) reads. Bed files for downstream analysis were generated by the bamtobed command of BEDTools v2.29.2 with the "-bedpe" option. BigWig files were generated from the BAM files using bamcoverage for raw count with the "--normalizeUsing CPM -bs 25" option of deepTools v3.5.0. The blacklist regions (https://www.encodeproject.org/files/ENCFF999QPV/) were excluded.

Peak calling was performed using MACS v2.1.1 with the "--nomodel --shift -100 --extsize 200 --keep-dup all" option after shifting NFR reads with the offset by +4 bp in the + strand and by -5 bp in the - strand. Then, confident peak sets in each cell type were obtained by the IDR method (https://www.encodeproject.org/software/idr/) using two replicates.

PBAT data processing

Public read data processing of the methylation levels was performed as described previously [131]. In brief, all reads were processed with Trim-Galore! v0.4.1/cutadapt v1.9.1 with the "--clip_R1 4," "--trim1" and "-a AGATCGGAAGAGC" options. Output reads were mapped onto the mouse genome, GRCm38.p6, using Bismark v0.22.1 [77]/Bowtie2 v2.3.4.1 with the "--pbat" option. All public WGBS data were obtained from DDBJ or NCBI SRA ftp sites and processed as described above. Conversion rates were calculated as follows: output reads after Trim-Galore were mapped onto the lambda phage DNA sequence using Bismark v0.22.1/Bowtie2 v2.3.4.1 with the "--pbat" option. From the Bismark's statistics, conversion rates were determined as 1 - ([total mC counts] / [total C and mC counts]). All CpG sites with a read depth of between 4 and 200 were used for the %mC calculations.

3-prime RNA sequencing data processing

Raw 3' RNA-seq data were directly used with Salmon v1.4.0 [110] with default parameters and --noLengthCorrection to quantify the expression of GENCODE vM25 features on GRCm38.p6. Gene-level expression estimates were aggregated from transcript-level abundance using tximport v1.16.1 [116].

In situ Hi-C data processing

Sequences were first trimmed using fastp v0.21.0 [96] with default options and the --detect_adapter_for_pe flag. Trimmed sequences were then processed using HiCUP v0.8.0 [103] with default options and the di-tag length range set to 0–800, with bowtie v2.4.2 as the aligner. hicup2juicer was then used to produce pairs files, which were subsequently ingested with Juicer tools v1.22.01 [105] for the creation of .hic files. The same set of pairs files were also used to create multi-resolution cooler files using cooler v0.8.10 [84] with default options. Additionally, HiCSR commit b13ac41 [132] was used to de-noise 10 kb-resolution contact maps for visualization. In particular, pooled mESC data from [12] after 10× down-sampling were used for training with default parameters; inference was then performed using default parameters. FAN-C v0.9.13 [95] was finally used for the normalization (with default parameters) and subsequent visualization of the enhanced 10 kb matrices, including virtual 4C profiles.

NET-CAGE data processing

Sequences were first trimmed using fastp v0.21.0 and then aligned with STAR 2.7.6a [133] using default options. Uniquely mapped reads were converted to coverage bigWig tracks with G-bias correction using CAGEr v1.32.0 [80] with default options. Tag clusters were identified using CAGEfightR v1.7.6 [79] with pooledCutoff = 0.1 and mergeDist = 20 for unidirectional clusters as well as balanceThreshold = 0.8 for bidirectional clusters. These clusters were subsequently filtered to require at least 1 sample demonstrating an expression level exceeding 1 TPM. Unidirectional clusters (putative promoters) were removed if they overlapped bidirectional clusters (putative enhancers), and the two region sets were subsequently combined to identify coordinately regulation enhancer-promoter co-transcription across stages. In particular, Kendall correlation was used to find putative enhancers within 1 mb of putative promoters that exhibited correlated expression patterns, with TPM as the expression unit.

Global Hi-C metrics

HiCRes v1.1 [101] with default parameters was used for the resolution of contact maps following the definition in [127]. Matrix similarity scores were computed using HiCRep.py v0.2.3 [100] with --binSize=50000 --dBPMax=5000000 --h=3. Contact probability decay (i.e., the average contact frequency across different genomic separation distances) was assessed using the compute-expected and logbin-expected modules from cooltools v0.4.0 [86] at all resolutions, in both cis and trans. 3D models of individual chromosomes were produced using CSynth commit 26e21fb [37] with balanced 50 kb cis matrices, whose coordinates are normalized to achieve unit backbone length (i.e., the sum of Euclidean distance between adjacent beads being 1); and the size of these predicted structures are taken to be the volume of their 3D convex hulls.

Compartment-related analysis

For analyses involving data across multiple studies, eigendecomposition was performed at 100 kb resolution using the call-compartments module from cooltools v0.4.0 with GC content for orientating the track sign to achieve a positive correlation. For analyses strictly focusing on data generated within this study, dcHiC commit 7b1727f [88] was used with default parameters to perform simultaneous compartment score calculation across all samples at 50 kb resolution to facilitate statistical comparison across cell types while integrating replicate data. Though the values produced by dcHiC showed high correlation with those generated by cooltools, dcHiC was not applied to public datasets due to a lack of replication in certain datasets. Quantile-binned saddle plots were produced using dcHiC-generated compartment scores and the outputs of compute-expected described above at 50 kb resolution. Binarization of compartment score tracks was carried out using A := score > 0 and B := score < 0. PCA of compartment scores to contrast lineages was done using 100 kb resolution data and bins non-masked in all samples. The average size of compartments was assessed using an auto-correlation function, where the signal profile is shifted and correlated against the original, using the acf function from R library stats 4.0.3 with na.action = na.pass.

Subcompartment-related analysis

8-state subcompartment labels were assigned to 50 kb bins with balanced contact frequencies using CALDER commit 32220e8 [67]. The strength of epigenetic signals in each subcompartment was subsequently examined by converting enrichment values to Z-scores genome-wide, after which the average across all bins with the same label was computed. Significant differences in subcompartment proportions were evaluated using the prop.test function from R library stats 4.0.3.

TAD-related analysis

Insulation scores were computed at 10 kb resolution with a window size of 100 kb using the diamond-insulation module of cooltools v0.4.0. Consensus TADs in each dataset were derived by taking the set of bins with boundary prominence scores >0.2 in at least half the cell types present and subsequently pairing neighboring boundaries, with those exceedingly 2mb filtered out, consistently yielding ~4000–5000 domains for each dataset. The significance and strength of TAD-TAD interactions were evaluated using a non-central hypergeometric (NCHG) test implemented as a part of the Chrom3D pipeline [81]. Biological replicates (the two deepest ones in case there were more than two) were then used to identify highly reproducible TAD-TAD interactions using IDR2D v1.4.0 [104] with default parameters. In particular, TAD-TAD interactions with NCHG p-value > 0.01 were first filtered out, and then the odds ratio was used as the ranking statistic for IDR analysis, with the final filter criteria being IDR p-value < 0.01. Treating significant TAD-TAD interactions as edges of a graph, cliques were identified using the max_cliques function from R library igraph v1.2.6 [134]. The over-representation of A-A vs B-B clique interactions was compared against an expected value based on the proportion of A vs B TADs across all TADs, with the identity of compartment assignment of TADs based on having more 25 kb bins labelled as one compartment versus the other. Confidence intervals were derived from bootstrapping the set of clique interactions. The degree of TAD boundary conservation was evaluated using a permutation test, where the number of boundaries being shared across cell types was compared against a background derived from merging the list of boundaries and shuffling cell type labels. Additionally, 9 other TAD identification algorithms [83,99,102,107,108,112-114,127] were used with default parameters to validate trends observed with insulation scores, all at 50 kb resolution.

Histone mass spectrometry analysis

Single histone modification abundances are summed from their individual occurrences as well as co-occurrences (e.g., H3K27me3 = H3K27me3 + H3K27me3&H3K36me1 + H3K27me3&H3K36me2 + H3K27me3&H3K36me3). PCA of these relative abundance measures for all quantifiable H3 modifications (at least one sample exhibiting abundance >0.1%) were used as input for PCA using the prcomp function from the R library stats v4.0.3 with default parameters to assess epigenome-wide tendencies. Abundance measures were further Z-score transformed for hierarchical clustering using the hclust function from R library stats v4.0.3 with default parameters.

Normalization of epigenetic signals

Histone mass spectrometry-derived abundances were used to scale corresponding ChIP-seq tracks by directly multiplying the library-size normalized (counts/million mapped reads) values with the relative abundance. For targets lacking mass spectrometry data (e.g., transcription factors), we applied S3V2-IDEAS commit b7cc2d5 [109] to derive scaling factors using default parameters at a bin size of 200 bp.

ATAC-seq analysis

The union set of peaks across all cell types was taken as features against which reads were counted, and the resulting count matrix was further normalized via FPKM to account for variations in peak widths and sequence depth. PCA was then performed on the 10000 most variable peaks to assess global accessome trends. The 2000 most variable peaks were additionally clustered by using the hclust function from R library stats v4.0.3 with default parameters; visual inspection of the resulting dendrogram suggested 7 as a reasonable number of clusters for cutting. Global openness was assessed by first fitting a two-component gaussian mixture model to the log2(FPKM + 1) distribution across the union peak set and then assessing the number of sites exceeding the higher component's mean versus those below the lower component's mean.

Motif enrichment analysis

Over-representation of known transcription factor motifs was assessed in an ensemble manner by combining multiple frameworks (e.g., HOMER, MEME) as implemented in GimmeMotifs v0.15.3 [97] using default options. Differential enrichment of motifs between different region sets (e.g., open sites with distinct chromatin states) was examined using the maelstrom module of GimmeMotifs with default options.

Enhancer-promoter pairing

Cis-regulatory elements were associated with putative target genes using "activity-by-contact" (ABC) commit 7fd69b0 [43]. KR-normalized matrices at 5 kb resolution were combined with H3K27ac and ATAC-seq data to calculate ABC scores quantile-normalized to K562 data, after which a stringent cut-off of 0.02 was applied — corresponding to 70% recall and 60% precision based on previous CRISPRi-FlowFISH validation [43]. Alternatively, enhancer-promoter pairs identified based on co-regulated NET-CAGE tag clusters, as described above, were assessed for their degree of coordination. Specifically, a permutation test was used to compare the number of co-expressed (>1 TPM in a specific cell type) enhancer-promoter pairs versus that of background sets generated by sampling from all tag clusters. Differential interactions between enhancer-promoter pairs identified by ABC scores were investigated using R library HiCDCPlus v0.99.12 [135] using default parameters at 10 kb resolution. The degree of coordinated differential promoter interaction and differential expression was quantified through the application of RRHO2 v1.0 [136] to gene lists ranked by DESeq2 test statistics; for promoters involved in multiple ABC E-P pairs, the mean test statistic was used for ranking.

ChIP-seq analysis

The domain size distributions of histone modifications were determined using MCORE [137] with the maximum shift size set to the chromosome lengths and other parameters kept at their defaults. Resulting cross-correlation values between replicates were averaged using a cubic spline via the function smooth.spline from R library stats v4.0.3 with default parameters, after which Gardner transformations were applied to decompose the decay spectrum into component exponential functions corresponding to different domain sizes and quantify their contribution. Differential ChIP-seq analysis was performed using DiffBind [91] for targets with narrow signals and csaw for broad ones. DiffBind v3.0.13 was applied with union peak sets resized to 500 bp around the summits of MACS peak calls and other options kept at their defaults using both edgeR [138] and DESeq2 [90] for the underlying statistical framework, after which only concordant results were retained (e.g., up-regulated with both methods). Unless otherwise stated, "constitutive"/"conserved" peaks refer to the intersection of MACS peak calls between cell types. csaw v1.24.3 [139] was applied with default settings with edgeR as the underlying statistical framework at both a coarse (2 kbp windows with a 500 bp step size for H3K27me3; 10 kbp windows with a 2 kbp step size for H3K9me2) and a fine resolution (500 bp windows with a 100 bp step size for H3K27me3; 1 kb windows with a 200 bp step size for H3K9me2), after which the results were consolidated, allowing for a gap size of 100 bp. The domain expansion/contraction kinetics were characterized using ChromTime commit a332dbb [140] with default settings in broad mode, with a post-hoc filter applied to exclude regions <10 kb. Aggregate plots were generated using the module computeMatrix from deepTools v3.5.0 with default options, in scale-regions mode for domains and reference-point mode for focal features such as peaks. Differential H3K9me3 promoters (+/- 1kb from TSS) were defined using the mass spectrometry-derived coefficient-normalized log2-transformed FPKM signal with the threshold (log2(FPKM) >1 in either cell type and log2(FPKM) difference >1).

Epigenome-based clustering of cis-regulatory elements

The log2(enrichment over input) values of ChIP-seq signals and log2(FPKM + 1) for ATAC-seq signals in promoters (+/- 2.5kb from TSS) or reproducible accessible sites identified using ChromA v2.1.1 [82] (resized to +/- 500 bp surrounding the summit) were used as input for dimension reduction through UMAP v0.5.1 [117] and subsequently clustered through HDBSCAN v0.8.27 [98]. For UMAP, manhattan distances were used for promoters and correlation distances for open sites; a grid search over min_dist of [0.0, 0.01, 0.1], n_neighbors of [15, 30, 50] and n_components of 2–10 were all subjected to HDBSCAN clustering to identify epigenetically distinct clusters via visual inspection. For HDBSCAN, a grid search over min_cluster_size and min_samples over [50, 100, 200, 500, 1000, 2000, 5000, 10000] were tested. In a semi-supervised fashion, individual clusters were isolated and subjected to further sub-clustering until the embedding no longer exhibited distinct segregation of data points for any individual epigenetic signal.

Pathway enrichment analysis

Associations of specific gene lists with particular biological pathways were evaluated using the gost function from R library gprofiler2 v0.2.0 [141] with default options. The enrichment of pathways towards the extremes of ranked gene lists, on the other hand, was assessed using the fgseaMultilevel function from R library fgsea 1.17.1 [142] with the boundary parameter eps set to 0 and others kept at their default values; redundant terms were collapsed by using collapsePathways with an adjusted p-value threshold of 0.05. To obtain gene lists ranked by multiple metrics (e.g., differential expression and promoter interaction), the mean test statistic was used to rank genes independently for each metric, and an aggregated ranking was then obtained using p-values produced by the aggregateRanks function from R library RobustRankAggreg v1.1 [143].

Overlap enrichment analysis

The overlap between genomic regions and annotated intervals was examined using Fisher's exact tests as implemented in the R library LOLA v1.19.1 [144]. Ensembl Regulatory build annotations v20180516 were sourced directly from Ensembl; RepeatMasker annotations were obtained from the rmsk table hosted on the UCSC Genome Browser. ENCODE cCRE annotations were downloaded from SCREEN v13 (http://screen.encodeproject.org/).

Pile-up analysis

Interaction between specific regions (e.g., promoters of a similar chromatin state) were quantified using the ObsExpSnipper function from cooltools v0.4.0 with default parameters and using the aforementioned diagonal-wise expected values. For pile-up of domains (e.g., TADs or broad H3K9me3 domains) rescaled to the same size, coolpup.py v0.9.7 [85] was used with the option --rescale and optionally --local when assessing on-diagonal patterns, and with all other options kept at their defaults.

Lamin B1-related analysis

EDD v1.1.19 [92] was used to identify lamina-associated domains from lamin B1 ChIP-seq with a bin size of 10 kb, gap penalty set to 20, and all others options kept at their defaults. LADetector v8122016 [55] was used instead for lamin B1 DamID, with a bin size of 10 kb and max dip size of 25 kb. Generalized linear models with 50 basis functions were used to visualize chromosome-scale patterns using REML for smoothness selection as implemented in the gam function of R library mgcv v1.8-31 [145].

Partially methylated domains-related analysis

PMDs were identified by calculating median mCG/CG values using a 100 kb sliding window and identifying those falling below 85%; after merging adjacent regions, those wider than 500 kb were called as PMDs. The binary status of whether a bin falls within a GSC PMD or not was modelled using three methods: (1) gradient boosted tree (gbm), (2) neural network (nnet), and (3) elastic net (glmnet), each with 10x10 cross validation using a 70/30 train/test split as implemented in the R library caret v6.0-86 [146]. Model performance for predicting PMDs was then assessed on the held-out test set using the roc function from R library pROC v1.16.2 [147].

Mapping to the Y chromosome

Ampliconic sequences on the murine Y chromosome were retrieved from an earlier report describing its assembly [66], and were directly used as the reference for alignment. Otherwise, data was processed as described in “PBAT data processing”.

Statistical considerations

P-values were mapped to symbols as follows: 0 (****) 0.0001 (***) 0.001 (**) 0.01 (*) 0.05 (ns) 1. Wilcoxon rank-sum tests and T-tests were carried out using the functions wilcox.test and t.test, respectively, from the R library stats v4.0.3. Bootstrap confidence intervals were computed using the function boot with 100000 replicates followed by boot.ci from the R library boot 1.3-28 [148] using default options. For all box plots (i.e., box-and-whiskers plots), the lower and upper hinge correspond to the first and third quartile, and the upper whiskers extend to the largest value % 1.5 * IQR and vice versa for the lower whiskers.

Data Availability

The accession number for all the sequencing data generated in this study is GSE183828 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183828) (the GEO database). Scripts used to generate the presented results and additional raw data underlying figures are available at https://github.com/bhu/germ_nucleome.

References

1 Gurdon, J. B. & Wilmut, I. Nuclear transfer to eggs and oocytes. *Cold Spring Harb Perspect Biol* **3**, doi:10.1101/cshperspect.a002659 (2011).

2 Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* **20**, 535-550, doi:10.1038/s41580-019-0132-4 (2019).

3 Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114, doi:10.1038/nature21711 (2017).

4 Du, Z. *et al.* Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235, doi:10.1038/nature23263 (2017).

5 Ke, Y. *et al.* 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381 e320, doi:10.1016/j.cell.2017.06.029 (2017).

6 Battulin, N. *et al.* Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol* **16**, 77, doi:10.1186/s13059-015-0642-0 (2015).

7 Wang, Y. *et al.* Reprogramming of Meiotic Chromatin Architecture during Spermatogenesis. *Mol Cell* **73**, 547-561 e546, doi:10.1016/j.molcel.2018.11.019 (2019).

8 Patel, L. *et al.* Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat Struct Mol Biol* **26**, 164-174, doi:10.1038/s41594-019-0187-0 (2019).

9 Alavattam, K. G. *et al.* Attenuated chromatin compartmentalization in meiosis and its maturation in sperm development. *Nat Struct Mol Biol* **26**, 175-184, doi:10.1038/s41594-019-0189-y (2019).

10 Vara, C. *et al.* Three-Dimensional Genomic Structure and Cohesin Occupancy Correlate with Transcriptional Activity during Spermatogenesis. *Cell reports* **28**, 352-367 e359, doi:10.1016/j.celrep.2019.06.037 (2019).

11 Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219-226, doi:10.1038/nature23884 (2017).

12 Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572 e524, doi:10.1016/j.cell.2017.09.043 (2017).

13 Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat Genet* **50**, 238-249, doi:10.1038/s41588-017-0030-7 (2018).

14 Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, 1380-1388, doi:10.1038/s41588-019-0479-7 (2019).

15 Saitou, M. & Hayashi, K. Mammalian in vitro gametogenesis. *Science* **374**, eaaz6830, doi:10.1126/science.aaz6830 (2021).

16 Griswold, M. D. Spermatogenesis: The Commitment to Meiosis. *Physiol Rev* **96**, 1-17,

doi:10.1152/physrev.00013.2015 (2016).

17 Spiller, C., Koopman, P. & Bowles, J. Sex Determination in the Mammalian Germline. *Annu Rev Genet* **51**, 265-285, doi:10.1146/annurev-genet-120215-035449 (2017).

18 Wen, L. & Tang, F. Human Germline Cell Development: from the Perspective of Single-Cell Sequencing. *Mol Cell* **76**, 320-328, doi:10.1016/j.molcel.2019.08.025 (2019).

19 Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710-719, doi:10.1016/j.stem.2014.05.008 (2014).

20 Tang, W. W., Kobayashi, T., Irie, N., Dietmann, S. & Surani, M. A. Specification and epigenetic programming of the human germ line. *Nat Rev Genet* **17**, 585-600, doi:10.1038/nrg.2016.88 (2016).

21 Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519-532, doi:S0092-8674(11)00771-9 [pii]

10.1016/j.cell.2011.06.052 (2011).

22 Ohta, H. *et al.* In vitro expansion of mouse primordial germ cell-like cells recapitulates an epigenetic blank slate. *EMBO J* **36**, 1888-1907, doi:10.15252/embj.201695862 (2017).

23 Ohta, H. *et al.* Cyclosporin A and FGF signaling support the proliferation/survival of mouse primordial germ cell-like cells in vitrodagger. *Biol Reprod* **104**, 344-360, doi:10.1093/biolre/ioaa195 (2021).

24 Kanatsu-Shinohara, M. *et al.* Long-term proliferation in culture and germline transmission of mouse male germline stem cells. *Biol Reprod* **69**, 612-616, doi:10.1095/biolreprod.103.017012

biolreprod.103.017012 [pii] (2003).

25 Ishikura, Y. *et al.* In Vitro Derivation and Propagation of Spermatogonial Stem Cell Activity from Mouse Pluripotent Stem Cells. *Cell reports* **17**, 2789-2804, doi:10.1016/j.celrep.2016.11.026 (2016).

26 Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519-523, doi:nature06968 [pii]

10.1038/nature06968 (2008).

27 Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590-604, doi:S0092-8674(12)00409-6 [pii]

10.1016/j.cell.2012.03.026 (2012).

28 Boroviak, T., Loos, R., Bertone, P., Smith, A. & Nichols, J. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat Cell Biol* **16**, 516-528, doi:10.1038/ncb2965 (2014).

29 Kurimoto, K. *et al.* Quantitative Dynamics of Chromatin Remodeling during Germ Cell Specification from Mouse Embryonic Stem Cells. *Cell Stem Cell* **16**, 517-532, doi:10.1016/j.stem.2015.03.002 (2015).

30 Jameson, S. A. *et al.* Temporal transcriptional profiling of somatic and germ cells reveals biased lineage priming of sexual fate in the fetal mouse gonad. *PLoS Genet* **8**, e1002575, doi:10.1371/journal.pgen.1002575

PGENETICS-D-11-02591 [pii] (2012).

31 Evans, E. P., Ford, C. E. & Lyon, M. F. Direct evidence of the capacity of the XY germ cell in the mouse to become an oocyte. *Nature* **267**, 430-431 (1977).

32 Miyauchi, H. *et al.* Bone morphogenetic protein and retinoic acid synergistically specify female germ-cell fate in mice. *EMBO J* **36**, 3100-3119, doi:10.15252/embj.201796875 (2017).

33 Nagaoka, S., I. *et al.* ZGLP1 is a determinant for the oogenic fate in mice. *Science* **367**, doi:10.1126/science.aaw4115 (2020).

34 Guenatri, M., Bailly, D., Maison, C. & Almouzni, G. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J Cell Biol* **166**, 493-505, doi:10.1083/jcb.200403109 (2004).

35 Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128**, 747-762 (2007).

36 Paulsen, J. *et al.* Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat Genet* **51**, 835-843, doi:10.1038/s41588-019-0392-0 (2019).

37 Todd, S. *et al.* CSynth: an interactive modelling and visualization tool for 3D chromatin structure. *Bioinformatics* **37**, 951-955, doi:10.1093/bioinformatics/btaa757 (2021).

38 Du, Z. *et al.* Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Mol Cell* **77**, 825-839 e827, doi:10.1016/j.molcel.2019.11.011 (2020).

39 Luo, Z. *et al.* Reorganized 3D Genome Structures Support Transcriptional Regulation in Mouse Spermatogenesis. *iScience* **23**, 101034, doi:10.1016/j.isci.2020.101034 (2020).

40 Farhangdoost, N. *et al.* Chromatin dysregulation associated with NSD1 mutation in head and neck squamous cell carcinoma. *Cell reports* **34**, 108769, doi:10.1016/j.celrep.2021.108769 (2021).

41 Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744-751, doi:10.1038/s41586-020-2093-3 (2020).

42 Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell* **48**, 849-862, doi:10.1016/j.molcel.2012.11.001 (2012).

43 Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664-1669, doi:10.1038/s41588-019-0538-0 (2019).

44 Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56, doi:10.1038/nature24281 (2017).

45 D'Oliveira Albanus, R. *et al.* Chromatin information content landscapes inform transcription factor

and DNA interactions. *Nat Commun* **12**, 1307, doi:10.1038/s41467-021-21534-4 (2021).

46 Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet* **42**, 1093-1100, doi:ng.708 [pii]
10.1038/ng.708 (2010).

47 Western, P. S., Miles, D. C., van den Bergen, J. A., Burton, M. & Sinclair, A. H. Dynamic regulation of mitotic arrest in fetal male germ cells. *Stem Cells* **26**, 339-347 (2008).

48 Polovnikov, K., Belan, S., Imakaev, M., Brand, H. B. & A., M. L. Fractal polymer with loops recapitulates key features of chromosome organization. *bioRxiv*, doi:https://doi.org/10.1101/2022.02.01.478588 (2022).

49 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951, doi:10.1038/nature06947 (2008).

50 Poleshko, A. *et al.* Genome-Nuclear Lamina Interactions Regulate Cardiac Stem Cell Lineage Restriction. *Cell* **171**, 573-587 e514, doi:10.1016/j.cell.2017.09.018 (2017).

51 Yattah, C. *et al.* Dynamic Lamin B1-Gene Association During Oligodendrocyte Progenitor Differentiation. *Neurochem Res* **45**, 606-619, doi:10.1007/s11064-019-02941-y (2020).

52 Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-613, doi:10.1016/j.molcel.2010.03.016 (2010).

53 Robson, M. I. *et al.* Tissue-Specific Gene Repositioning by Muscle Nuclear Membrane Proteins Enhances Repression of Critical Developmental Genes during Myogenesis. *Mol Cell* **62**, 834-847, doi:10.1016/j.molcel.2016.04.035 (2016).

54 Chen, X. *et al.* The visualization of large organized chromatin domains enriched in the H3K9me2 mark within a single chromosome in a single cell. *Epigenetics* **9**, 1439-1445, doi:10.4161/15592294.2014.971633 (2014).

55 Harr, J. C. *et al.* Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. *J Cell Biol* **208**, 33-52, doi:10.1083/jcb.201405110 (2015).

56 Bian, Q., Khanna, N., Alvikas, J. & Belmont, A. S. beta-Globin cis-elements determine differential nuclear targeting through epigenetic modifications. *J Cell Biol* **203**, 767-783, doi:10.1083/jcb.201305027 (2013).

57 Fukuda, K. *et al.* Regulation of mammalian 3D genome organization and histone H3K9 dimethylation by H3K9 methyltransferases. *Commun Biol* **4**, 571, doi:10.1038/s42003-021-02089-y (2021).

58 Mochizuki, K. *et al.* Repression of germline genes by PRC1.6 and SETDB1 in the early embryo precedes DNA methylation-mediated silencing. *Nat Commun* **12**, 7020, doi:10.1038/s41467-021-27345-x (2021).

59 Kubo, N. *et al.* DNA methylation and gene expression dynamics during spermatogonial stem cell differentiation in the early postnatal mouse testis. *BMC Genomics* **16**, 624, doi:10.1186/s12864-015-

1833-5 (2015).

60 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322, doi:nature08514 [pii]

10.1038/nature08514 (2009).

61 Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-775, doi:10.1038/ng.865 (2011).

62 Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res* **22**, 246-258, doi:10.1101/gr.125872.111 (2012).

63 Schroeder, D. I. *et al.* The human placenta methylome. *Proc Natl Acad Sci U S A* **110**, 6037-6042, doi:10.1073/pnas.1215145110 (2013).

64 Salhab, A. *et al.* A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol* **19**, 150, doi:10.1186/s13059-018-1510-5 (2018).

65 Shirane, K., Miura, F., Ito, T. & Lorincz, M. C. NSD1-deposited H3K36me2 directs de novo methylation in the mouse male germline and counteracts Polycomb-associated silencing. *Nat Genet* **52**, 1088-1098, doi:10.1038/s41588-020-0689-z (2020).

66 Soh, Y. Q. *et al.* Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800-813, doi:10.1016/j.cell.2014.09.052 (2014).

67 Liu, Y. *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat Commun* **12**, 2439, doi:10.1038/s41467-021-22666-3 (2021).

68 Johnstone, S. E. *et al.* Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer. *Cell* **182**, 1474-1489 e1423, doi:10.1016/j.cell.2020.07.030 (2020).

69 Kanatsu-Shinohara, M. *et al.* Generation of pluripotent stem cells from neonatal mouse testis. *Cell* **119**, 1001-1012 (2004).

70 Poleshko, A. *et al.* H3K9me2 orchestrates inheritance of spatial positioning of peripheral heterochromatin through mitosis. *Elife* **8**, doi:10.7554/eLife.49278 (2019).

71 Sanulli, S. *et al.* HP1 reshapes nucleosome core to promote phase separation of heterochromatin. *Nature* **575**, 390-394, doi:10.1038/s41586-019-1669-2 (2019).

72 Keough, K. C. *et al.* An atlas of lamina-associated chromatin across twelve human cell types reveals an intermediate chromatin subtype. *bioRxiv*, doi:https://www.biorxiv.org/content/10.1101/2020.07.23.218768v3.full (2021).

73 Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36**, 3573-3599, doi:10.15252/embj.201798004 (2017).

74 Cuadrado, A. *et al.* Specific Contributions of Cohesin-SA1 and Cohesin-SA2 to TADs and Polycomb

Domains in Embryonic Stem Cells. *Cell reports* **27**, 3500-3510 e3504, doi:10.1016/j.celrep.2019.05.078 (2019).

75 Ishikura, Y. *et al.* In vitro reconstitution of the whole male germ-cell development from mouse pluripotent stem cells. *Cell Stem Cell* **28**, 2167-2179 e2169, doi:10.1016/j.stem.2021.08.005 (2021).

76 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

77 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572, doi:10.1093/bioinformatics/btr167 (2011).

78 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

79 Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R. & Sandelin, A. CAGEfightR: analysis of 5'-end data using R/Bioconductor. *BMC Bioinformatics* **20**, 487, doi:10.1186/s12859-019-3029-5 (2019).

80 Haberle, V., Forrest, A. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res* **43**, e51, doi:10.1093/nar/gkv054 (2015).

81 Paulsen, J. *et al.* Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol* **18**, 21, doi:10.1186/s13059-016-1146-2 (2017).

82 Gabitto, M. I. *et al.* Characterizing chromatin landscape from aggregate and single-cell genomic assays using flexible duration modeling. *Nat Commun* **11**, 747, doi:10.1038/s41467-020-14497-5 (2020).

83 Matthey-Doret, C. *et al.* Computer vision for pattern detection in chromosome contact maps. *Nat Commun* **11**, 5795, doi:10.1038/s41467-020-19562-7 (2020).

84 Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311-316, doi:10.1093/bioinformatics/btz540 (2020).

85 Flyamer, I. M., Illingworth, R. S. & Bickmore, W. A. Coolpup.py: versatile pile-up analysis of Hi-C data. *Bioinformatics* **36**, 2980-2985, doi:10.1093/bioinformatics/btaa073 (2020).

86 Venev, S. *et al.* open2c/cooltools: v0.4.1. Zenodo. (2021).

87 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

88 Wang, J., Chakraborty, A. & Ay, F. dcHiC: differential compartment analysis of Hi-C datasets. *bioRxiv*, https://doi.org/10.1101/2021.1102.1102.429297 (2021).

89 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

90 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

91 Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393, doi:10.1038/nature10730 (2012).

92 Lund, E., Oldenburg, A. R. & Collas, P. Enriched domain detector: a program for detection of wide genomic enrichment domains robust against local variations. *Nucleic Acids Res* **42**, e92, doi:10.1093/nar/gku324 (2014).

93 Stovner, E. B. & Saetrom, P. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392-4393, doi:10.1093/bioinformatics/btz232 (2019).

94 Yuan, Z. F. *et al.* EpiProfile 2.0: A Computational Platform for Processing Epi-Proteomics Mass Spectrometry Data. *J Proteome Res* **17**, 2533-2541, doi:10.1021/acs.jproteome.8b00133 (2018).

95 Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol* **21**, 303, doi:10.1186/s13059-020-02215-9 (2020).

96 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

97 Bruse, N. & Heeringen, S. J. V. GimmeMotifs: an analysis framework for transcription factor motif analysis. *bioRxiv*, https://doi.org/10.1101/474403 (2018).

98 Campello, R., Moulavi, D. & Sander, J. in *Advances in Knowledge Discovery and Data Mining* 160-172 (Springer, 2013).

99 Xing, H., Wu, Y., Zhang, M. Q. & Chen, Y. Deciphering hierarchical organization of topologically associated domains through change-point testing. *BMC Bioinformatics* **22**, 183, doi:10.1186/s12859-021-04113-8 (2021).

100 Lin, D., Sanders, J. & Noble, W. S. HiCRep.py : Fast comparison of Hi-C contact matrices in Python. *Bioinformatics*, doi:10.1093/bioinformatics/btab097 (2021).

101 Marchal, C., Singh, N., Corso-Díaz, X. & Swaroop, A. HiCRes: a computational method to estimate and predict the resolution of HiC libraries. *bioRxiv*, https://doi.org/10.1101/2020.1109.1122.307967 (2020).

102 Levy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386-392, doi:10.1093/bioinformatics/btu443 (2014).

103 Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310, doi:10.12688/f1000research.7334.1 (2015).

104 Krismer, K., Guo, Y. & Gifford, D. K. IDR2D identifies reproducible genomic interactions. *Nucleic Acids Res* **48**, e31, doi:10.1093/nar/gkaa030 (2020).

105 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

106 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-

2008-9-9-r137 (2008).

107 An, L. *et al.* OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol* **20**, 282, doi:10.1186/s13059-019-1893-y (2019).

108 Dali, R., Bourque, G. & Blanchette, M. RobusTAD: A Tool for Robust Annotation of Topologically Associating Domain Boundaries. *bioRxiv*, https://doi.org/10.1101/293175 (2018).

109 Xiang, G., Giardine, B. M., Mahony, S., Zhang, Y. & Hardison, R. C. S3V2-IDEAS: a package for normalizing, denoising and integrating epigenomic datasets across different cell types. *Bioinformatics*, doi:10.1093/bioinformatics/btab148 (2021).

110 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

111 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

112 Cresswell, K. G., Stansfield, J. C. & Dozmorov, M. G. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics* **21**, 319, doi:10.1186/s12859-020-03652-w (2020).

113 Soler-Vila, P., Cusco, P., Farabella, I., Di Stefano, M. & Marti-Renom, M. A. Hierarchical chromatin organization detected by TADpole. *Nucleic Acids Res* **48**, e39, doi:10.1093/nar/gkaa087 (2020).

114 Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44**, e70, doi:10.1093/nar/gkv1505 (2016).

115 Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7. Zenodo. (2021).

116 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

117 Mcinnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

118 Ohta, H., Yomogida, K., Yamada, S., Okabe, M. & Nishimune, Y. Real-time observation of transplanted 'green germ cells': proliferation and differentiation of stem cells. *Dev Growth Differ* **42**, 105-112 (2000).

119 Anton, T., Bultmann, S., Leonhardt, H. & Markaki, Y. Visualization of specific DNA sequences in living mouse embryonic stem cells with a programmable fluorescent CRISPR/Cas system. *Nucleus* **5**, 163-172, doi:10.4161/nucl.28488 (2014).

120 Okamoto, I. *et al.* Evidence for de novo imprinted X-chromosome inactivation independent of meiotic inactivation in mice. *Nature* **438**, 369-373, doi:10.1038/nature04155 (2005).

121 Miura, K. in *Bioimage Data Analysis Workflows* 9-32 (Springer International Publishing, 2020).

122 Sidoli, S., Bhanu, N. V., Karch, K. R., Wang, X. & Garcia, B. A. Complete Workflow for Analysis of

Histone Post-translational Modifications Using Bottom-up Mass Spectrometry: From Histone Extraction to Data Analysis. *Journal of Visualized Experiments*, doi:10.3791/54112 (2016).

123 Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature protocols* **1**, 729-748, doi:10.1038/nprot.2006.98 (2006).

124 Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959-962, doi:10.1038/nmeth.4396 (2017).

125 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

126 Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56-65, doi:10.1016/j.ymeth.2017.04.004 (2017).

127 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

128 Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat Genet* **51**, 1369-1379, doi:10.1038/s41588-019-0485-9 (2019).

129 Murata, M. *et al.* Detecting expressed genes using CAGE. *Methods Mol Biol* **1164**, 67-85, doi:10.1007/978-1-4939-0805-9_7 (2014).

130 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific reports* **9**, 9354, doi:10.1038/s41598-019-45839-z (2019).

131 Shirane, K. *et al.* Global Landscape and Regulatory Principles of DNA Methylation Reprogramming for Germ Cell Specification by Mouse Pluripotent Stem Cells. *Dev Cell* **39**, 87-103, doi:10.1016/j.devcel.2016.08.008 (2016).

132 Dimmick, M. C., Lee, L. J. & Frey, B. J. HiCSR: a Hi-C super-resolution framework for producing highly realistic contact maps. *Bioinformatics*, doi:https://doi.org/10.1101/2020.02.24.961714.

133 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

134 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695** (2006).

135 Sahin, M. *et al.* HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. *Nat Commun* **12**, 3366, doi:10.1038/s41467-021-23749-x (2021).

136 Cahill, K. M., Huo, Z., Tseng, G. C., Logan, R. W. & Seney, M. L. Improved identification of concordant and discordant gene expression signatures using an updated rank-rank hypergeometric overlap approach. *Scientific reports* **8**, 9588, doi:10.1038/s41598-018-27903-2 (2018).

137 Molitor, J., Mallm, J. P., Rippe, K. & Erdel, F. Retrieving Chromatin Patterns from Deep Sequencing

Data Using Correlation Functions. *Biophys J* **112**, 473-490, doi:10.1016/j.bpj.2017.01.001 (2017).

138 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

139 Lun, A. T. & Smyth, G. K. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res* **44**, e45, doi:10.1093/nar/gkv1191 (2016).

140 Fiziev, P. & Ernst, J. ChromTime: modeling spatio-temporal dynamics of chromatin marks. *Genome Biol* **19**, 109, doi:10.1186/s13059-018-1485-2 (2018).

141 Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res* **9**, doi:10.12688/f1000research.24956.2 (2020).

142 Korotkevich, G. *et al.* Fast gene set enrichment analysis. doi:https://doi.org/10.1101/060012.

143 Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573-580, doi:10.1093/bioinformatics/btr709 (2012).

144 Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-589, doi:10.1093/bioinformatics/btv612 (2016).

145 Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Statist Soc B* **73**, 3-36, doi:https://doi.org/10.1111/j.1467-9868.2010.00749.x (2011).

146 Kuhn, M. Building Predictive Models in R Using the caret Package. *J Stat Soft* **28**, 1-26, doi:https://doi.org/10.18637/jss.v028.i05 (2008).

147 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77, doi:10.1186/1471-2105-12-77 (2011).

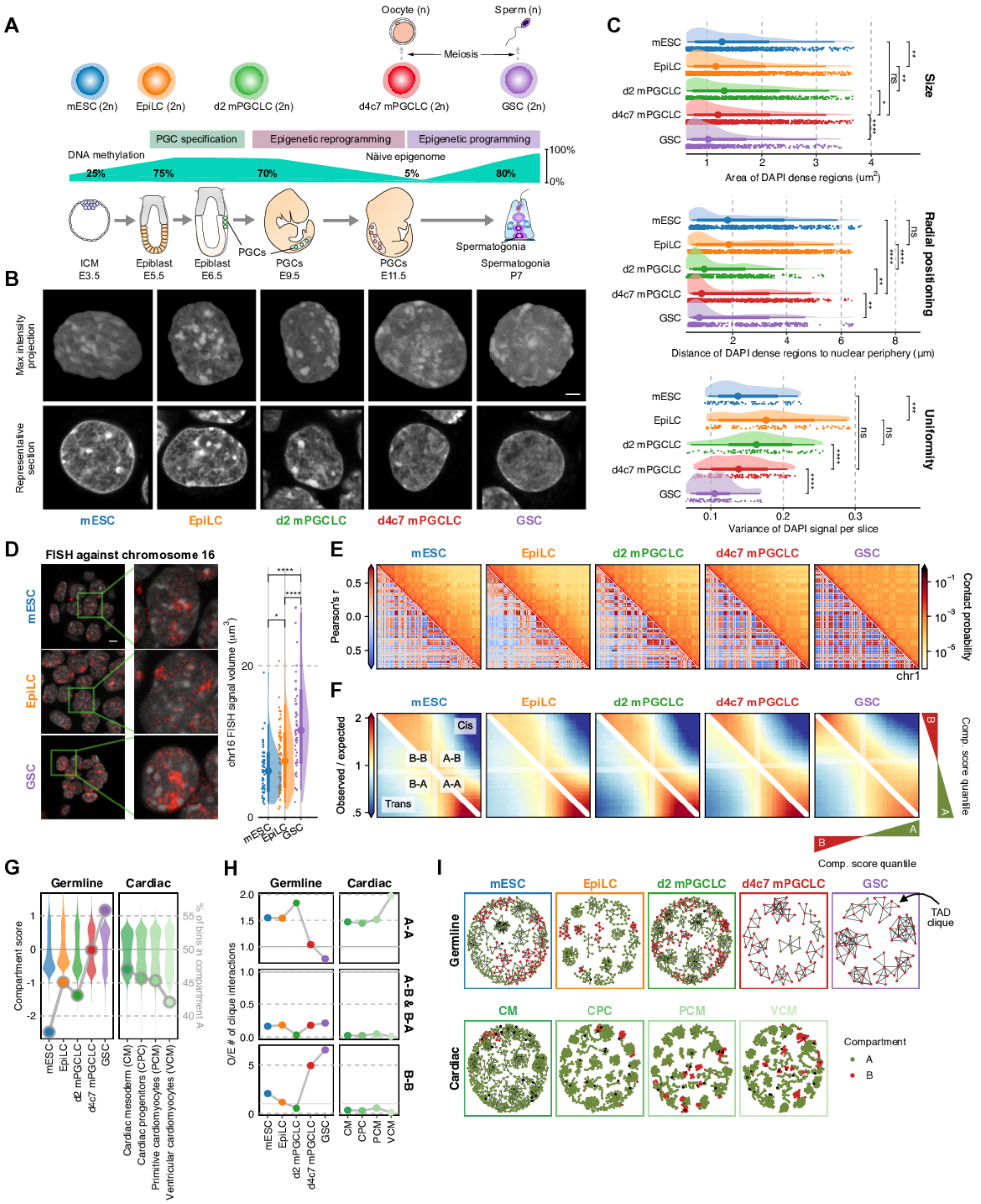
148 Davison, A. C. & Hinkley, D. V. *Boostrap methods and their application*. (Cambridge University Press, 1997).

149 Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178-194, doi:10.1016/j.stem.2015.06.014 (2015).

150 Karimi, M. M. *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* **8**, 676-687, doi:10.1016/j.stem.2011.04.004 (2011).

FIGURES

Figure 2.1. 3D genome programming.



(A) Scheme for mouse germ-cell development *in vitro* (top) and *in vivo* (bottom), with dynamics of genome-wide DNA methylation levels (middle).

(B) Maximum intensity projections (top) and representative sections (bottom) of typical nuclei of the indicated cell types stained with DAPI. Scale bars, 3 μm.

(C) Areas of DAPI-dense regions (top), distance of DAPI-dense regions from the nuclear periphery (middle), and variance of DAPI signals (bottom). The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. Number of DAPI dense regions = 950/1450/839/1535/736 and number of slices = 90/115/95/135/110 for mESC/EpiLC/d2/d4c7 mPGCLC/GSC. Significances are computed using Wilcoxon rank-sum tests, p-values from top to bottom 4.37e-3, 1.62e-3, 2.99e-2, 2.03e-10, 4.03e-1, <2.2e-16, 1.31e-3, 8.94e-3, 1.06e-4, 5.62e-2, 4.63e-5, 7.65e-13. P-value symbol brackets: **** = [0, 0.0001); *** = [0.0001, 0.001]; ** = [0.001, 0.01); * = [0.01, 0.05); ns = [0.05, 1].

(D) (left) Fluorescence in situ hybridization (FISH) against chromosome 16 (red) with DAPI staining (grey). Z-stacked representative images are paired with magnified views. (right) Distributions of surface volumes for chr16. The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. Number of cells = 51/68/53 for mESC/EpiLC/GSC. Scale bars, 5 μm. Significances are computed using Wilcoxon rank-sum tests, p-values from left to right: 4.16e-2, 4.33e-6, 8.68e-9.

(E) Hi-C maps of chromosome 1. (upper right triangle) 250 kb-resolution balanced contact probability matrices; (lower left triangle) matching Pearson's correlation matrices.

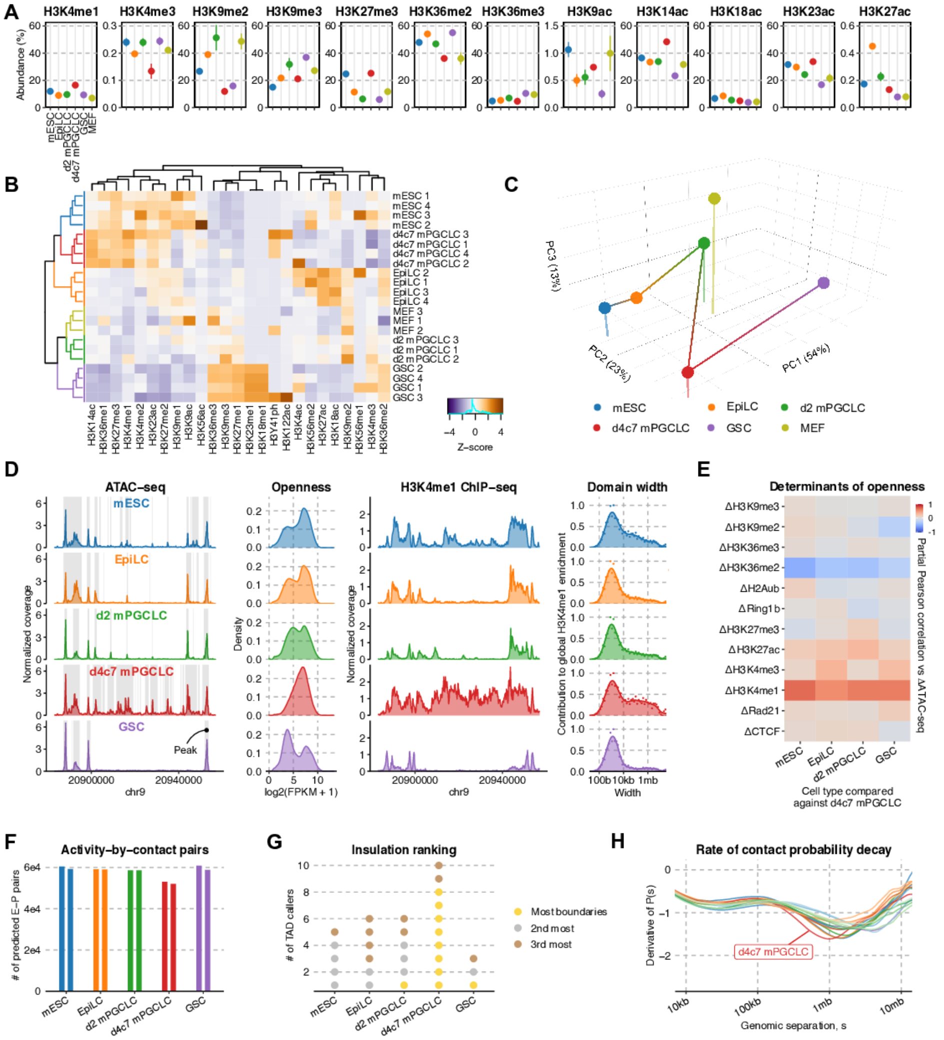
(F) Compartmentalization saddle plots for the average interaction frequency between pairs of 50 kb genomic bins belonging to various compartment-score quantiles in cis (upper right triangle) and trans (lower left triangle).

(G) Transitions in euchromatin-vs-heterochromatin bias during the development of different lineages (cardiomyocyte differentiation [14]) at 100 kb resolution. (left axis: violin plots) Distribution of compartment scores; (right axis: dots) ratio of A:B compartment bins.

(H) Enrichment of TAD-TAD interactions involved in max cliques (size ≥3) during the development of different lineages. A dispersal of active hubs was specifically observed during epigenetic reprogramming. Inter-compartmental TAD-TAD interactions are under-represented in all cases.

(I) Network representation of TAD cliques and their compartment identity during germ cell and cardiomyocyte differentiation.

Figure 2.2. Epigenome profiles and CTCF insulation.



(A) Relative abundance (%) of key histone modifications as measured by mass spectrometry. The point marks the mean while error bars indicate standard errors. Three biological replicates in each cell type were analyzed.

(B) UHC of H3 modification abundances. Numeric suffixes indicate biological replicates.

(C) PCA of average H3 modifications abundances in each cell type.

(D) Chromatin accessibility landscape throughout germline development. (left) ATAC-seq coverage tracks at a representative locus, with peaks highlighted; (second left) distribution of read counts per each in the union peak set; (second right) H3K4me1 ChIP-seq coverage tracks at the same locus; (right) Distribution of domain widths for H3K4me1-enriched regions based on cross-correlation, as implemented in MCORE.

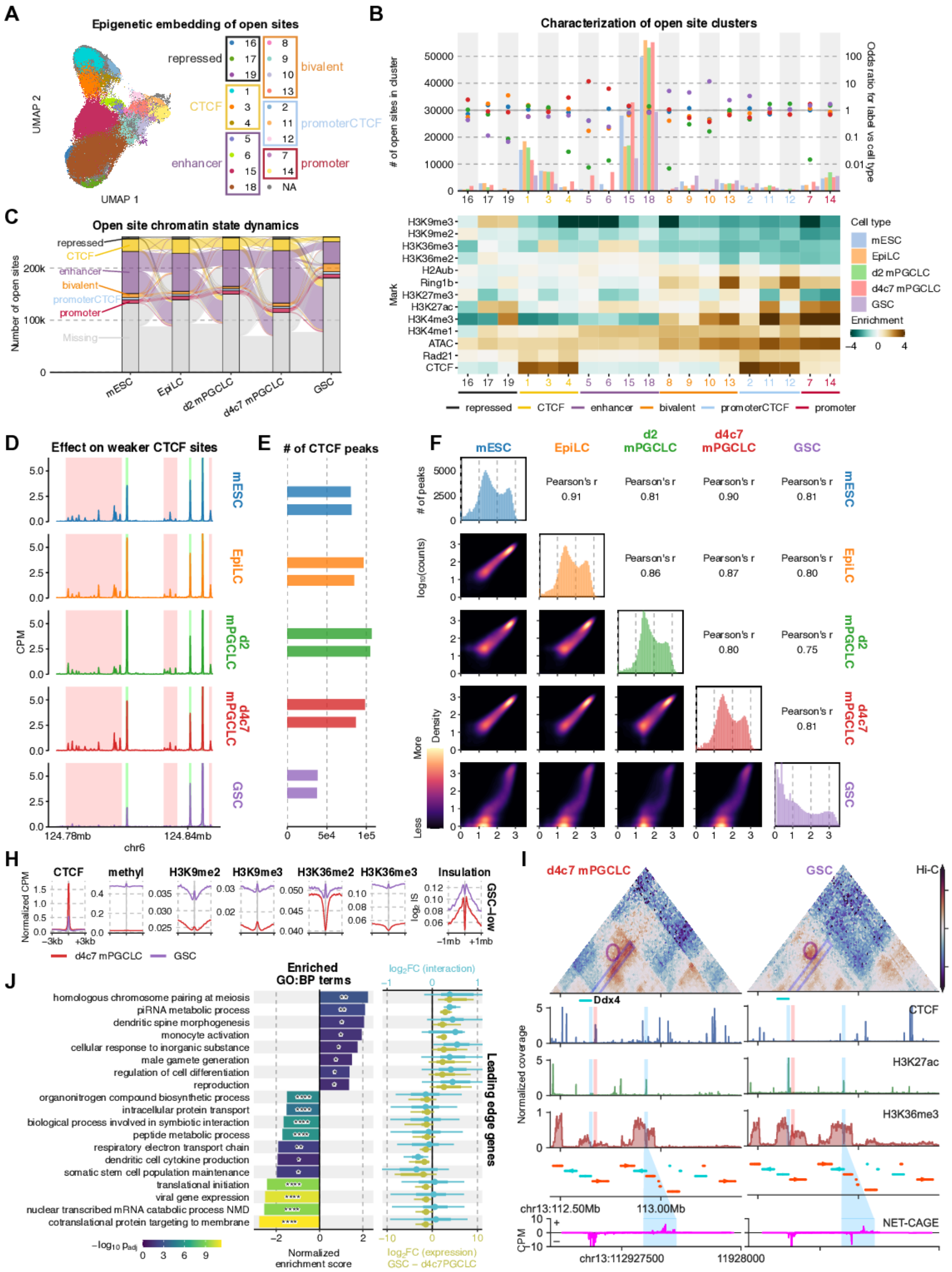
(E) Partial Pearson correlation matrix for inter-cell type ATAC-seq differences against d4c7 mPGCLCs versus differences in other epigenetic signals.

(F) Number of E-P pairs with ABC score > 0.02 [43]. Two biological replicates in each cell type were analyzed.

(G) Cell type insulation ranking. 10 different TAD-calling algorithms were used to determine the cell types rank in terms of insulation (gold: most insulated; silver: 2nd most insulated; bronze: 3rd most insulated).

(H) Slope of contact decay (P(s)) curves as a function of genomic separation in log-log space for the germline, neural induction [12], B cell reprogramming [13], and cardiomyocyte differentiation [14] datasets.

Figure 2.3. Open-site characterizations and CTCF release.



(A) 2D UMAP embedding based on epigenetic signals in ATAC-seq peaks for each cell type, with labels derived from semi-supervised HDBSCAN.

(B) Association between open-site clusters and cell types. (top) Number of open sites per cell type in each cluster (left axis: bars) and their enrichment as odds ratios (right axis: dots); (bottom) enrichment of epigenetic signals in each cluster.

(C) Dynamics of open site classes. Classification of the same open sites peak are compared between adjacent stages and shown as flows. Open sites that could not be reliably clustered or were not called as peaks are labelled as "Missing."

(D) ChIP-seq coverage tracks of CTCF in each cell type.

(E) Number of CTCF peaks called in each cell type. GSCs have considerably fewer CTCF peaks. Two biological replicates in each cell type were analyzed.

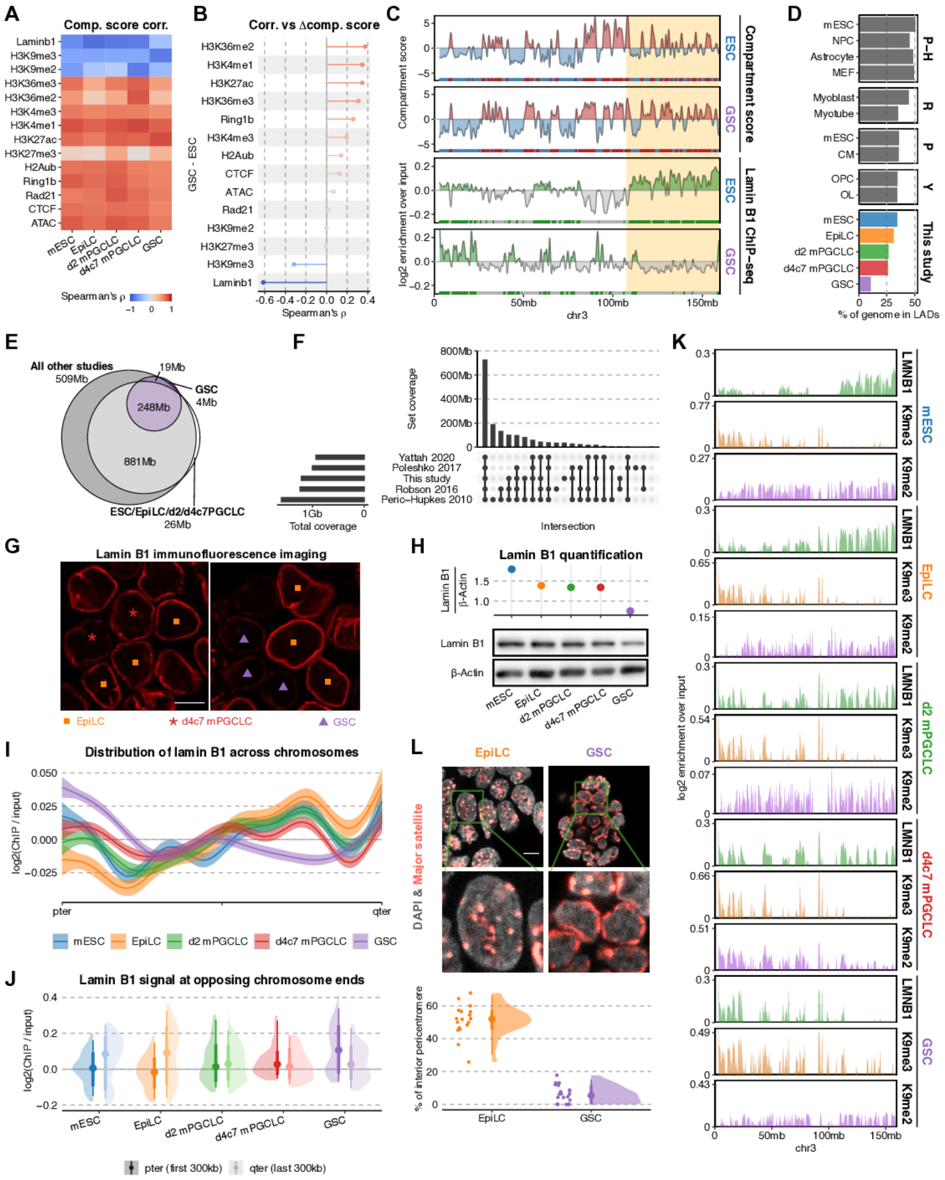
(F) Correlograms of CTCF binding in the union peak set. (Upper right panels) Pearson's correlation coefficients between log2 transformed signals. (Diagonal) Histograms of CTCF signal intensity in the union peak set. (Lower left panels) 2D density plots of CTCF binding in pairs of cell types.

(G) Aggregate plots of ChIP-seq enrichment for various targets and insulation score (IS) around CTCF-binding sites depleted in GSCs as compared to d4c7 mPGCLCs. n = 39408.

(H) 3D epigenetic landscape re-wiring near *Ddx4*. Observed/expected contact maps at 10 kb resolution for d4c7 mPGCLCs and GSCs are shown alongside select ChIP-seq and NET-CAGE coverage tracks. A strong insulating CTCF peak (highlighted in red) upstream of the *Ddx4* TSS (upstream blue highlight) is lost in GSCs, facilitating the interaction between the *Ddx4* promoter and an active enhancer (downstream blue highlight) demonstrating pronounced bidirectional nascent transcription (bottom).

(I) GSEA using genes ranked by concomitant differential expression and promoter interaction. (left) ABC-defined E-P pairs overlapping GSC-depleted CTCF peaks are used to rank genes based on coordinated E-P interaction and expression differences; (right) $\log_2$ fold changes for leading-edge genes of enriched gene sets. Significances computed using pre-ranked multilevel GSEA, p-values from top to bottom: 0.00106, 0.00343, 0.0173, 0.0269, 0.0382, 0.0439, 0.0109, 0.0439, 5.19e-6, 6.26e-5, 1.79e-6, 1.91e-6, 0.00931, 0.0454, 0.0125, 6.24e-10, 2.17e-11, 6.24e-10, 6.18e-12.

Figure 2.4. Generation of minimal LADs.



(A) Correlation between compartment score and ChIP-seq enrichment at 50 kb resolution.

(B) Correlation between differential compartment score and differential ChIP-seq enrichment between mESCs and GSCs at 50kb resolution.

I Representative chromosome-wide distributions of compartment score and lamin B1 enrichment for mESCs and GSCs.

(D) LAD occupancies in different cell types [50-53].

I Venn diagram of LADs called in GSCs, union of LADs called in all other cell types in this study, and union of LADs identified from all other studies [50-53].

(F) UpSet plot for the union set of LADs in different studies [50-53]. A majority of regions correspond to constitutive LADs.

(G) IF analysis for lamin B1 in (left) EpiLCs and d4c7 mPGCLCs, as well as (right) EpiLCs and GSCs. Symbols for each cell type are as indicated. Scale bars, 10 μm.

(H) Western blot for lamin B1 in different cell types (bottom) and quantification normalized by β-actin (top).

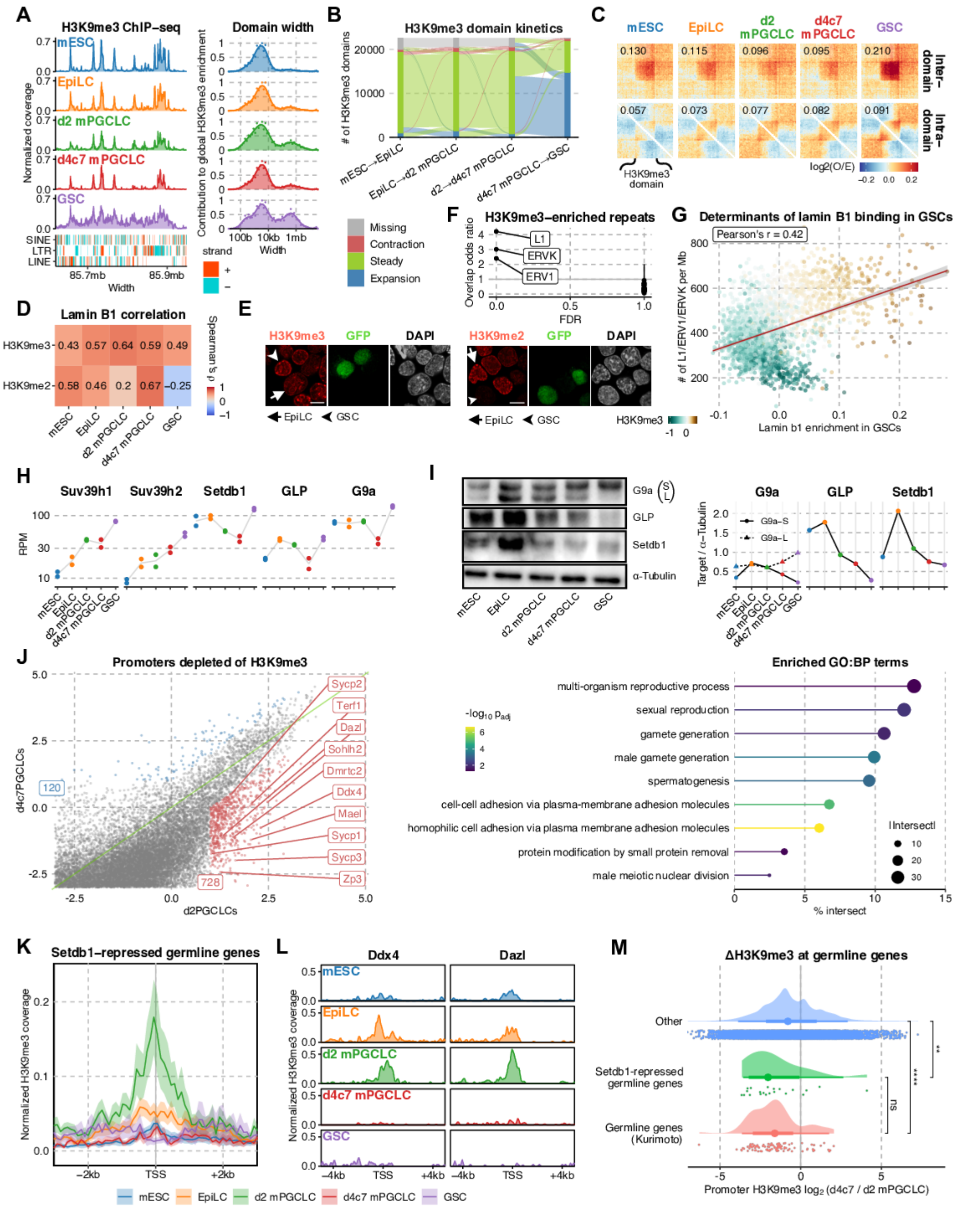
(I) Average distributions of lamin B1 enrichment across all chromosomes (1–19, X). Ribbons correspond to 95% confidence intervals of fitted GAMs.

(J) Lamin B1 ChIP-seq enrichment in the first (left/p-ter) and the last (right/q-ter) 300 Kb of each chromosome. The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. Number of chromosomes = 20 (autosomes and chromosome X).

(K) Representative chromosome-wide distributions of ChIP-seq enrichment for lamin B1 and H3K9me3/me2.

(L) (top) Representative images of FISH against major satellite repeats in EpiLCs and GSCs. Scale bars, 10 μm; (bottom) percentage of the pericentromeres detached from the nuclear lamina in EpiLCs and GSCs. The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. Number of cells = 18/22 for EpiLC/GSC.

Figure 2.5. Heterochromatin re-organization.



(A) (left) H3K9me3 ChIP-seq tracks, with TEs in different classes shown below; (right) Distribution of domain widths for H3K9me3-enriched regions based on cross-correlation, as implemented in MCORE.

(B) Spatial-temporal dynamics of H3K9me3 domains (>10 Kb) analyzed using ChromTime.

(C) (Top) Enrichment of interaction between (upper) and within (lower) broad H3K9me3 domains (>50 Kb; identified in GSCs and overlap peaks in all other cell types).

(D) Correlation between H3K9me2/3 and lamin B1 ChIP-seq enrichment.

(E) IF analysis for H3K9me3 (left) and H3K9me2 (right) in EpiLCs and GSCs. Arrowheads: GFP$^+$ GSCs; arrows: EpiLCs. Scale bars, 10 μm.

(F) Odds ratio and significance of overlap between H3K9me3 domains conserved across all cell types and different repeat families. Error bars denote 95% confidence intervals.

(G) Scatter plot of lamin B1 enrichment in GSCs vs the aggregated density of select TEs (L1, ERV1 and ERVK) in 1mb bins, with points colored by H3K9me3 enrichment in GSCs.

(H) Expression of H3K9 methyltransferases as measured by RNA-seq [23,25,149]. Two biological replicates in each cell type were analyzed.

(I) (left) Western blot for G9a, GLP, Setdb1 and α-tubulin; (right) quantification normalized by α-Tubulin.

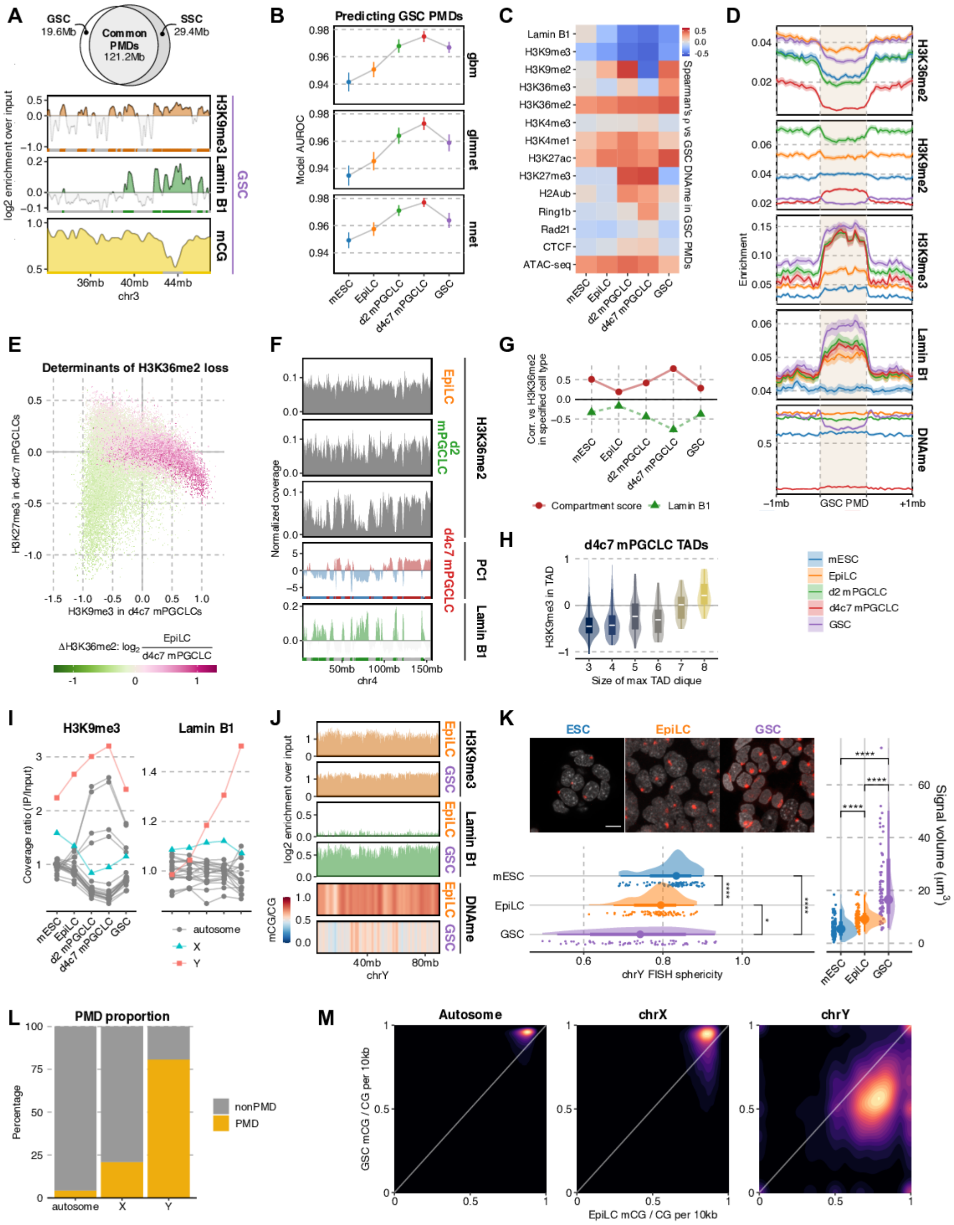
(J) (left) Scatter plot of H3K9me3 enrichment across all promoters in d2 mPGCLCs and d4c7 mPGCLCs, with 728 genes (red) showing substantially higher H3K9me3 levels in d2 mPGCLCs than d4c7 mPGCLCs; (right) pathway enrichment of the 728 genes using g:Profiler.

(K) Aggregate plot of H3K9me3 around the TSSs of *Setdb1*-repressed germline genes [150]. The thick line marks the mean while the upper and lower limits indicate standard errors.

(L) Normalized H3K9me3 tracks around the TSSs of *Dazl* and *Ddx4*.

(M) Distribution of differences in promoter H3K9me3 between d2 and d4c7 mPGCLCs for germline genes [29], *Setdb1*-repressed germline genes [150] and other genes. The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. From top to bottom, Number of genes = 19559, 21, 99. Significances are computed using Wilcoxon rank-sum tests, p-values from top to bottom: 2.36e-3, 1.14e-9, 4.43e-1.

Figure 2.6. Mechanism of PMD formation via balancing H3K36me2 vs H3K9me-marked LADs and Y chromosome hypomethylation.



(A) (top) Overlap of PMDs between spermatogonia [59] and GSCs; (bottom) Representative locus demonstrating colocalization of H3K9me3 and lamin B1 enrichment wit DNA hypomethylation.

(B) The area under the receiver operating characteristic (AUROC) of classifiers predicting 50 kb bins as either PMD or not in GSCs using each cell type's own epigenome. Error bars denote 95% confidence intervals.

(C) Correlation of GSCs' DNA methylation levels in GSC LADs with epigenetic signals in different cell types.

(D) Aggregate plots of H3K36me2, H3K9me2, H3K9me3, and lamin B1 enrichment as well as DNA methylation around PMDs in GSCs. The thick line marks the mean while the upper and lower limits indicate standard errors.

(E) Scatter plot of d4c7 mPGCLCs' H3K9me3 and H3K27me3 enrichment in 50 kb bins colored by differential H3K36me2 (EpiLCs–d4c7 mPGCLCs).

(F) Representative chromosome-wide distributions of compartment score, lamin B1 enrichment, and H3K36me2 coverage.

(G) Correlation between H3K36me2 and compartment scores or lamin B1 enrichment in 50 kb bins.

(H) Relationship between the max clique size involving a given TAD and the average H3K9me3 enrichment in that TAD in d4c7 mPGCLCs. Number of TADs with specific max clique sizes, from left to right: 798/269/94/35/18/11. The central band of boxplots indicate median values, while the lower and upper hinge correspond to the first and third quartile, and the upper whiskers extend to the largest value % 1.5 * IQR and vice versa for the lower whiskers.

(I) IP/input ratio of H3K9me3 and lamin B1 alignments per chromosome.

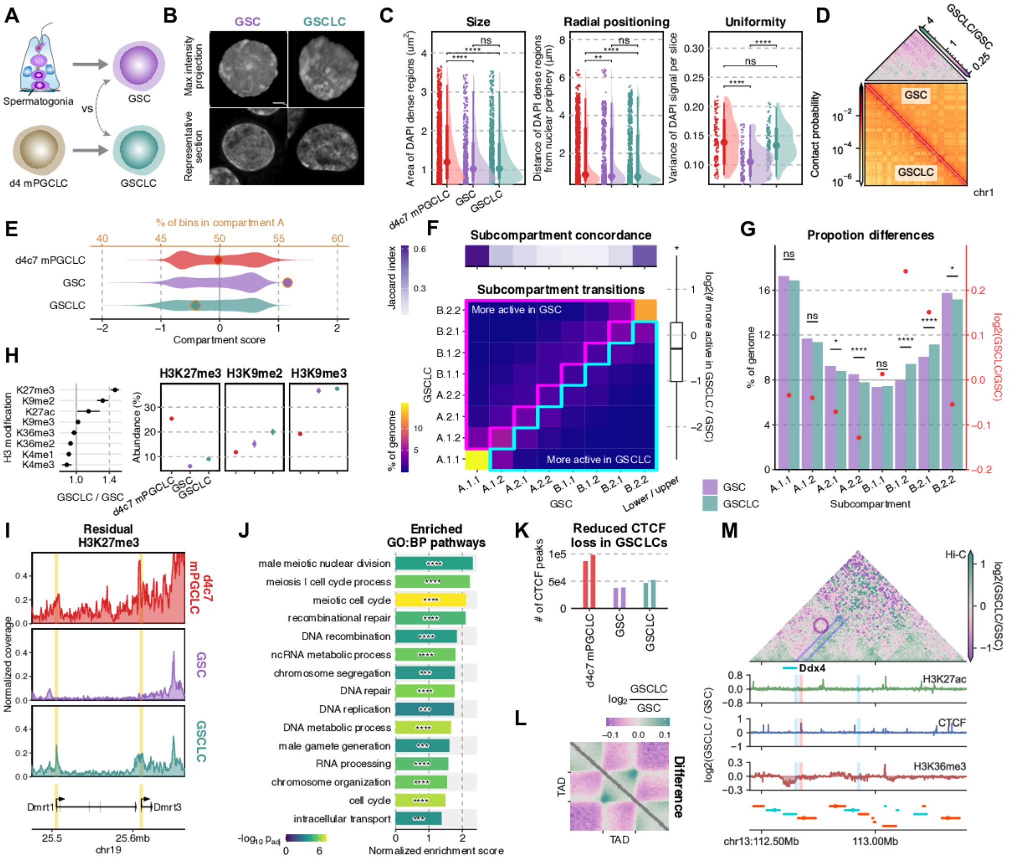
(J) Enrichment tracks of H3K9me3 and lamin B1 as well as DNA methylation in EpiLCs and d4c7 mPGCLCs on chromosome Y.

(K) (top) FISH against the Y chromosome; (bottom) sphericity of the Y chromosome FISH signals; (right) distributions Y chromosome surface volumes. The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. Number of cells = 89/76/69 for mESC/EpiLC/GSC. Scale bar, 10 µm. Significances are computed using Wilcoxon rank-sum tests, p-values from

(L) Proportion of the genome occupied by PMDs in GSCs with stratification by chromosome.

(M) 2D density plots of DNA methylation level (mCG/CG) between EpiLCs and GSCs in 10 kb bins.

Figure 2.7. Nucleome differences between GSCs and GSCLCs.



(A) Scheme for the derivation of GSCs and GSCLCs.

(B) Maximum intensity projections (top) and representative sections (bottom) of typical nuclei of GSCs and GSCLCs stained with DAPI. Scale bars, 3 μm.

(C) Areas of DAPI-dense regions (left), distance of DAPI-dense regions from the nuclear periphery (middle), and variance of DAPI signals (right). The point marks the median while the thick and thin lines correspond to 66% and 95% intervals, respectively. Number of DAPI dense regions = 1535/736/1227 and number of slices = 135/110/120 for d4c7 mPGCLC/GSC/GSCLC. Significances are computed using Wilcoxon rank-sum tests, p-values from left to right: 2.03e-10, 1.69e-9, 0.123, 0.00894, 8.02e-8, 0.0707, 7.65e-13, 0.417, 2.07e-12.

(D) (bottom) 250 kb resolution balanced contact probability matrices of chromosome 1 in GSCs (upper) and GSCLCs (lower); (top) fold change (GSCLCs/GSCs) of contact probability, showing an attenuation of distal interactions in GSCLCs.

(E) Distribution of compartment scores (bottom axis: violin plots) and ratio of A:B compartment bins (top axis: dots) at 100 kb resolution.

(F) Differential subcompartmentalization between GSCs and GSCLCs at 50 kb resolution. (top) Jaccard index between genomic bins belonging to each subcompartment in GSCs vs GSCLCs. (bottom) Comparison of subcompartment labels between cell types reveals a greater proportion of the genome belongs to the upper triangle, in line with GSCLCs being more repressive. (right) Quantification of matched bins in the upper vs lower triangle.

(G) Comparison of overall subcompartment proportions in GSCs vs GSCLCs. Most significant changes are again observed mostly for the intermediate states and not active euchromatin (A.1) or constitutive heterochromatin (B.2). Significances are computed using two-proportions z-tests, p-values from left to right: 0.0829, 0.107, 0.0169, 3.32e-5, 0.683, 1.09e-16, 7.46e-9, 0.0112.

(H) (left) Fold change (GSCLCs/GSCs) of different H3 modifications as measured by mass spectrometry, with confidence intervals denoting standard errors; (right) full data for select modifications. Three biological replicates in each cell type were analyzed.

(I) Normalized H3K27me3 coverage tracks around *Dmrt1* and *Dmrt3*.

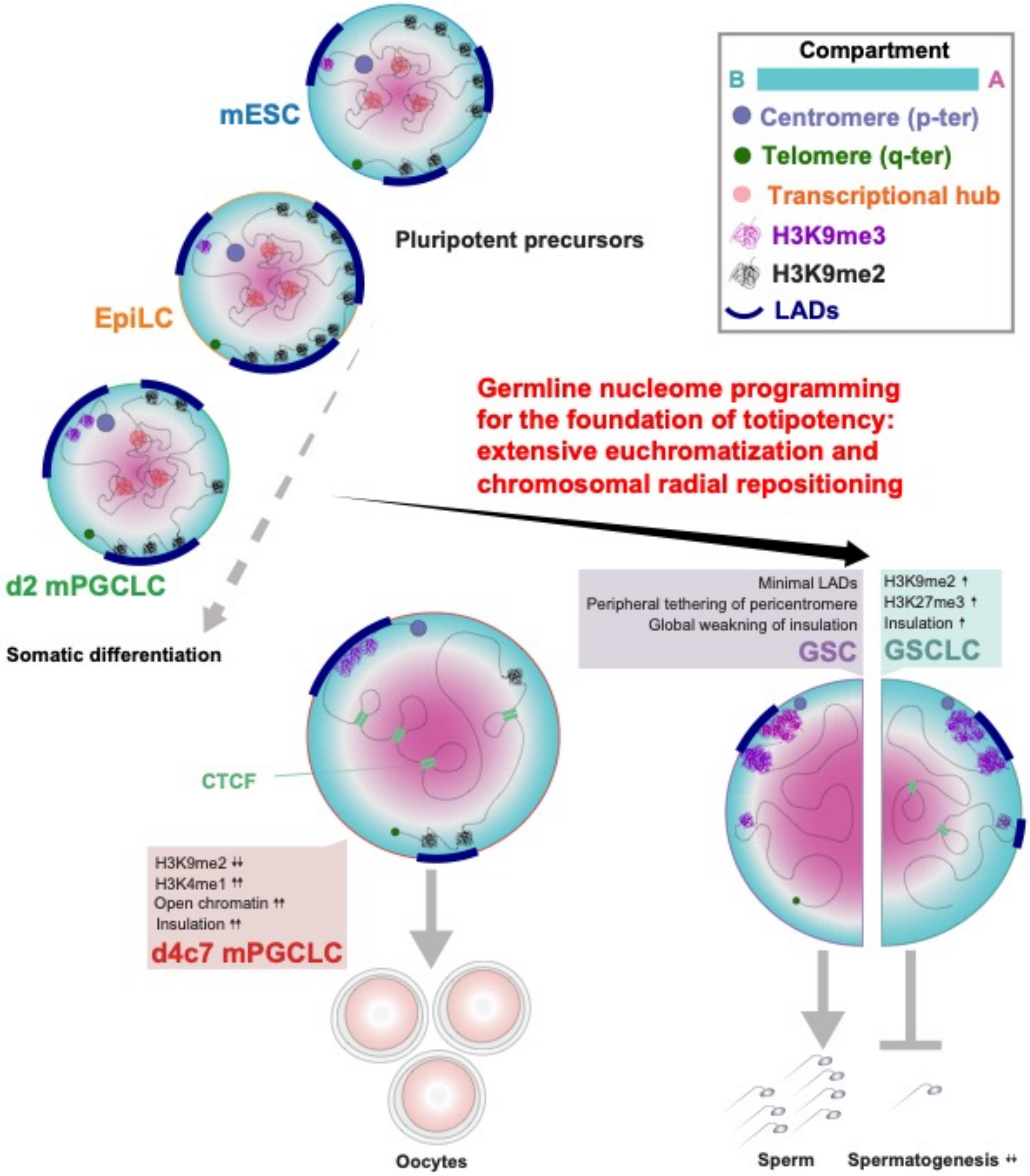
(J) GSEA results for promoters ranked by preferential enrichment in GSCLCs as compared to GSCs. Significances computed using pre-ranked multilevel GSEA, p-values from top to bottom: 7.11e-5, 3.31e-6, 5.15e-8, 1.44e-5, 9.71e-5, 3.06e-6, 0.000159, 1.42e-6, 0.000513, 3.66e-7, 0.000185, 8.43e-6, 3.31e-6, 3.7e-7, 0.000194.

(K) Number of CTCF peaks in each cell type. Two biological replicates in each cell type were analyzed.

(L) Pile-up plots of intra-TAD interactions in GSCs and GSCLCs.

(M) 3D epigenetic landscape rewiring near *Ddx4.* Differential (GSCLCs/GSCs) contact maps and ChIP-seq coverage at the *Ddx4* locus are shown. The insulating CTCF peak separating *Ddx4* from one of its enhancers is not completely removed in GSCLCs.

Figure 2.8. A model for the nucleome programming during mouse germ-cell development.



Unlike somatic fates, germline nucleome programming entails extensive euchromatization, which is associated with radial re-positioning of pericentromeres and peripheral de-attachment elsewhere. Augmented insulation helps to maintain transcriptional fidelity during global DNA hypomethylation in PGCs (PGCs bear oogenic potential as well). Insulators are subsequently erased en masse to activate gametogenic program during the PGCs-to-spermatogonia/SSC development. Faulty nucleome maturation involving intermediate compartment states leads to impaired spermatogenic capacity.

Non-mutational epigenetic reprogramming will come to be accepted as a bona fide enabling characteristic that serves to facilitate the acquisition of hallmark capabilities, distinct from that of genomic DNA instability and mutation. Notably, it can be anticipated that non-mutational epigenetic reprogramming will prove to be integrally involved in enabling the provisional new hallmark capability of phenotypic plasticity

Douglas Hanahan

Chapter 3

H3K27me3 spreading organizes canonical PRC1 chromatin architecture to regulate developmental transcriptional program

Having established the importance of nuclear architecture throughout germline development in the previous chapter, we now turn our attention to possible faulty 3D epigenome dynamics in disease. Polycomb group proteins, in particular, have a storied past as the archetypal mediator of epigenetic transcriptional regulation and serves critical functions in the proper orchestration of various early developmental programs.[65] For instance, germ-cell differentiation entails dynamic expression of germline-specific Polycomb-related factors such as SCML2 and EZHIP whose loss can lead to infertility.[66,67] Accumulating evidence suggests that Polycomb-mediated regulation can take shape both within local chromatin domains as well as across large genomic separations via distal chromatin looping. Recent results highlighting the importance of these processes in early development thus lead us to re-examine the effect of epigenetic dysregulation on 3D genome organization in several disease states studied by our labs.

Broadly, the Polycomb Repressive Complexes exist as critical machineries regulating transcription through inducing facultative heterochromatinization by means of depositing histone modifications such as H3K27me3 and H2AK119ub, the target of which are often key developmental regulators dynamically expressed during the course of differentiation. But the set of associated factors also form an intricate interaction network given their dual reader-writer capacities for diverse epigenetic markers. As the counterpart to Polycomb, a collection of SET domain-containing enzymes catalyzing H3K36 methylation such as NSD1 mediates the deposition of euchromatic H3K36 methylome, in turn shaping the facultative heterochromatin landscape through their mutual antagonism. With the factors governing this euchromatin-heterochromatin balance dynamically shifted during development and perturbed in a variety of diseases, we here holistically assess the impact of these events on the 3D epigenome through applying Hi-C, ChIP-seq, RNA-seq to assay stem and cancer cells.

H3K27me3 spreading organizes canonical PRC1 chromatin architecture to regulate developmental programs

Brian Krug[1,27], Bo Hu[1,2,3,27], Haifen Chen[1,2, 27], Shriya Deshmukh[4], Xiao Chen[5,6], Kristjan H. Gretarsson[5], Nisha Kabir[1,7], Augusto Faria Andrade[1], Elias Jabbour[4], Ashot S. Harutyunyan[1], John J. Y. Lee[8,9,10], Maud Hulswit[1], Damien Faury[11], Caterina Russo[11], Xinjing Xu[5], Michael J. Johnston[12,13,14], Audrey Baguette[15], Nathan A. Dahl[16,17], Alexander G. Weil[18], Benjamin Ellezam[19], Rola Dali[2], Mathieu Blanchette[20], Khadija Wilson[21], Benjamin A. Garcia[21,22], Marco Gallo[12,13,14], Michael D. Taylor[8,9,10,23,24], Claudia L. Kleinman[1,7], Jacek Majewski[1,2#], Nada Jabado[1,4,11,25#*], Chao Lu[5,26#]

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 1B1, Canada

[2]McGill University Genome Centre, Montreal, QC H3A 0G1, Canada

[3]Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan.

[4]Division of Experimental Medicine, Department of Medicine, McGill University, Montreal, QC H4A 3J1, Canada

[5]Department of Genetics and Development, Columbia University Irving Medical Center, New York, NY 10032, USA

[6]Marine College, Shandong University, Weihai, 264209, China

[7]The Lady Davis Institute, Jewish General Hospital, Montreal, QC H3T 1E2, Canada

[8]The Arthur and Sonia Labatt Brain Tumor Research Center, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada

[9]Developmental & Stem Cell Biology Program, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada

[10]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5G 1L7, Canada

[11]Department of Pediatrics, McGill University, The Research Institute of the McGill University Health Center, Montreal, QC H4A 3J1, Canada

[12]Arnie Charbonneau Cancer Institute, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada

[13]Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada

[14]Department of Molecular Biology and Biochemistry, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada

[15]Quantitative Life Sciences, McGill University, Montreal, QC H3A 1B9, Canada

[16]Morgan Adams Foundation Pediatric Brain Tumor Research Program, Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA

[17]Pediatric Neuro-Oncology, Center for Cancer and Blood Disorders, Children's Hospital Colorado, Aurora, CO 80045, USA

[18]Department of Pediatric Neurosurgery, Centre Hospitalier Universitaire Sainte-Justine, Université de Montréal, Montréal, QC H3T 1C5, Canada

[19]Department of Pathology, Centre Hospitalier Universitaire Sainte-Justine, Université de Montréal, Montréal, QC H3T 1C5, Canada

[20]School of Computer Science, McGill University, Montreal, QC H3A 2A7, Canada

[21]Department of Biochemistry and Biophysics and Penn Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104-6073, USA

[22]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO, USA, current institution

[23]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada

[24]Division of Neurosurgery, The Hospital for Sick Children, Toronto, ON M5G 1L7, Canada

[25]The Research Institute of the McGill University Health Center, Montreal, QC H4A 3J, Canada

[26]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY 10032, USA

[27]These authors contributed equally.

#Co-senior authors

*Corresponding author:

nada.jabado@mcgill.ca

Abstract

Polycomb Repressive Complex 2 (PRC2) deposits H3K27me3 to recruit canonical PRC1 (cPRC1) that maintains repressive heterochromatin. Higher order cPRC1-mediated chromatin interactions characterize early development and resolve during cell differentiation. Here, we use two opposing models of H3K27me3 dysregulation to elucidate how these long-range loops affect gene expression and cellular phenotypes. Aggressive gliomas driven by histone H3 Lys-27-Met (H3K27M) mutations or EZH-Interacting Protein (EZHIP) expression confine H3K27me3 deposition to PRC2 nucleation sites, while loss of the H3K36 methyltransferase NSD1 in pluripotent stem cells leads to its unrestrained spread from these sites. In H3K27M mutant tumours, focal H3K27me3 deposition concentrates chromatin occupancy of cPRC1 complexes, which results in long-range chromatin interactions anchored in polycomb bodies, mirroring patterns found in stem cells. Conversely, spread of H3K27me3 due to NSD1 loss dilutes cPRC1 deposition and disrupts polycomb body architecture. In H3K27M cells, H3K27me3 confinement sustains repression of genes tethered to the polycomb bodies, promoting self-renewing progenitor states required for tumour development. Maintenance of progenitor states depends on cPRC1 interaction with H3K27me3, as chemical allosteric modulation of chromodomains alleviates repression of transcription and promotes differentiation. These results suggest that H3K27me3 spread from CGIs modulates polycomb target gene expression to orchestrate developmental transitions through the maintenance or dissolution of cPRC1-mediated 3D chromatin interactions. Imbalances in the spread of H3K27me3 altering genome architecture represent important pathogenic mechanisms and therapeutic targets in diseases including H3K27M-mutant cancers and NSD1-associated neurodevelopmental syndromes.

Introduction

Polycomb Repressive Complexes 1 and 2 (PRC1, PRC2) are essential and evolutionarily conserved chromatin modifying complexes that regulate cell fate transitions in tissue development and homeostasis.[1,2] PRC1/2 repress expression of target genes that include transcription factors and developmental regulators establishing lineage identity. PRC2 is the "writer" that catalyzes histone H3 lysine 27 mono, di and tri-methylation (H3K27me1/2/3) post-translational modifications (PTMs), through methyltransferase subunits EZH2 or EZH1 in association with the other core subunits EED, SUZ12 and RBBP4/7.[3] PRC1 complexes are classified as canonical PRC1 (cPRC1) and variant PRC1 (vPRC1). cPRC1 can act as "readers" of H3K27me3 through Chromobox (CBX) subunits and organize long-range chromatin interactions to repress gene expression. In this context, chromatin compaction and aggregation are achieved through self-association of cPRC1 PHC1/2/3 subunits and phase separation properties of CBX2 subunits[4], which form repressive condensates termed polycomb bodies. vPRC1 act as the main writers of the repressive PTM histone H2A K119 ubiquitylation (H2AK119ub), through the RING1A/B ligase core subunits common to all PRC1.[5,6]

Stable silencing of polycomb target genes is achieved through multiple cooperative effects from PRC2, cPRC1 and vPRC1 complexes.[5,6] PRC2 is recruited to chromatin at nucleation sites, corresponding to unmethylated CpG Islands (CGIs) at target gene promoters or cis-regulatory elements. Following nucleation, H3K27me1, me2 and me3 are sequentially spread over adjacent chromatin domains. PRC2 subunits coordinately sense regulatory stimuli and the local chromatin environment, including antagonistic histone PTMs, to guide its nucleation and spread. Factors including H3K36me2/3 impair PRC2 activity and demarcate boundaries between active euchromatin and H3K27me3 domains.[3,7] The loss of H3K36me2 in cells lacking the H3K36 dimethylase Nuclear Receptor Binding SET Domain Protein 1 (NSD1) therefore results in increased spread of H3K27me3.[8] Genome-wide patterns of H3K27me3 are cell-type specific and dynamically regulated during cell differentiation. Focal confinement of H3K27me3 at many developmental genes characterizes primed pluripotent stem cells (PSCs), corresponding to early embryonic development. These sites are occupied by cPRC1, which congregate in space to form polycomb bodies.[2,9] Interestingly, dissolution of these cPRC1-mediated 3D structures is observed during exit from pluripotency despite the presence of H3K27me3, and, in various progenitor cells, this is associated with expression of differentiation-specifying polycomb target genes.[10,11]

Multiple cancers and genetic disorders present alterations to H3K27me3 genomic distribution, implicating the importance of tightly regulated PRC2 activity in normal physiology.[12] Midline high-grade gliomas (HGGs) and posterior-fossa group A ependymomas (PFA-EPN) are brain tumours characterized by histone H3 Lys-27-Met substitutions (H3K27M)[13], or by ectopic expression of EZH Inhibitory Protein (EZHIP), which shares structural resemblance to H3K27M.[14] Mechanistically, H3K27M and EZHIP converge to dramatically restrict H3K27me3 spreading beyond PRC2 nucleation sites.[15,16] These chromatin alterations associate with impaired tumour cell differentiation, and their oncogenic potential is restricted to narrow developmental windows in select cycling progenitor states.[17,18] Moreover, tumour cells harboring these alterations are critically dependent on residual H3K27me3[19,20], distinguishing H3K27M/EZHIP-associated malignancies from those displaying complete loss of PRC2 function.[14] The patterns of PRC2 distribution and H3K27me3 deposition strongly resemble those of primed PSCs, wherein focal H3K27me3 domains repress target genes to preserve self-renewal. To elucidate how patterns of H3K27me3 deposition can affect chromatin architecture, gene expression and cellular phenotypes, we used models with enforced H3K27me3 confinement (H3K27M, EZHIP-expressing tumours) and those with pervasive spreading of H3K27me3 (NSD1 loss) as tools to develop a mechanistic understanding of how spatial distributions of H3K27me3 impacts normal developmental processes and how its imbalanced spreading may promote disease states.

Results

Confined H3K27me3 deposition associates with enhanced chromatin looping

To address whether the level of spread of H3K27me3 impacts chromatin architecture, we compared global profiles of chromatin states and architecture by integrating Chromatin Immunoprecipitation and Cleavage Under Targets and Release Under Nuclear (ChIP/CUT&RUN-seq) and Hi-C chromatin conformation capture data in brain tumor and developmental contexts (**Fig 3.1**). First, we profiled glioma cell lines and tumours expressing H3K27M (4 cell lines, 4 tumours) or EZHIP (3 cell lines, 4 tumours). Isogenic patient HGG tumour-derived cell lines in which we either knocked out endogenous H3K27M mutant alleles (KO) or expressed H3K27M were generated to delineate the mutation's effect in disease-relevant contexts.[15,21,22] We also compared PFA-EPN-derived cell cultures expressing EZHIP when maintained in hypoxia to those losing EZHIP in normal oxygen levels.[20] Next, we analyzed developmental contexts that promote H3K27me3 deposition and spreading (**Fig 3.1a**). We tested human induced pluripotent stem cell (hiPSC) differentiation to neural progenitor cells (NPCs) and sampled published datasets characterizing mouse embryonic stem cell (mESC) differentiation to NPCs[11] and mESCs cultured in naïve (2i/LIF media) or primed (serum media) pluripotency conditions[23]. Entrance to primed pluripotency from the naïve ground state was previously described to confine H3K27me3 and heighten polycomb body architecture, both of which diminish when exiting pluripotency upon differentiation.[23] To evaluate genome-wide distributions across each model system, we adapted a ChIP-seq quality control measure called Fragment Cluster Score (FCS) that is meant to assess the concentration of reads mapping to forward and reverse strands with different genomic separation.[24] We validated that this metric not only accurately captures the dichotomy of confined versus diffuse H3K27me3 patterns and confirms the consistently punctate binding of CTCF, but that it also quantifies the magnitude of confinement based on simulating H3K27me3 profiles of varying breadth (**Fig C.1a-d**). This approach effectively showed consistent increase in H3K27me3 confinement in PSCs and H3K27M/EZHIP-expressing tumours (**Fig 3.1a-c**). When applied to published ChIP-seq data from mouse embryonic brain[25], this metric captured the expected reduction in H3K27me3 confinement along developmental progression (**Fig C.1e**), resembling patterns of PSC to NPC transitions. This analysis confirmed that H3K27me3 spreading is a feature of *in vivo* differentiation and early development.

Based on Hi-C data, we quantified and aggregated pairwise 3D interactions between CGIs enriched for H3K27me3. We found that such interactions are specifically enriched in H3K27M/EZHIP-expressing gliomas and primed PSCs showing confined H3K27me3 spread (**Fig 3.1d**). Notably, H3K27M did not appreciably alter the number of CTCF ChIP-seq peaks or contact frequencies among these sites (**Fig C.2a-b**), suggesting that these interactions are independent of conventional cohesin/CTCF-associated loops. At the Homeobox D (HOXD) cluster, H3K27M mutant cells display multiple distal loop contacts between H3K27me3 peaks, and these discrete interactions are lost in H3K27M-KO lines where H3K27me3 spreads over extended domains (**Fig 3.1e**). Strengthened interaction between two otherwise insulated domains is especially evident in the 3D structures predicted from the same locus (**Fig 3.1f**). These types of chromatin loops can span tens of megabases, cross TAD boundaries, and anchor multiple sites at one location, in stark contrast to typical loop extrusion associated structures, which are generally on a sub-megabase scale.[9,11,26] When assessing which transcriptional regulator's binding sites are over-represented among regions showing differential chromatin interactions between isogenic H3K27M/KO comparisons, PRC2 components emerged as the most strongly associated with H3K27M-specific interactions (**Fig C.2c**). Notably, changes in compartment or insulation scores across isogenic pairs derived from different cell lines showed limited concordance (**Fig C.2d**), suggesting that H3K27M-induced changes in H3K27me3 do not substantially alter compartmentalization and TAD architectures. In support of this hypothesis, when comparing the Hi-C profiles of various brain tumors subtypes, we found that compartment and insulation scores could not segregate H3K27M tumors from other malignancies (**Fig C.3**).

To assess whether the association between H3K27me3 confinement and 3D chromatin architecture extends beyond H3K27M/EZHIP-mediated malignancies, we investigated published datasets of other cancers or genetic alterations which may affect patterns of deposition of H3K27me3. Loss of histone H1 genes impairs H3K27me3 spread in lymphoma models.[27] Conversely, loss of the tumour-suppressor BAP1, an H2AK119 deubiquitylase, results in the opposite effect to increase H3K27me3 spread in mESCs.[28] In both scenarios, using published datasets of matched comparisons of H3K27me3 profiles and Hi-C, we observed that greater H3K27me3 confinement associated with increased contact frequencies between CGIs (**Fig C.4**). These results suggest that H3K27me3 spread altering chromatin architecture is a broadly applicable organizing principle in contexts of oncogenic transformation, and that limited H3K27me3 spread is universally linked with greater 3D contacts between H3K27me3 sites.

Confined H3K27me3 concentrates canonical PRC1 to drive polycomb body compaction

To delineate mechanisms of polycomb loop architecture, we next charted PRC1 localization and activity. cPRC1 links Chromobox (CBX2/4/6/7/8) recognition of H3K27me3 to spatial chromatin organization via SAM domain oligomerization between PHC1, PHC2 or PHC3 subunits and phase separation properties of the CBX2 subunit.[4,6] We profiled chromatin binding of PRC1 complexes by ChIP-seq of the RING1B core subunit, CBX2 that is unique to cPRC1 and robustly expressed across our cell lines (**Fig C.5a**), and the H2AK119ub mark to further assess the effect of the catalytic activity of PRC1 in our models.

Using a peak-calling approach, we noted that a minority of H3K27me3 and RING1B peaks overlap, marking dual PRC2 and PRC1 enrichment while CBX2 peaks largely overlap with these intersected sites. This indicates cPRC1 are enriched at H3K27me3/PRC2 foci, while vPRC1 complexes lacking CBX subunits but also containing RING1B are widely distributed outside of PRC2 domains (**Fig 3.2a-b, C.5b-c**). Sites with H3K27me3, RING1B and CBX2 overlap displayed significantly higher H3K27M-enriched contact frequencies compared to other peak categories (**Fig 3.2a**), suggesting a link between cPRC1 and the formation of polycomb-based chromatin loops in H3K27M cells. We subsequently sought to quantify how H3K27M affects cPRC1 signal intensity. When H3K27me3 spread is confined, the retainment of H3K27me3 enrichment at specific CGIs correlated with several-fold higher signal of RING1B and CBX2 enrichment (**Fig 3.2c**), indicating they are concentrated specifically in regions where H3K27me3 is confined. We observed this redistribution in other H3K27M glioma cells and by expressing H3K27M in WT glioma cells, showing the mutation drives these effects (**Fig C.6**). In contrast, cPRC1 redistribution did not appreciably alter tumour H2AK119ub profiles (**Fig 3.2b-d, C.6a, c, h**), corroborating observations that this PTM is largely deposited by vPRC1 complexes in somatic cell types.[29] Moreover, substantial RING1B enrichment was observed at H3K27ac-marked sites devoid of H3K27me3, displaying 3D interactions characteristic of active enhancers, indicative of vPRC1 recruitment to these loci (**Fig C.5d**).[10]

NSD1 antagonism of H3K27me3 spread preserves cPRC1 loop architecture in stem cells

Links between H3K27me3 confinement and polycomb-associated chromatin looping motivated our investigation of opposing scenarios of heightened PRC2 spread. As H3K36me2 limits H3K27me3 deposition, NSD1 mutations or inactivation diminish H3K36me2 and increase H3K27me3 spreading in stem cells and cancers.[8,30,31] Germline heterozygous loss of NSD1 also defines Sotos syndrome, wherein patients display precocious developmental progression and overgrowth,[32,33] with yet unclear effects of NSD1 loss on chromatin architecture. We profiled chromatin conformation in matched pairs of WT or mutant PSCs harboring either homozygous NSD1 loss in mouse ESCs, or heterozygous loss in human induced PSCs (iPSCs). Loss of NSD1 in both mouse and human PSCs led to H3K36me2 loss and increased H3K27me3 spread, resulting in subsequent weakening of interactions between H3K27me3 sites (**Fig 3.3a-d, C.7a-b**). In WT stem cells, peak overlap between H3K27me3 and RING1B largely corresponded with CBX2 peaks, representing the anchors of loops that dissolved upon NSD1-KO (**Fig 3.3e-f, C.7c-d**). Concomitant with coordinated changes in H3K36me2 and H3K27me3, we observed reduced cPRC1 binding upon NSD1 loss in regions showing H3K27me3 enrichment in WT cells (**Fig 3.3g-h, C.7e-g**). Taken together, NSD1 loss resulted in decreased H3K36me2 deposition, which led to extended spread of H3K27me3, cPRC1 dilution from CGIs, and decreased contact frequencies specifically at sites of PRC2/cPRC1 co-occupancy. Therefore, polycomb bodies in pluripotent stem cells are disrupted when H3K36me2 deposition is impaired, as this mark contributes to preserving cPRC1 concentration by promoting focal H3K27me3 confinement to PRC2 nucleation sites.

Distinctive features of polycomb bodies in pluripotency and cancer

We next integrated profiles for a greater variety of chromatin modifiers and PTMs to comprehensively characterize the diversity of regulatory region chromatin states in our model systems. To this effect, we generated ChIP-seq datasets of several active (H3K27ac, H3K4me3, H3K36me2/3), repressive (H3K27me3, PRC2/1, H3K9me3) and architectural (CTCF, cohesin complex with SMC1 member) features for our model systems and sampled several published datasets from mESCs. Akin to single-cell transcriptomics, we considered genomic intervals (promoters & CGIs) instead of cells as individual observations, whereas enrichment of different epigenetic marks were treated as quantitative features rather than expression of various genes.[34] In this framework, the application of dimension reduction (UMAP) followed by clustering (HDBSCAN) delineated four broad categories of cis-regulatory regions with distinct chromatin states (**Fig 3.4a**). This classification confirmed that H3K27me3-enriched sites with simultaneous RING1B and CBX2 binding constitute a distinct chromatin state in both mESCs and hiPSCs (**Fig 3.4b**). Accordingly, we next divided CGIs and promoters into 4 main clusters/classes; 1) Active (H3K27ac, H3K4me3 co-enriched), 2) cPRC1 (CBX2, RING1B, H3K27me3 co-enriched), 3) PRC2 (H3K27me3 and SUZ12 co-enriched, CBX2 absent) and 4) Other (lacking distinctive enrichment). Although clusters 2-to-4 broadly represent non-expressed genes, cPRC1 were repressed to a greater degree than PRC2 targets (**Fig. C.8a**). There was a strong degree of conservation in the promoter chromatin state assignments across species, with neurodevelopmental genes consistently over-represented among cPRC1 targets (**Fig C.8b-d**). Assessment of intra-class distal interaction revealed that loss of NSD1 consistently results in dissolution of long-range cPRC1 loops in hiPSCs and mESCs alike (**Fig 3.4c**). Moreover, reduced cPRC1-mediated interactions strongly correlated with target gene up-regulation (**Fig 3.4d-f**). Dual promoter enrichment of H3K27me3 and H3K4me3 has been described as a state of bivalency, wherein productive transcription is absent but can be rapidly activated upon change in stimuli.[35] As expected, cPRC1 targets in PSCs and NPCs showed H3K4me3/H3K27me3 co-enrichment (**Fig C.8g**), and H3K4me3 enrichment had no impact on differential cPRC1-centered interactions upon differentiation from PSCs into NPCs (**Fig C.8h**).

We similarly classified CGIs and promoters in three H3K27M tumour-derived cell lines (**Fig 3.5a-b**). cPRC1 targets in the HGG lines exhibited substantial overlap with those identified in hiPSCs, and an even stronger enrichment with those identified in human NPCs (**Fig C.8e-f**). However, in gliomas, we observed a bimodal distribution of H3K4me3 at cPRC1 targets, leading us to further subdivide them into H3K4me3+ (enriched) and H3K4me3- (absent) sub-clusters (**Fig 3.5c**). When analyzing chromatin interactions, we observed that H3K27M-dependent cPRC1-mediated loops were a feature specific to the sub-cluster absent for H3K4me3 (**Fig 3.5c-d, C.9e**). cPRC1 3D interactions at sites devoid of H3K4me3 enrichment were consistently elevated by the H3K27M mutation among our three isogenic glioma cell lines (**Fig C.9e**). These features were also evident in primary H3K27M HGGs and PFA-EPNs tumours, when compared to normal fetal and adult brain, and to brain tumours wild-type for H3K27M and lacking EZHIP expression (**Fig C.10a**), confirming the relevance of our cell line models to the patient disease context. We also found that H3K4me3- cPRC1 sites specifically gained H3K4me3 enrichment when H3K27M was lost (**Fig 3.5e**). Taken together, these results indicate that cPRC1 3D loops in H3K27M mutant cells are associated with diminished H3K4me3 deposition, and their dissolution following H3K27M loss allows acquisition of H3K4me3, suggesting this may promote transcription.

Aggregation of polycomb bodies represses target expression

To examine the effects of cPRC1 3D interactions observed in H3K27M mutant glioma cell lines on target gene transcription in primary tumours, we surveyed bulk and single-cell RNA sequencing datasets from H3K27M and histone-WT HGGs, and normal human fetal brain. Using consensus promoter classification derived from the intersection of three H3K27M cell lines, bulk tumour RNA-seq revealed that genes associated with active promoters are similarly expressed at high level across tumor types, whereas cPRC1 H3K4me3- targets are specifically repressed in H3K27M glioma tumour cells (**Fig 3.5f**). In contrast, H3K4me3+ cPRC1 targets were expressed at similar levels, suggesting that promoter occupied by cPRC1 can remain accessible for expression when not compacted by loop architecture in H3K27M tumours. Using tumour scRNA-seq to assess intercellular heterogeneity, we find that H3K4me3- cPRC1 targets were homogeneously silenced in H3K27M tumour cells, whereas the same gene set was variably expressed in H3WT HGG and fetal brain cells (**Fig C.10b-c**). We additionally partitioned subpopulations using known signatures of various cell types, which revealed that H3K4me3- cPRC1 targets are uniformly repressed across diverse cell types ranging from progenitors to differentiated cells only among H3K27M mutants (**Fig C.10c**).

To test the functional significance of repression of cPRC1 loop-associated genes, we examined experimental *in vivo* models of H3K27M's tumourigenic potential. We previously showed that

H3K27M is required to maintain tumour-forming competence in patient-derived HGG orthotopic models in immunodeficient mice (PDOX models).[15] Following KO of H3K27M in primary HGG lines, PDOX tumour development is either absent or occurs with substantially greater latency and less penetrance, while H3K27M lines form lethal high-grade tumours in a majority of mice. Despite the absence of tumour development, H3K27M-KO lines engrafted in the murine brain, determined by luciferase imaging of labeled cells and transcriptome analysis (**Fig 3.5g-i, C.10d**). Similar to our *in vitro* findings[17], H3K27M loss alleviates impaired differentiation, allowing glioma progenitor cells to acquire markers of mature glial cells. We recovered from mice brains matched pairs of engrafted H3K27M and KO cells and used scRNA-seq to profile *in vivo* cell states in the brain based on our group's reference atlas of the developing brain.[17] In all three distinct tumour cell lines, H3K27M tumour cell populations showed decreased representation of mature glial cells and increased fractions of glial progenitor-like cells, including those resembling radial glial cells (RGCs), which are progenitors at the top of the hierarchy in neural development (**Fig 3.5h-j, C.10e**); Notably, these RGCs were largely lost or substantially diminished in KO engrafted cell populations, which instead showed greater representation of more mature glial cell types such as astrocytes and oligodendrocytes. Consistent with results from primary tumours, we found that H3K4me3- cPRC1 targets were uniformly repressed in H3K27M cell type subpopulations while their up-regulation accompanies differentiation of KO cells (**Fig 3.5i, C.10e**). These findings support that H3K27M mutations impair differentiation of early glial progenitors in the presumed OPC lineage of origin.[17,36] Loss of the mutation leads to polycomb body dissolution, increased expression of cPRC1 target genes, and restoration of differentiation competence. Taken together, we observe that cPRC1-mediated loop architecture contributes to stable silencing of developmental target genes in both models of PSCs and tumour development.

We next assessed the transcriptomic consequences of cPRC1 loop dissolution using different cell culture media aimed to promote further cellular differentiation. Glioma progenitors maintained in media conditions favoring neural stem cell self-renewal (stem cell media) largely maintain stemness features as they do not receive environmental cues to differentiate. We have previously shown that, while H3K27M dramatically impairs tumour cell differentiation, this effect only emerged in *in vitro* cultures when applying culture media promoting glial fate acquisition (differentiation media), compared to isogenic KO cells.[17] Accordingly, in stem cell media, select genes (e.g., PRDM13) becomes modestly up-regulated in KO lines after loss of cPRC1 loops (**Fig 3.6a-b**). In differentiation media, we saw more pronounced differences in transcript expression, with notably more potent de-repression of cPRC1 targets in KO lines as compared to H3K27M cells across three cell lines (**Fig 3.6c, C.11a**).

Obstructing CBX chromodomain recognition of H3K27me3 alleviates differentiation blockade

Having established the regulatory importance of cPRC1, we sought to test whether it is causal to target repression and H3K27M-induced impairment of differentiation. cPRC1 recruitment and residence time on chromatin depends on chromodomain affinity for H3K27me3 substrate recognition.[37] Chemical probes have been developed that selectively obstruct the chromodomain reading of H3K27me3 by various CBX proteins. These allosteric modulators (CBX-AM) can precisely dilute cPRC1 occupancy at H3K27me3-marked targets, promoting their expression.[38]

We thus tested whether these compounds could overcome cPRC1 target repression and reverse differentiation impairment in H3K27M tumours. To this effect, we added the CBX-AM compound UNC4976 to H3K27M cell cultures upon application of differentiation stimuli. Transcriptome analyses of treated H3K27M cultures compared to the vehicle-treated controls showed specific de-repression of cPRC1 targets (**Fig 3.6d, C.11b**). We additionally confirmed by ChIP-seq that CBX-AM treatment specifically led to reduction of CBX2/RING1B enrichment at cPRC1 cluster sites, without perturbing RING1B enrichment at active promoters (RING1B here likely represents vPRC1) (**Fig 3.6e**). To further validate that increased expression of developmental regulators such as DLX1 in KO and CBX-AM treated H3K27M cells effectively promotes differentiation (**Fig 3.6f**), we profiled cells in culture media for markers of differentiation in the presence or absence of UNC4976. We compared acquisition of the astrocyte marker Glial Fibrillary Acidic Protein (GFAP) in DIPGXIII and the OPC marker SRY-Box Transcription Factor 10 (SOX10) in BT245 between these conditions and to what we had observed when H3K27M was knocked out. CBX-AM treatment was associated with increased

GFAP and SOX10 expression to levels comparable to those observed in KO lines (**Fig 3.6g, C.11c-d**). cPRC1 dilution and target de-repression therefore endow competence for greater differentiation to mature glia. Consistently, there was a significant correlation of differentially expressed genes between CBX-AM treatment and H3K27M-KO (**Fig C.11e**). Taken together, H3K27M-driven focal H3K27me3 confinement and subsequent cPRC1 assembly at these loci via its chromodomain reading of H3K27me3 form the basis of stable target repression. Therefore, by impairing cPRC1 aggregation either through spreading of H3K27me3 following H3K27M KO, or obstruction of CBX reading of H3K27me3 by CBX-AM compounds, we can alleviate the repression of target genes associated with polycomb bodies, thereby restoring cell differentiation (**Fig 3.6h**).

Discussion

Loss and gain of function alterations to PRC2/1 can promote cancers or genetic disorders, yet a unifying molecular mechanism underlying their roles in disease pathogenesis is lacking. As shown here, H3K27M and EZHIP tumour drivers converged on the restriction of H3K27me3 spread, confining its deposition to PRC2 nucleation sites. These effects drive cPRC1 recruitment and concentration at a subset of H3K27me3 sites that promote interactions linked to the formation of distal loop architecture across the genome. These polycomb bodies repress transcription of key genes involved in development to impair glial differentiation, accounting for features resembling stalled development in these tumours.[17,18] Conversely, we show that human and mouse PSCs losing NSD1 activity can abnormally spread H3K27me3 beyond nucleation sites following decreased H3K36me2 deposition. The pervasive H3K27me3 domain expansion in turn dilutes cPRC1 deposition and results in loss of 3D interactions characteristic of WT PSCs, allowing for de-repression of select polycomb target genes (**Fig 3.7**). We further show that compounds preventing CBX subunit recognition of H3K27me3 can reverse cPRC1 concentration on chromatin, de-repress target genes and endow competence for differentiation that mirrors effects from H3K27M loss. We also observed altered H3K27me3 spreading linked to changes in interaction frequencies in H1 knockout lymphoma models and BAP1 knockout mESCs. These effects are thus plausible contributors to oncogenic transformation in specific progenitor states with limited PRC2 activity.[27] We propose this can help explain the narrow developmental windows permissive to the oncogenic effects of H3K27M[39] or EZHIP.[18,20]

The extent of H3K27me3 spread from PRC2 nucleation sites is shown here to dictate the formation of higher order chromatin architecture among diverse cells. We can thus broadly categorize H3K27me3 patterns into a dichotomy of confined or diffuse patterns, with the degree of confinement predicting the extent of distal interactions. This model may help to account for unexplained observations in regulatory roles of polycomb complexes. Two distinct PRC2 subcomplexes display either heightened occupancy at targets (PRC2.1) or spread (PRC2.2) depending on their accessory subunits. Exclusive formation of PRC2.2 allows the expression of strong polycomb targets in hiPSCs, despite increased spread of this complex mediated through its positive feedback with H2AK119Ub (Youmans 2020, science 2021). Forced enrichment of PRC2.1 has been shown to be associated with preservation of target gene repression in iPSCs[40], possibly through maintenance of polycomb bodies promoted by focal H3K27me3 deposition, as we show. When PRC2.2 spreads H3K27me3 beyond CGIs, we speculate that relatively increased expression of these polycomb targets may be a consequence of cPRC1 dilution and resolution of the 3D loops. We further hypothesize that vPRC1/ H2AK119Ub /PRC2.2 spread mediates weaker transcriptional repression, priming these lowly expressed genes for rapid induction upon the right stimuli during cellular differentiation. This may account for how H3K27M and EZHIP prevent further cellular differentiation by retaining H3K27me3 at CGIs.

Our data also indicate that H3K27M mutations shape distal loop architecture independent of the broadly distributed H2AK119ub. Previous reports showed that RING1B presence, but not its catalytic activity, is required for maintenance of polycomb architecture in mESCs[9], and mice deficient in CBX2's compaction properties are defective in PRC1 repression.[41] However, other work challenges the view that cPRC1 contributes to gene repression in PSCs and emphasizes the role of vPRC1 in silencing transcription.[42,43] Here we used several somatic cell models relevant to disease pathology to show that the extent of genome-wide H3K27me3 spread is central to cPRC1-driven repression. Future work may enable approaches to determine whether cPRC1-driven local promoter chromatin compaction is sufficient to repress transcription, or whether distal loop architecture is critically required for this effect.

Notably, we identified that chromatin of polycomb bodies caused by H3K27M can exclude H3K4me3 deposition. This architectural feature lends understanding to proposed models of bi-phasic switches in promoter activity that can arise from competition between Polycomb group and Trithorax group members.[44]

Our findings suggest a mechanism for the phenotypic resemblance between Sotos syndrome (heterozygous NSD1 loss of function) and Weaver syndrome-related overgrowth disorders that include mutations in EZH2, SUZ12 or EED subunits of PRC2.[45,46] Loss of H3K27me3 could likely dilute cPRC1 repression in progenitor cells and parallel the effects of NSD1 mutations. Precocious polycomb body dissolution may thus contribute to Sotos syndrome patients' characteristic accelerated growth rates from early childhood and advanced morphological and molecular aging markers.[47] Hox gene clusters are prototypical polycomb targets that are altered in cancers and Sotos syndrome.[33,48,49] They begin as silenced in pluripotency and then segmentally lose PRC2/1 repression during spatial axial patterning and temporal maturation.[50] H3K27M tumours express select anterior segments of Hox clusters, which can evolve towards developmentally 'aged' states after loss of H3K27M-driven compaction of posterior segments (represented in cPRC1, H3K4me3- genes). We propose that the inaccessibility of master regulatory posterior Hox codes may contribute to stalling differentiation of tumours with restricted H3K27me3 spread.

Many cancers depend on PRC2/1 activity, motivating the design of targeted therapies against them. Understanding cPRC1 functional complexity will require the study of subunit expression across cell types, their stochiometric composition and the variety of cofactors recruiting them to chromatin. These careful considerations will help identify therapeutic agents modulating these chromatin structures with the potential to overcome aberrant polycomb silencing and self-renewal in numerous cancers.

Methods

Patient samples and clinical information. This study was approved by the Institutional Review Board of the respective institutions from which the samples were collected. We thank Keith Ligon and

Michelle Monje for generously sharing primary tumor cell lines established from patients with high-grade glioma.

Cell culture. Tumor-derived cell lines were maintained in Neurocult NS-A proliferation media (StemCell Technologies) supplemented with bFGF (10 ng/mL) (StemCell Technologies), rhEGF (20 ng/mL) (StemCell Technologies), and heparin (0.0002%) (StemCell Technologies) on plates coated in poly-L-ornithine (0.01%) (Sigma) and laminin (0.01 mg/mL) (Sigma). Lines were cultured to become differentiated glioma cells by adaptation to media of DMEM-F12 (Wisent) supplemented with 10% FBS (Wisent) for 10-14 days on coated plates. Measurements of GFAP/SOX10 differentiation in CBX-AM experiments used the following differentiation media; BT245 (DMEM-F12 (Wisent), NeuroCult SM1 Without Vitamin A (StemCell Technologies), N-2 supplement (Thermo Fisher), T3 hormone 3 ug/mL (Sigma), rhPDGF-AA 40 ng/mL (R&D Systems)) or DIPGXIII (DMEM-F12 (Wisent), FBS 10% (Wisent), rhCNTF 50 ng/mL (R&D Systems), rhBMP4 50 ng/mL (R&D Systems)). All lines tested negative for mycoplasma contamination, checked monthly using the MycoAlert Mycoplasma Detection Kit (Lonza). Tumour-derived cell lines were confirmed to match original samples by STR fingerprinting. The A1 mouse ES WT and NSD1-KO lines[51] (background C57BL/6 × 129S4/SvJae F1) were obtained from David Allis lab and maintained on gelatin-coated plates in Knockout DMEM (Gibco) supplemented with 15% ES-cell-qualified FBS (Gemini), 0.1 mM 2-mercapoethanol, 2mM L-glutamine (Life technologies) and LIF. The NCRM-1 hiPSC line was obtained from Tom Durcan and cultured on Matrigel (Corning) coated plates in mTeSR Plus media (StemCell Technologies). Editing of NSD1+/- clonal lines was performed using Integrated DNA Technolgoies Alt-R S.p. Cas9 Nuclease V3 and Alt-R CRISPR gRNAs, followed by deep sequencing of the target locus to validate mutations.

Mouse orthotopic xenograft. All mice were housed, bred and subjected to listed procedures according to the McGill University Health Center Animal Care Committee and were in compliance with the guidelines of the Canadian Council on Animal Care. Brain tumour cell cultures were transduced with lentiviruses constitutively expressing GFP and luciferase and selected by flow cytometry. Female NSG mice (4-6 weeks) (The Jackson Laboratory) were used for xenograft experiments using 12-20 mice per experimental group. Cell lines were engrafted using 7x10^5 cells in the caudate putamen (BT245, HSJ019) or the pons (DIPGXIII) by Robot Stereotaxic machine (Neurostar). Mice were imaged for luciferase signal and monitored for neurological symptoms of brain tumours, including weight loss, epilepsy, altered gait and lethargy. Mice were euthanized when clinical endpoint is reached, followed by removal of the brain. Tissue was sectioned into pieces and a portion of tumour or normal brain was dissociated using the MACS Brain Tumor Dissociation Kit (Miltenyi Biotec). Dissociated cells were cryopreserved in CryoStor CS10 (StemCell Technologies), followed by thawing, PBS washing and sorting for GFP signal using a BD FACSAria Fusion flow cytometer machine at the McGill University Health Centre Research Institute platform. Collected GFP+ cells were used as input for scRNA library prep beginning with 5000 viable cells per sample. Each experimental group was profiled using 2-3 animals for collection.

Chromatin immunoprecipitation sequencing. Cells (cell lines or dissociated tumor cells) were fixed with 1% formaldehyde for 10 minutes at room temperature. Cells were lysed in Cell Lysis Buffer for 30 minutes on ice. Nuclei were pelleted at 5000xG for 10 minutes at 4°C, and were resuspended in 200 µL/10M cells Nuclei Lysis Buffer for 45 minutes on ice. Lysed nuclei were sonicated on a BioRuptor UCD-300 for 18-30 cycles, 30s on 30s off, centrifuged every 15 cycles, chilled by 4°C water cooler. Samples were checked for sonication efficiency to 150-500bp size range by gel electrophoresis. Following centrifugation of samples at 12000xG for 10 minutes at 4°C, supernatants were collected and the chromatin was diluted in RIPA buffer to reduce SDS level to 0.1%. Before ChIP reaction 2% of sonicated drosophila S2 cell chromatin was spiked-in the samples for quantification by exogenous reference genome normalization. ChIP reaction for histone modifications was performed on a Diagenode SX-8G IP-Star Compact using Diagenode automated Ideal ChIP-seq Kit. 25ul Dynabeads Protein A beads (Invitrogen) Dynabeads M-280 Sheep anti-Mouse IgG beads (Invitrogen) were washed and then incubated with antibodies (anti-H3K27me3 (1:40, CST 9733), anti-H3K4me3 (1:40, CST 9751), anti-H2AK119ub (1:40, CST 8240), anti-H3K27ac (1:100, Diagenode C15410196), anti-H3K36me2 (1:50, CST 2901), anti-H3K36me3 (1:50, Diagenode C15200183), anti-H3K9me3 (1:50, Active Motif 39161)). One-two million cells of sonicated cell lysate combined with protease inhibitors for 10 hr, followed by 20 min wash cycle with provided wash buffers. ChIP reactions for SUZ12, RING1B, CBX2, SMC1 and CTCF were performed as follows: antibodies (anti-SUZ12 (1:150, CST 3737), anti-RING1B (1:200, Active Motif 39663), anti-CBX2 (Bethyl A302-524 1:200), anti-CTCF (1:400, Diagenode C15410210), anti-SMC1 (1:200, CST 4802)) were conjugated by incubating with 40ul protein A or G beads at 4°C for 6 hours, then chromatin from 5-10 million cells was added in RIPA buffer, incubated at 4°C overnight, washed using buffers RIPA, RIPA+500mM NaCl, LiCl and TE. Reverse cross linking took place on a heat block at 65°C for 4 hr. ChIP samples were then treated with 2ul RNase Cocktail at 65°C for 30 min followed by 2ul Proteinase K at 65°C for 30 min. Samples were then purified with QIAGEN MiniElute PCR purification kit as per manufacturers' protocol. In parallel, input samples (chromatin from about 50,000 cells) were reverse crosslinked, and DNA was isolated following the same protocol. Library preparation was carried out using Kapa HTP Illumina library preparation reagents. Half of ChIP DNA was used in End Repair and A-tailing reaction mix, followed by Illumina TruSeq DNA UD Index ligation (3 μL of 1/10 dilution) for 20 minutes at 20°C. The entire bead-purified ligation sample was amplified by 9-12 cycles of PCR. Size selection was performed after PCR using 0.6x/0.8x ratios of AMPure XP beads to collect 250-450bp fragments. ChIP libraries were sequenced using Illumina HiSeq 2000, 2500 or 4000 or NovaSeq 6000 platforms at 50 or 100 bp single reads. For mESC ChIP-seq of RING1B and CBX2, the protocol was adapted from Lee et al. (2006).[52] For each immunoprecipitation, 30M cells were dissociated, resuspended in media and crosslinked in 1% paraformaldehyde for 3 minutes at room temperature. Cells were lysed in lysis buffer 1, lysis buffer 2, lysis buffer 3 and 100uL 10% Triton X-100. Samples were sonicated with the Covaris M220 machine (Peak Power 75, Duty Factor 10, 200 cycles for 25 minutes). For antibody conjugation, 75uL of Dynabeads were washed and antibodies (5uL anti-RING1B, Active Motif 39663, or 10uL anti-CBX2 Bethyl A302-524A) were added. Beads were resuspended in lysis buffer 3, and sonicated supernatant was added and rotated overnight at 4°C. Between magnet captures, beads were washed in consecutive buffers; Low salt, High salt, LiCl, and TE with 50mM NaCl. Chromatin was eluted from beads, along with input chromatin sample. Next RNA and protein were digested and DNA was recovered by Qiagen

PCR purification kit. Eluted DNA was captured for library preparation and sequencing as previously described. Analysis used 1-3 replicates per condition in matched experimental designs.

CUT&RUN-sequencing. Reagents and protocol were based on the Epicypher commercial protocol. Briefly, 5x10^5 cells per sample were dissociated, washed and bound to CUTANA Concanavalin A coated Paramagnetic Beads (Epicypher). Antibodies were bound to cells overnight using 0.5 uL per sample (CBX2 ab1 CST 18687, CBX2 ab2 CST E3N6A, CBX2 ab3 Novus NBP247524, CBX4 CST 30559, CBX8 CST 14696, RING1B CST 5694, H3K27me3 CST 9733, SUZ12 CST 3737, EED Abcam ab4469, IgG control Thermo Fisher 02-6102). Digestion of target chromatin used CUTANA pAG-MNase, followed by DNA collection. Libraries were generated using Kapa HTP Illumina library preparation reagents using 11-12 cycles of PCR, followed by dual 0.6-0.8x size selection using AMPure XP magnetic beads. Analysis used 1-3 replicates per condition in matched experimental designs.

High-throughput chromosome conformation capture. In situ Hi-C libraries were generated from samples, as described previously with minor modifications.[26] Briefly, in situ Hi-C was performed in 7 steps. (1) Cell cultures or frozen tissue were dissociated into single cell suspensions of approximately 2-5M cells, washed in PBS and crosslinked with 1% formaldehyde for 10 minutes and pelleted. (2) Digestion of DNA used a 4-cutter restriction enzyme (DpnII, 100U) within intact permeabilized nuclei, (3) Filling in and biotinylating the resulting 5′ overhangs and ligating the blunt ends was performed. (4) DNA was sheared to a size of 300-500bp using Covaris LE220 machine (Covaris) at the conditions; Fill Level:10, Duty Cycle: 15, PIP: 500, Cycles/Burst: 200, Time: 58 seconds. Successful shearing was verified using agarose gel separation. (5) Pulling down biotinylated ligation junctions with streptavidin beads was performed. (6) The libraries were amplified from beads in 200 µl of PCR amplification mastermix (KAPA HiFi Hotstart ReadyMix, and KAPA Primer Mix), divided into 50uL aliquots, and run on the program: 98°C 45 seconds, 98°C 15 seconds, 60°C 30 seconds, 65°C 45 seconds, repeating to second step for 8-12 cycles, then 72°C 5 minutes. (7) Amplified libraries were subjected to AMPure XP bead dual size selection from 0.7x-0.9x. Libraries were sequenced for approximately 350M reads using either Illumina HiSeqX PE150 or NovaSeq6000 S4 v1.5 PE150 platforms. Analysis used 1-3 replicates per condition in matched experimental designs, or comparison of brain tumour groups using 3-4 samples per catergory.

Bulk RNA sequencing. Total RNA was extracted from cell pellets and tumours using the AllPrep DNA/RNA/miRNA Universal Kit (Qiagen) according to instructions from the manufacturer. Library preparation was performed with ribosomal RNA (rRNA) depletion according to instructions from the manufacturer (Epicentre) to achieve greater coverage of mRNA and other long non-coding transcripts. Paired-end sequencing (100 bp) was performed on the Illumina HiSeq 2500 or 4000 platform. Analysis used a minimum of two biological replicates per experimental condition.

Single cell RNA-seq. The concentration of the single-cell suspension was assessed with a Trypan blue count. Approximately 5000 cells per sample were loaded on the Chromium Single Cell 3′ system (10X Genomics). GEM-RT, DynaBeads cleanup, PCR amplification and SPRIselect beads cleanup were performed using Chromium Single Cell 3′ Gel Bead kit. Indexed single-cell libraries were generated using the Chromium Single Cell 3′ Library kit and the Chromium i7 Multiplex kit. Size, quality, concentration and purity of the complementary DNAs and the corresponding 10X library were evaluated by the Agilent 2100 Bioanalyzer system. The 10X libraries were sequenced in the Illumina 2500 sequencing platform.

Droplet Digital PCR. RNA was extracted from cells using the Aurum Total RNA Mini Kit (Bio-Rad). cDNA was generated using iScript Reverse Transcription Supermix (Bio-Rad). Target concentration was determined using the QX200 ddPCR EvaGreen Supermix assay (Bio-Rad) using 20 uL per reaction using 5 ng of cDNA, using manufacturer's protocol cycling conditions with a 58 degrees annealing temperature and 40 cycles. Droplets were assayed using the QX200 Droplet Reader (Bio-Rad) and scored for positive signal using QuantaSoft Software (Bio-Rad). The concentration of positive droplets per target was normalized to the concentration of GAPDH. For quantification of CUT&RUN library enrichment, the above protocol was followed using 1 ng of library DNA. Primers are listed in supplemental information.

Histone mass spectrometry. The complete workflow for histone extraction, LC/MS, and data analysis was previously described.[53,54] Briefly, cell pellets (~1 × 10^6 cells) were lysed, histone precipitated and protein estimated by Bradford assay. Approximately 20 μg of histone extract was then digested with trypsin and a cocktail of isotopically-labeled synthetic histone peptides was spiked in at a final concentration of 250 fmol/μg, followed by propionic anhydride derivatization. nanoLC was performed using a Thermo ScientificTM Easy nLCTM 1000 equipped with a 75 μm × 20 cm in-house packed column using Reprosil-Pur C18-AQ (3 μm; Dr. Maisch GmbH, Germany). Peptides were resolved using a two-step linear gradient from 5 to 33% B over 45 min, then from 33 to 90% B over 10 min at a flow rate of 300 nL/min. The HPLC was coupled online to an Orbitrap Elite mass spectrometer operating in the positive mode using a Nanospray FlexTM Ion Source (Thermo Scientific) at 2.3 kV. Two full MS scans (m/z 300–1100) were acquired in the orbitrap mass analyzer with a resolution of 120,000 (at 200 m/z) every 8 DIA MS/MS events using isolation windows of 50 m/z each (e.g., 300–350, 350–400, ...,650–700). MS/MS spectra were acquired in the ion trap operating in normal mode. Fragmentation was performed using collision-induced dissociation (CID) in the ion trap mass analyzer with a normalized collision energy of 35. Raw files were analyzed using EpiProfile.[55] Analysis used a minimum of two biological replicates per experimental condition.

Western blotting. Histone lysates were extracted using the Histone Extraction kit (Abcam). Lysate protein concentration was determined with the Bradford assay reagent (Bio-Rad). Ten micrograms of protein was separated on SDS-PAGE gels (10% acrylamide) and wet-transferred to a PVDF membrane (GE Healthcare). Membrane blocking was performed with 5% skimmed milk in Tris-buffered saline (50 mM Tris, 150 mM NaCl, 0.1% Tween-20, pH 7.4) (TBST) for 1 hour. Membranes were incubated overnight with primary antibody (GAPDH Advanced ImmunoChemical Inc 2-RGM2 1:1000 dilution, or SOX10 ab212843 1:1000 dilution) in 1% skimmed milk in TBST. Membranes were washed three times in TBST, and the secondary antibody (ECL anti-mouse IgG horseradish peroxidase linked whole antibody) (GE Healthcare) was applied for 1 h in 1% skimmed milk in TBST. Membranes were washed three times and the signal was resolved with Amersham ECL Prime Western Blotting Detection Reagent

(GE Healthcare) and imaged on a ChemiDoc MP Imaging System (Bio-Rad). Analysis used a minimum of two biological replicates per experimental condition.

Immunofluorescence. Cells were plated in a Nunc Lab-Tek II Chamber slide system (ThermoFisher Scientific). Slides were fixed with 4% paraformaldehyde in PBS for 20 min at room temperature, followed by washing three times with PBS. Cells were permeabilized by Triton X-100 (0.05% DIPGXIII, 0.2% BT245), 2% BSA, 5% normal goat serum (NGS) in PBS followed by three PBS washes. Slides were blocked with 5% NGS in PBS for 1 h, followed by overnight incubation with primary antibody (anti-GFAP rabbit monoclonal antibody CST 12389 at 1:200 dilution, or anti-SOX10 ThermoFisher Scientific 703439 at 1:400 dilution) in blocking solution. Cells were washed three times with PBS and incubated for 1 h with 1:1000 dilution of goat anti-rabbit IgG cross-adsorbed secondary antibody, Alexa Fluor 488 (ThermoFisher Scientific) in blocking solution. Slides were washed three times in PBS and Prolong Gold antifade reagent with DAPI (Invitrogen) was applied. Slides were photographed with a Zeiss LSM780 Laser Scanning Confocal Microscope at ×63 magnification. Each image had the protein of interest (SOX10/GFAP) quantified by fluorescence signal normalized to nucleus count value using ImageJ software. Analysis used minimum of two biological replicates and 5 image replicates per experimental condition.

ChIP-seq data processing. Raw sequences were first trimmed using fastp v0.22.0 with default settings before alignment using bwa-mem2 v2.2.1 to a combined reference of hg38+dm6 or mm10+dm6 with default settings.[56,57] After identification of duplicates using picard v2.26.2's "MarkDuplicates" module with default settings, alignments with MAPQ>3 mapping to each species was extracted into separate BAM files using samtools v1.14's "view" module.[58] Alignments overlapping various kinds of genomic intervals (e.g., uniform 10kb windows, promoters, CpG islands) were subsequently tabulated using bedtools v2.30.0's "intersect" module.[59] Depth-normalized coverage tracks were generated using deepTools v3.5.1's bamCoverage module with parameters "--normalizeUsing CPM --centerReads -e 200".[60] Visualization of aggregate ChIP-seq signal around particular genomic regions was performed using the computeMatrix module of deepTools v3.5.1 in "reference-point" mode with parameters "-bs 1000 -b 1000000 -a 1000000 --referencePoint center".

Definition of regions with ChIP-seq enrichment. Regions with focal ChIP-seq enrichment with respect to input controls were identified using MACS v2.2.7.1 with settings "—broad --broad-cutoff 0.1".[61] To avoid biases, H3K27me3-enriched regions in select panels were alternatively defined as the top 1000 CpG islands with the greatest number of H3K27me3 alignments within 5kb from the midpoint; the union set was then taken between diffuse and confined conditions to further reduce bias towards one condition over another. Comparative ChIP-seq enrichment (e.g., differential CBX2 binding) across conditions were assessed using DiffBind v3.4.0.[62]

Quantification of ChIP-seq signal confinement. Parameters capturing experimental ChIP-seq data was extracted using the "learn" module of ChIPs v2.4 with default parameters and H3K27me3 ChIP-seq data as well as peak calls from K27M pHGG cell line BT245.[63] To simulate spreading, peaks intervals were enlarged from summits by specified widths and then used together with previously learned parameters to generate synthetic ChIP-seq datasets using the "simreads" module of ChIPs v2.4. Signal breadth ("confinement") of experimental and simulated ChIP-seq datasets was quantified using the fragment cluster score metric from ssp v1.2.1 with default parameters,[24] which captures the degree of clustering for forward-reverse read pairs with specific genomic separation; for the purposes of quantifying H3K27me3 spread, we opted to focus on a separation distance of 10kb based on the typical width of H3K27me3 enriched regions.

Chromatin-state classification of genomic regions. The union set of GENCODE (v36/vM25) annotated TSSs and mid points of CpG islands from the UCSC Genome Browser were expanded to +/-2.5kb and used as the reference set of regulatory regions for chromatin state classification. The log fold enrichment of ChIP-seq over input for alignments overlapping these intervals were then computed for various targets, producing a table with the rows being promoter / CGI intervals and columns being different ChIP-seq signal sources; this mirrors the expression count matrix of RNA-seq datasets with rows being genes and columns being bulk single cell barcodes or bulk sample identifiers. Analogous to scRNA-seq, dimension reduction was then performed on this data matrix using UMAP with correlation as the distance metric, minimum distance set to 0.01, and neighborhood size varied among (15, 30, 50, 100) depending on the number of datapoints.[64] HDBSCAN was subsequently used for clustering of similar datapoints in UMAP embeddings (of dimensions 5-10) with comparable chromatin states, with minimum cluster and sample sizes varying from (500, 1000, 5000) depending on the number of datapoints.[65] Dimension reduction and clustering was performed iteratively to refine subclusters (e.g., partitioning of all data points into either cluster A and B, and cluster A is then divided into subclusters A1 & A2, and so on).

Hi-C data processing. Raw sequences were processed using Juicer v1.6 with default parameters against hg38 or mm10.[66] Additionally, Juicer .hic files were converted to cooler .mcool files using hic2cool v0.8.3, after which bias vectors were re-computed using the balance module of cooler v0.8.11 with default settings.[67] To additionally avoid confounding effects contributed by cancer samples with abnormal karyotypes, SV-aware bias vectors were calculated using OneD normalization from R library dryhic 0.0.0.9100 with default settings.[68]

Global, compartment, and domain level analysis. Compartment scores were computed using the call-compartments module of cooltools v0.4.1 on 100kb resolution OneD-normalized contact maps with GC content as the phasing track. Boundary scores were computed at 50kb resolution using RobusTAD v1.0 with default parameters, taking max(left, right boundary score.[69] To assess the inter-sample similarity of compartment and boundary scores, genome-wide binned signals per sample were arranged into matrix form (i.e., column = genomic bin, row = sample, value = insulation or compartment score) and used as input for UMAP embedding with correlation as the metric. Global similarity between pairs of Hi-C contact matrices were also directly assessed through HiCRep.py 0.2.6 with "--binSize 50000 --h 5 --dBPMax 5000000", with the resulting correlation coefficients used for distance matrix computation.[70] Silhouette scores were similarly computed from the same data matrix using inter-sample (1 - Pearson's r) as the distance, with the goal of identifying which tumor subtype(s) display highly consistent compartment/insulation signatures (or lack thereof).

Assessment of loop strength. Aggregate peak analysis (APA) / pile-up of interaction profiles was performed using coolpup.py v0.9.5 with default settings to assess the average interaction strength between pairs of genomic intervals such as promoter/CGIs or ChIP-seq peaks.[71] Alternatively, the looping strength of individual pairs of regions (with genomic separation in the range of 20kb-2mb, unless otherwise stated) was measured using chromosight v1.6.1's quantify module at 10kb resolution with default settings, scoring whether pairs of genomic regions contribute more "loop-like" patterns on Hi-C contact maps.[72] For promoter/CGI centric analyses, the average chromosight-measured loop score between a given promoter/CGI and neighboring ones (with genomic separation in the range of 20kb-2mb, unless otherwise stated) is taken to assess measures such as intra-class looping strength (e.g., the strength of interaction among cPRC1 targets).

Differential interactions. Loop score differences between two conditions were calculated based on subtracting chromosight-quantified values for one contact map from another. To independently identify transcriptional regulators whose binding sites strong overlap with differential loop anchors, BART3D v1.0 was ran for Hi-C maps from each of the three isogenic pHGG cell lines comparing K27M vs K27M-KO[73]; the ranking (by significance) of transcriptional regulators most strongly associated with differential interaction was then integrated using RobustRankAggreg v1.1 to finally identify which transcriptional regulators were the most consistently enriched across multiple cell lines.[74] Pile-up plots of differential interactions

Bulk RNA-seq data processing. Raw sequences were trimmed using fastp with default settings, after expression quantification was performed using salmon v1.4.0 with default settings against GENCODE annotations (v36/vM25). Transcript-level counts were collated to genes using tximport to produce gene-level count matrices.[75]

Differential gene expression analysis. DESeq2 v1.34.0 was ran with default settings with gene expression count matrices to identify differentially expressed genes.[76] Gene set enrichment analysis was performed using fgsea v1.20.0 with default settings and taking Wald test statistics as the ranking metric.[77] Gene set over-representation analysis was instead carried out using Enrichr v3.0.[78] Concordance of differential gene expression between different contrasts (e.g., K27M vs K27M-KO as compared to K27M vs K27M + CBX-AM) was evaluated using RRHO2 v1.0 with $-\log_{10}$(p-value) × sign of logFoldChange as the ranking statistic.

scRNA-seq data processing. Cell Ranger (10X Genomics, v3.1.0) was used with default parameters to demultiplex and align sequencing reads, distinguish cells from background, and obtain gene counts per cell. Alignment was performed using a joint hg19+mm10 genome reference build, coupled with Ensembl transcriptome build GRCh37 v.82 for hg19 and GRCm38 v.84 for mm10. Intronic counts were excluded. Human cells were extracted if cells were either assigned as human by Cell Ranger or the cell contained greater than 75% of total reads mapping to hg19 in order to obtain adequate numbers of cells per sample. Quality control and normalization was performed using the R package Seurat (v3.1.0).[79] Cells were filtered based on the following quality control metrics: mitochondrial content (indicative of cellular damage), number of genes and number of unique molecular identifiers (UMIs). Filtering thresholds were set on a per-sample basis where cells were excluded if they had greater than 50% of total reads mapping to mitochondrial read counts, had less than 500 total genes or UMIs, or were outside 2 standard deviations from the mean number of genes or UMIs, respectively. Libraries were scaled to 10,000 UMIs per cell and natural log-normalized. Log normalized counts were used for computing correlations of gene expression and assessing expression of specific genes. Samples were combined by cell line, without any additional transformation of the data.

Identification of nearest normal cell types in xenograft samples. To assign a nearest-normal cell type to individual cells, Spearman correlation of the log-normalized counts with a reference expression matrix was computed inn base R with parameter 'complete.obs' to compute covariances. The reference expression matrix was a developmental murine forebrain and pons single cell atlas with average expression values per cluster, as described previously.[17] For each cell, the cluster label with the highest correlation was assigned as the nearest normal cell type.

Quantification of cPRC1 target expression in single cell data. To assess for enrichment of cPRC1 gene signatures in single cell data, ssGSEA was run to assess enrichment of cPRC1 gene signatures in single cell data,[80] single-sample gene set enrichment analysis (ssGSEA) was run using raw counts per cell and gene sets derived from chromatin-state classification of promoters. Gene sets were derived from chromatin state promoter classification from three separate K27M pHGG cell lines, and only consistently classified promoters were kept (i.e., assigned to the same class in all three lines). ssGSEA code was adapted from the GVSA package using parameters 'alpha = 0.75, normalize = FALSE'.[81] For visualization, proportions were calculated as the fraction of cells of total cells. In cases where only glial cells were visualized, proportions were calculated using fractions of total glial cells.

Visualization. ChIP-seq coverage tracks were imported using rtracklayer and subsequently displayed using ggplot2 v3.3.5.[82,83] Gene annotations were similarly imported and shown through gggenes. Balanced Hi-C matrices were further processed using VEHiCLE with default settings before being imported via RcppCNPy v0.2.10 and similarly visualized using ggplot2.[84,85] 3D structures were predicted from balanced contact matrices using CSynth with default settings.[86] Intersections were shown through Euler diagrams with eulerr v6.1.1.

Statistical consideration. Unless otherwise stated, Wilcoxon rank-sum tests were used to compare the distribution of metrics between two conditions. When appropriate (e.g., matched isogenic cell lines), a paired instead independent test is performed. P-values are represented as: 0 (****) 0.0001 (***) 0.001 (**) 0.01 (*) 0.05 (ns) 1.

Public datasets accessed. H3K27me3 ChIP-seq and Hi-C datasets were sourced from: Bonev 2017 (GSE96107)[11], Gorkin 2020 (ENCODE)[25], McLaughlin 2019 (GSE124342)[23], Conway 2021 (GSE162739)[28], Yusufova 2020 (GSE143293)[27], Won 2016 (GSE77565).[87] K27M and K27M-KO pHGG ChIP-seq and RNA-seq data were partially sourced from: Harutyunyan 2019 (GenAP)[15], Harutyunyan 2020 (GSE147783)[22], Krug 2019 (GSE128745).[21] mESC ChIP-seq data was partially sourced from Chen 2022 (GSE186506).[88] Single cell RNA-seq datasets were sourced from: Jessa 2019 (GSE133531).[17]

Data availability. Sequencing files and bed files are available from the GEO repository; Sequencing data: (GSE205249) and (Tumour Hi-C: GSE186599). Sequencing depth and data quality is described in Supplemental Information. Processed data matrices and genomic tracks are available for browsing at https://cprc1.com:8888.

Code availability. Scripts used for data processing and figure creation has been deposited on GitHub at: https://github.com/bhu/prc1_loops.

References

1 Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171**, 34-57, doi:10.1016/j.cell.2017.08.002 (2017).

2 Loubiere, V., Martinez, A. M. & Cavalli, G. Cell Fate and Developmental Regulation Dynamics by Polycomb Proteins and 3D Genome Architecture. *Bioessays* **41**, e1800222, doi:10.1002/bies.201800222 (2019).

3 Yu, J. R., Lee, C. H., Oksuz, O., Stafford, J. M. & Reinberg, D. PRC2 is high maintenance. *Genes Dev* **33**, 903-935, doi:10.1101/gad.325050.119 (2019).

4 Plys, A. J. *et al.* Phase separation of Polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. *Genes Dev* **33**, 799-813, doi:10.1101/gad.326488.119 (2019).

5 Piunti, A. & Shilatifard, A. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nat Rev Mol Cell Biol* **22**, 326-345, doi:10.1038/s41580-021-00341-1 (2021).

6 Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat Rev Mol Cell Biol* **22**, 815-833, doi:10.1038/s41580-021-00398-y (2021).

7 Finogenova, K. *et al.* Structural basis for PRC2 decoding of active histone methylation marks H3K36me2/3. *Elife* **9**, doi:10.7554/eLife.61964 (2020).

8 Streubel, G. *et al.* The H3K36me2 Methyltransferase Nsd1 Demarcates PRC2-Mediated H3K27me2 and H3K27me3 Domains in Embryonic Stem Cells. *Mol Cell* **70**, 371-379 e375, doi:10.1016/j.molcel.2018.02.027 (2018).

9 Boyle, S. *et al.* A central role for canonical PRC1 in shaping the 3D nuclear landscape. *Genes Dev* **34**, 931-949, doi:10.1101/gad.336487.120 (2020).

10 Loubiere, V., Papadopoulos, G. L., Szabo, Q., Martinez, A. M. & Cavalli, G. Widespread activation of developmental gene expression characterized by PRC1-dependent chromatin looping. *Sci Adv* **6**, eaax4001, doi:10.1126/sciadv.aax4001 (2020).

11 Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e524, doi:10.1016/j.cell.2017.09.043 (2017).

12 Zhao, S., Allis, C. D. & Wang, G. G. The language of chromatin modification in human cancers. *Nat Rev Cancer* **21**, 413-430, doi:10.1038/s41568-021-00357-x (2021).

13 Schwartzentruber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**, 226-231, doi:10.1038/nature10833 (2012).

14 Jain, S. U. *et al.* PFA ependymoma-associated protein EZHIP inhibits PRC2 activity through a H3 K27M-like mechanism. *Nat Commun* **10**, 2146, doi:10.1038/s41467-019-09981-6 (2019).

15 Harutyunyan, A. S. *et al.* H3K27M induces defective chromatin spread of PRC2-mediated repressive H3K27me2/me3 and is essential for glioma tumorigenesis. *Nat Commun* **10**, 1262, doi:10.1038/s41467-019-09140-x (2019).

16 Jain, S. U. *et al.* H3 K27M and EZHIP Impede H3K27-Methylation Spreading by Inhibiting Allosterically Stimulated PRC2. *Mol Cell* **80**, 726-735 e727, doi:10.1016/j.molcel.2020.09.028 (2020).

17 Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nat Genet* **51**, 1702-1713, doi:10.1038/s41588-019-0531-7 (2019).

18 Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* **572**, 67-73, doi:10.1038/s41586-019-1158-7 (2019).

19 Mohammad, F. *et al.* EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas. *Nat Med* **23**, 483-492, doi:10.1038/nm.4293 (2017).

20 Michealraj, K. A. *et al.* Metabolic Regulation of the Epigenome Drives Lethal Infantile Ependymoma. *Cell* **181**, 1329-1345 e1324, doi:10.1016/j.cell.2020.04.047 (2020).

21 Krug, B. *et al.* Pervasive H3K27 Acetylation Leads to ERV Expression and a Therapeutic Vulnerability in H3K27M Gliomas. *Cancer Cell* **35**, 782-797 e788, doi:10.1016/j.ccell.2019.04.004 (2019).

22 Harutyunyan, A. S. *et al.* H3K27M in Gliomas Causes a One-Step Decrease in H3K27 Methylation and Reduced Spreading within the Constraints of H3K36 Methylation. *Cell Rep* **33**, 108390, doi:10.1016/j.celrep.2020.108390 (2020).

23 McLaughlin, K. *et al.* DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. *Cell Rep* **29**, 1974-1985 e1976, doi:10.1016/j.celrep.2019.10.031 (2019).

24 Nakato, R. & Shirahige, K. Sensitive and robust assessment of ChIP-seq read distribution using a strand-shift profile. *Bioinformatics* **34**, 2356-2363, doi:10.1093/bioinformatics/bty137 (2018).

25 Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744-751, doi:10.1038/s41586-020-2093-3 (2020).

26 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

27 Yusufova, N. *et al.* Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. *Nature* **589**, 299-305, doi:10.1038/s41586-020-3017-y (2021).

28 Conway, E. *et al.* BAP1 enhances Polycomb repression by counteracting widespread H2AK119ub1 deposition and chromatin condensation. *Mol Cell* **81**, 3526-3541 e3528, doi:10.1016/j.molcel.2021.06.020 (2021).

29 Barbour, H., Daou, S., Hendzel, M. & Affar, E. B. Polycomb group-mediated histone H2A monoubiquitination in epigenome regulation and nuclear processes. *Nat Commun* **11**, 5947, doi:10.1038/s41467-020-19722-9 (2020).

30 Lu, C. *et al.* Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science* **352**, 844-849, doi:10.1126/science.aac7272 (2016).

31 Papillon-Cavanagh, S. *et al.* Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas. *Nat Genet* **49**, 180-185, doi:10.1038/ng.3757 (2017).

32 Kurotaki, N. *et al.* Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat Genet* **30**, 365-366, doi:10.1038/ng863 (2002).

33 Tauchmann, S. & Schwaller, J. NSD1: A Lysine Methyltransferase between Developmental Disorders and Cancer. *Life (Basel)* **11**, doi:10.3390/life11090877 (2021).

34 Nagano, M. H., B.; Yokobayashi, S.; Majewski, J.; Saitou, M. Nucleome programming is required for the foundation of totipotency in mammalian germline development. *EMBO Journal*, doi:doi:10.15252/embj.2022110600 (2022).

35 Blanco, E., Gonzalez-Ramirez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends Genet* **36**, 118-131, doi:10.1016/j.tig.2019.11.004 (2020).

36 Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331-335, doi:10.1126/science.aao4750 (2018).

37 Huseyin, M. K. & Klose, R. J. Live-cell single particle tracking of PRC1 reveals a highly dynamic system with low target site occupancy. *Nat Commun* **12**, 887, doi:10.1038/s41467-021-21130-6 (2021).

38 Lamb, K. N. *et al.* Discovery and Characterization of a Cellular Potent Positive Allosteric Modulator of the Polycomb Repressive Complex 1 Chromodomain, CBX7. *Cell Chem Biol* **26**, 1365-1379 e1322, doi:10.1016/j.chembiol.2019.07.013 (2019).

39 Pathania, M. *et al.* H3.3(K27M) Cooperates with Trp53 Loss and PDGFRA Gain in Mouse Embryonic Neural Progenitor Cells to Induce Invasive High-Grade Gliomas. *Cancer Cell* **32**, 684-700 e689, doi:10.1016/j.ccell.2017.09.014 (2017).

40 Youmans, D. T., Gooding, A. R., Dowell, R. D. & Cech, T. R. Competition between PRC2.1 and 2.2 subcomplexes regulates PRC2 chromatin occupancy in human stem cells. *Mol Cell* **81**, 488-501 e489, doi:10.1016/j.molcel.2020.11.044 (2021).

41 Lau, M. S. *et al.* Mutation of a nucleosome compaction region disrupts Polycomb-mediated axial patterning. *Science* **355**, 1081-1084, doi:10.1126/science.aah5403 (2017).

42 Blackledge, N. P. *et al.* PRC1 Catalytic Activity Is Central to Polycomb System Function. *Mol Cell* **77**, 857-874.e859, doi:10.1016/j.molcel.2019.12.001 (2020).

43 Fursova, N. A. *et al.* Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression. *Mol Cell* **74**, 1020-1036.e1028, doi:10.1016/j.molcel.2019.03.024 (2019).

44 Sneppen, K. & Ringrose, L. Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent. *Nat Commun* **10**, 2133, doi:10.1038/s41467-019-10130-2 (2019).

45 Cohen, A. S. *et al.* Weaver Syndrome-Associated EZH2 Protein Variants Show Impaired Histone Methyltransferase Function In Vitro. *Hum Mutat* **37**, 301-307, doi:10.1002/humu.22946 (2016).

46 Tatton-Brown, K. & Rahman, N. The NSD1 and EZH2 overgrowth genes, similarities and differences. *Am J Med Genet C Semin Med Genet* **163C**, 86-91, doi:10.1002/ajmg.c.31359 (2013).

47 Brennan, K. *et al.* NSD1 mutations deregulate transcription and DNA methylation of bivalent developmental genes in Sotos syndrome. *Hum Mol Genet*, doi:10.1093/hmg/ddac026 (2022).

48 Wang, G. G., Cai, L., Pasillas, M. P. & Kamps, M. P. NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis. *Nat Cell Biol* **9**, 804-812, doi:10.1038/ncb1608 (2007).

49 Parreno, V., Martinez, A. M. & Cavalli, G. Mechanisms of Polycomb group protein function in cancer. *Cell Res* **32**, 231-253, doi:10.1038/s41422-021-00606-6 (2022).

50 Ringrose, L. & Paro, R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* **38**, 413-443, doi:10.1146/annurev.genet.38.072902.091907 (2004).

51 Weinberg, D. N. *et al.* The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* **573**, 281-286, doi:10.1038/s41586-019-1534-3 (2019).

52 Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc* **1**, 729-748, doi:10.1038/nprot.2006.98 (2006).

53 Karch, K. R., Sidoli, S. & Garcia, B. A. Identification and Quantification of Histone PTMs Using High-Resolution Mass Spectrometry. *Methods Enzymol* **574**, 3-29, doi:10.1016/bs.mie.2015.12.007 (2016).

54 Sidoli, S., Bhanu, N. V., Karch, K. R., Wang, X. & Garcia, B. A. Complete Workflow for Analysis of Histone Post-translational Modifications Using Bottom-up Mass Spectrometry: From Histone Extraction to Data Analysis. *J Vis Exp*, doi:10.3791/54112 (2016).

55 Yuan, Z. F. *et al.* EpiProfile 2.0: A Computational Platform for Processing Epi-Proteomics Mass Spectrometry Data. *J Proteome Res* **17**, 2533-2541, doi:10.1021/acs.jproteome.8b00133 (2018).

56 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

57 Vasimuddin, M., Misra, S., Li, H. & Aluru, S. in *IEEE IPDPS* 314-324 (2019).

58 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, doi:10.1093/gigascience/giab008 (2021).

59 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

60 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

61 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

62 Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393, doi:10.1038/nature10730 (2012).

63 Zheng, A. *et al.* A flexible ChIP-sequencing simulation toolkit. *BMC Bioinformatics* **22**, 201, doi:10.1186/s12859-021-04097-5 (2021).

64 McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *JOSS* **3**, doi:10.21105/joss.00861 (2018).

65 McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *JOSS* **2**, doi:10.21105/joss.00205 (2017).

66 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

67 Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311-316, doi:10.1093/bioinformatics/btz540 (2020).

68 Vidal, E. *et al.* OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Research* **46**, e49-e49, doi:10.1093/nar/gky064 (2018).

69 Dali, R., Bourque, G. & Blanchette, M. RobusTAD: A Tool for Robust Annotation of Topologically Associating Domain Boundaries. *bioRxiv*, 293175, doi:10.1101/293175 (2018).

70 Lin, D., Sanders, J. & Noble, W. S. HiCRep.py: fast comparison of Hi-C contact matrices in Python. *Bioinformatics* **37**, 2996-2997, doi:10.1093/bioinformatics/btab097 (2021).

71 Flyamer, I. M., Illingworth, R. S. & Bickmore, W. A. Coolpup.py: versatile pile-up analysis of Hi-C data. *Bioinformatics* **36**, 2980-2985, doi:10.1093/bioinformatics/btaa073 (2020).

72 Matthey-Doret, C. *et al.* Computer vision for pattern detection in chromosome contact maps. *Nat Commun* **11**, 5795, doi:10.1038/s41467-020-19562-7 (2020).

73 Wang, Z., Zhang, Y. & Zang, C. BART3D: Inferring transcriptional regulators associated with differential chromatin interactions from Hi-C data. *Bioinformatics*, doi:10.1093/bioinformatics/btab173 (2021).

74 Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573-580, doi:10.1093/bioinformatics/btr709 (2012).

75 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

76 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

77 Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv*, doi:10.1101/060012 (2021).

78 Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90, doi:10.1002/cpz1.90 (2021).

79 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

80 Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112, doi:10.1038/nature08460 (2009).

81 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7, doi:10.1186/1471-2105-14-7 (2013).

82 Wickham, H. *ggplot2*. (Springer, Cham, 2016).

83 Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-1842, doi:10.1093/bioinformatics/btp328 (2009).

84 Eddelbuettel, D. & Wu, W. RcppCNPy: Read-Write Support for NumPy Files in R. *JOSS* **1**, doi:10.21105/joss.00055 (2016).

85 Highsmith, M. & Cheng, J. VEHiCLE: a Variationally Encoded Hi-C Loss Enhancement algorithm for improving and generating Hi-C data. *Sci Rep* **11**, 8880, doi:10.1038/s41598-021-88115-9 (2021).

86 Todd, S. *et al.* CSynth: an interactive modelling and visualization tool for 3D chromatin structure. *Bioinformatics* **37**, 951-955, doi:10.1093/bioinformatics/btaa757 (2021).

87 Morrissy, A. S. *et al.* Divergent clonal selection dominates medulloblastoma at recurrence. *Nature* **529**, 351-357, doi:10.1038/nature16478 (2016).

88 Chen, H. *et al.* H3K36 dimethylation shapes the epigenetic interaction landscape by directing repressive chromatin modifications in embryonic stem cells. *Genome Res* **32**, 825-837, doi:10.1101/gr.276383.121 (2022).

Acknowledgements

This work was supported by a Large-Scale Applied Research Project grant from Genome Quebec, Genome Canada, the Government of Canada, and Ministère de l'Économie, de la Science et de l'Innovation du Québec, with the support of the Ontario Research Fund through funding provided by the Government of Ontario to N.J., J.M. and C.L.K.; the Canadian Institutes for Health Research (CIHR; grants PJT-156086 to C.L.K., and MOP-286756 and FDN-154307 to N.J.); the National Sciences and Engineering Research Council (NSERC; grant RGPIN-2016-04911 to C.L.K.). J.M. and N.J. is supported by funding from the United States National Institutes of Health (NIH) Grant P01-CA196539; C. L. is supported by funding from NIH Grant R35GM138181 and the Pew-Stewart Scholars for Cancer Research Award; B. A. G. is supported by NIH grants CA196539 and AI118891, and the St. Jude Children's Hospital Chromatin Consortium; B. H. is supported by the Canadian Institutes of Health Research (CIHR) Banting and Best Graduate Scholarship. C.L.K. is supported by a salary award from Fonds de recherche du Québec (FRQS). N.J. is a member of the Penny Cole Laboratory and holds a Canada Research Chair Tier 1 in Pediatric Oncology from CIHR. This work was performed within the context of the International Childhood Astrocytoma Integrated Genomic and Epigenomic (ICHANGE) consortium. Data analyses were enabled by compute and storage resources provided by Compute Canada and Calcul Québec. We are especially grateful for the generous philanthropic donation of the Charles Bruneau Foundation, the WeLoveYouConnie Foundation, and the Cedars/Sarah Cook funds.

Author contributions

B. K. and B. H. generated data and contributed to study design, data interpretation and manuscript preparation. H. C., X. C., N. K., A. B., M. J. J., R. D. and M. B contributed to sequencing data analysis and interpretation. S. D., K. H. G., A. F. A., E. J., A. H., J. J. Y. L., M. H., D. F., C. R and X. X. contributed to data generation, analysis and interpretation. N. A. D., A. G. W. and B. E. assisted with the collection of patient samples, study design and data interpretation. K. W. and B. A. G. led the histone proteomics data generation and analysis. M. G., M. D. T., C. L. K., J. M., N. J and C. L. contributed to study design, data interpretation, and manuscript preparation.

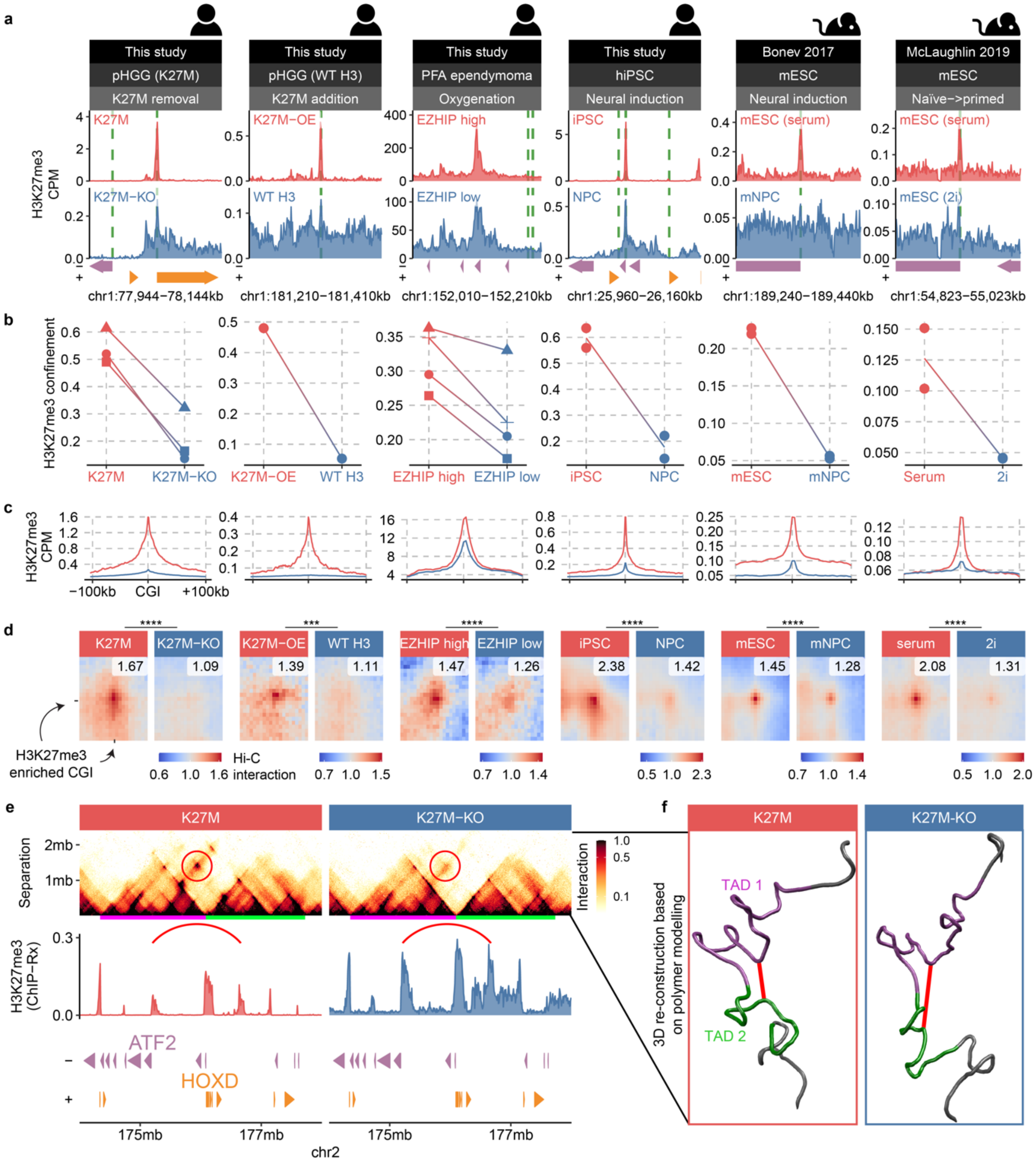
Competing interests

The authors declare no competing interests.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Nada Jabado (nada.jabado@mcgill.ca).

Figure 3.1. Restricted H3K27me3 occurs in brain tumor and developmental contexts and associates with stronger distal interactions between CpG islands



a. Genomic distribution of H3K27me3 (ChIP-seq coverage tracks in units of counts-per-million-alignments) at representative loci in brain tumor and developmental_contexts demonstrating distinction of confined versus diffuse profiles. Focal H3K27me3 enrichment preferentially occurs near regulatory regions such as promoters and CpG islands.

b. Measure of H3K27me3 ChIP-seq signal confinement (fragment cluster score at 1kb separation, computed using the tool "ssp", see methods) in diverse contexts confirming genome-wide distinction of confined versus diffuse profiles. Individual data points correspond to a replicate, with connected points indicating replicates from the same batch; connections not linking points indicate that multiple replicates were sequenced in a batch, and so the links are drawn between the average value per condition.

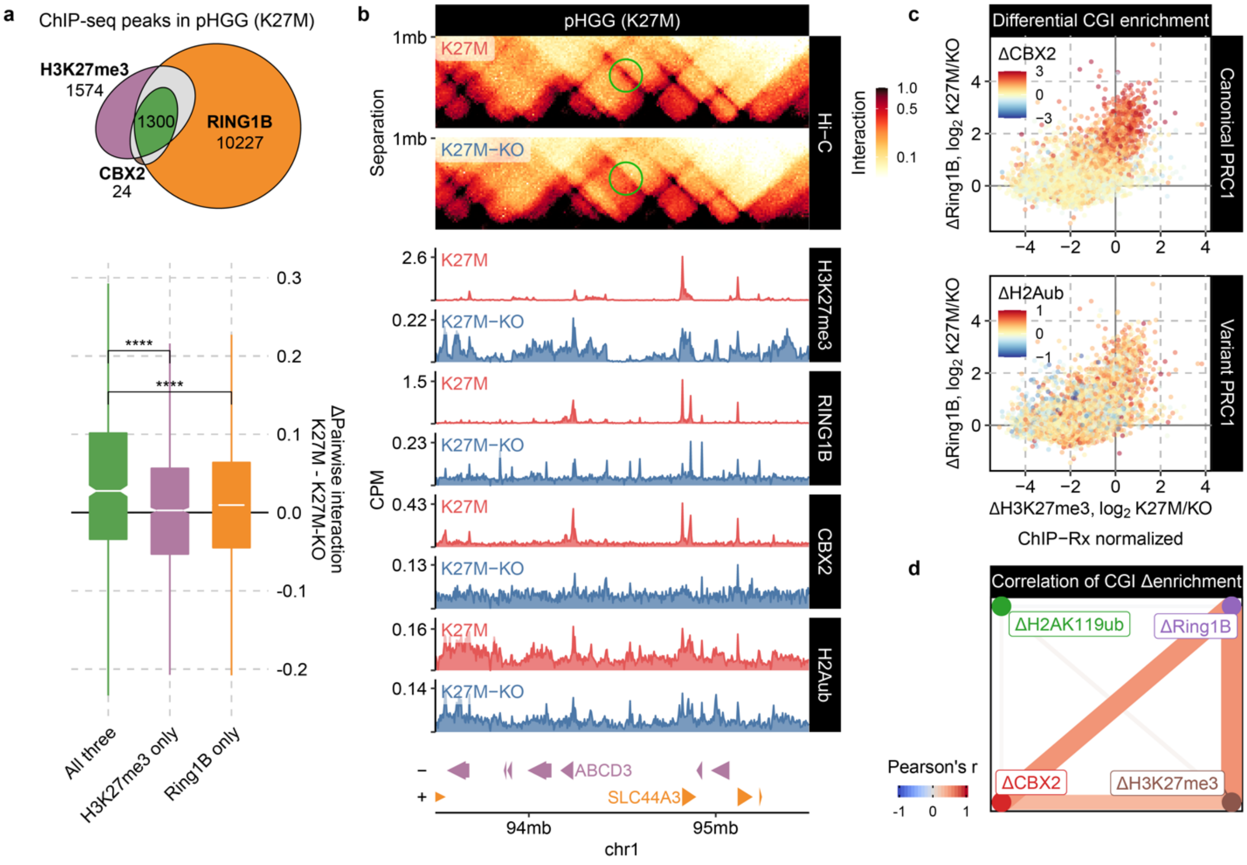
c. Metaplots of H3K27me3 aggregate ChIP-seq signals around H3K27me3-enriched CpG islands, normalized by total read depth. H3K27me3-enriched is defined as the union set of top 1000 CpG islands with the most H3K27me3 alignments in either condition.

d. Pile-up of Hi-C interaction among H3K27me3-enriched CpG islands, as defined in c., portraying average pairwise contact strength between such regions (in units of enrichment, i.e., observed / expected). Punctate enrichment signal in the center indicates elevated long-range interaction anchored at H3K27me3-enriched CGIs in cells with confined H3K27me3.

e. A representative locus in pHGG cell line BT245 (K27M versus K27M-KO) demonstrating the correspondence between confined H3K27me3 with elevated long-range interaction between H3K27me3-enriched regions. These distal loops can span megabases and across TAD boundaries, in this case disregarding the insulating HOXD cluster. The two separate domains are underlined in green and purple.

f. CSynth polymer simulation modelling the Hi-C contact maps shown in (f), visualizing the impact of H3K27me3 confinement in connecting distal chromatin segments with H3K27me3-enrichment, colored in red, from two otherwise insulated domains, here colored in green and purple.

Figure 3.2. Confined H3K27me3 induces cPRC1 concentration leading to compaction of Polycomb bodies in brain tumors



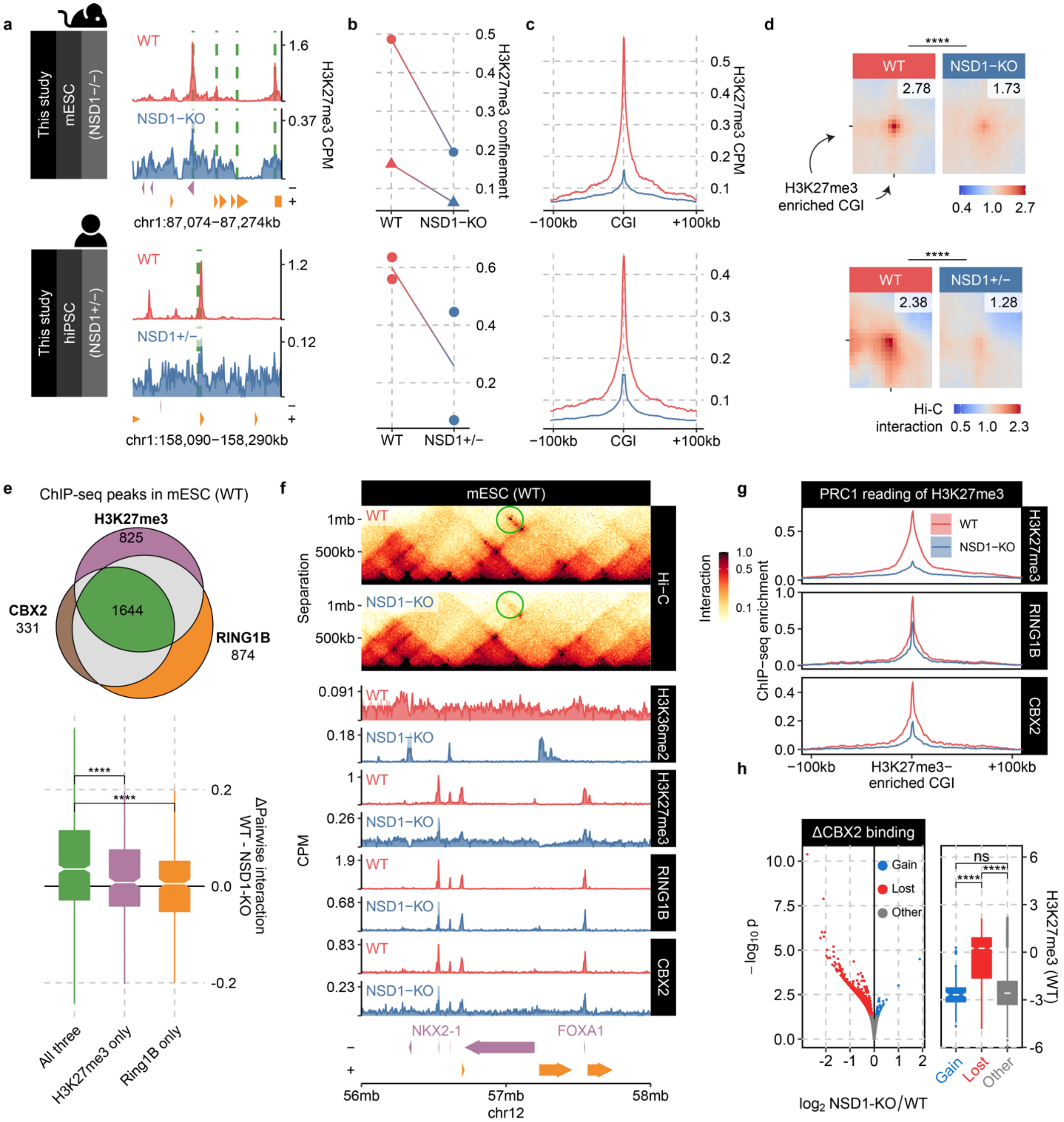
a. (top) Euler diagram of peak call at CGIs for H3K27me3 and PRC1 sub-units CBX2 & RING1B in K27M pHGG cell line BT245, confirming CBX2's reader function in localizing cPRC1 to H3K27me3-enriched regions; (bottom) differential long-range interaction strength (loop score computed for pairs of regions within 20kb-2mb) for various peak overlap subsets, revealing that sites marked by all three (H3K27me3, RING1B, CBX2) preferentially engage in strong distal interaction in K27M. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

b. Representative locus of differential interaction between cPRC1 binding sites bridging the promoters of ABCD3 and SLC44A3, along with ChIP-seq profiles of H3K27me3, RING1B, CBX2 and H2AK119ub. Whereas H3K27me3, RING1B, and CBX2 becomes demonstrably more diffuse in K27M-KO, H2AK119ub profile remains largely unchanged

c. log2 ratio of ChIP-seq alignments in CGIs between K27M and K27M-KO BT245 cells reveals coordinated change among H3K27me3, CBX2, and RING1B, indicating robust correlation between H3K27me3 confinement with enhanced cPRC1 recruitment (top). In contrast, coloring by differential H2AK119ub instead of CBX2 reveals lack of correlation with vPRC1 (bottom).

d. Correlation network of differential H3K27me3, RINGB1, CBX2 and H2AK119ub enrichment at CGIs, demonstrating the weak correlation between H2AK119ub changes and those of the other three, implicating cPRC1 rather than vPRC1 as the defining feature associated with H3K27me3 differences segregating K27M from K27M-KO. Edgewidths reflect the absolute value Pearson correlation coefficients.

Figure 3.3. Loss of NSD1 induces H3K27me3 spreading and dissolution of Polycomb bodies in PSCs



a. Genomic distribution of H3K27me3 (ChIP-seq coverage tracks in units of counts-per-million-alignments) at representative loci in WT and NSD1+/- or NSD1-/- PSCs, demonstrating distinction of confined versus diffuse profiles. Focal H3K27me3 enrichment preferentially occurs near regulatory regions such as promoters and CpG islands.

b. Measure of H3K27me3 ChIP-seq signal confinement (fragment cluster score at 1kb separation, computed using the tool "ssp", see methods) in diverse contexts confirming genome-wide distinction of confined versus diffuse profiles. Individual data points correspond to a replicate, with connected points indicating replicates from the same batch; connections not linking points indicate that multiple replicates were sequenced in a batch, and so the links are drawn between the average value per condition.

c. Metaplots of H3K27me3 aggregate ChIP-seq signals around H3K27me3-enriched CpG islands, normalized by total read depth. H3K27me3-enriched is defined as the union set of top 1000 CpG islands with the most H3K27me3 alignments in either condition.

d. Pile-up of Hi-C interaction among H3K27me3-enriched CpG islands, as defined in c., portraying average pairwise contact strength between such regions (in units of enrichment, i.e., observed / expected). Punctate enrichment signal in the center indicates elevated long-range interaction anchored at H3K27me3-enriched CGIs in cells with confined H3K27me3.

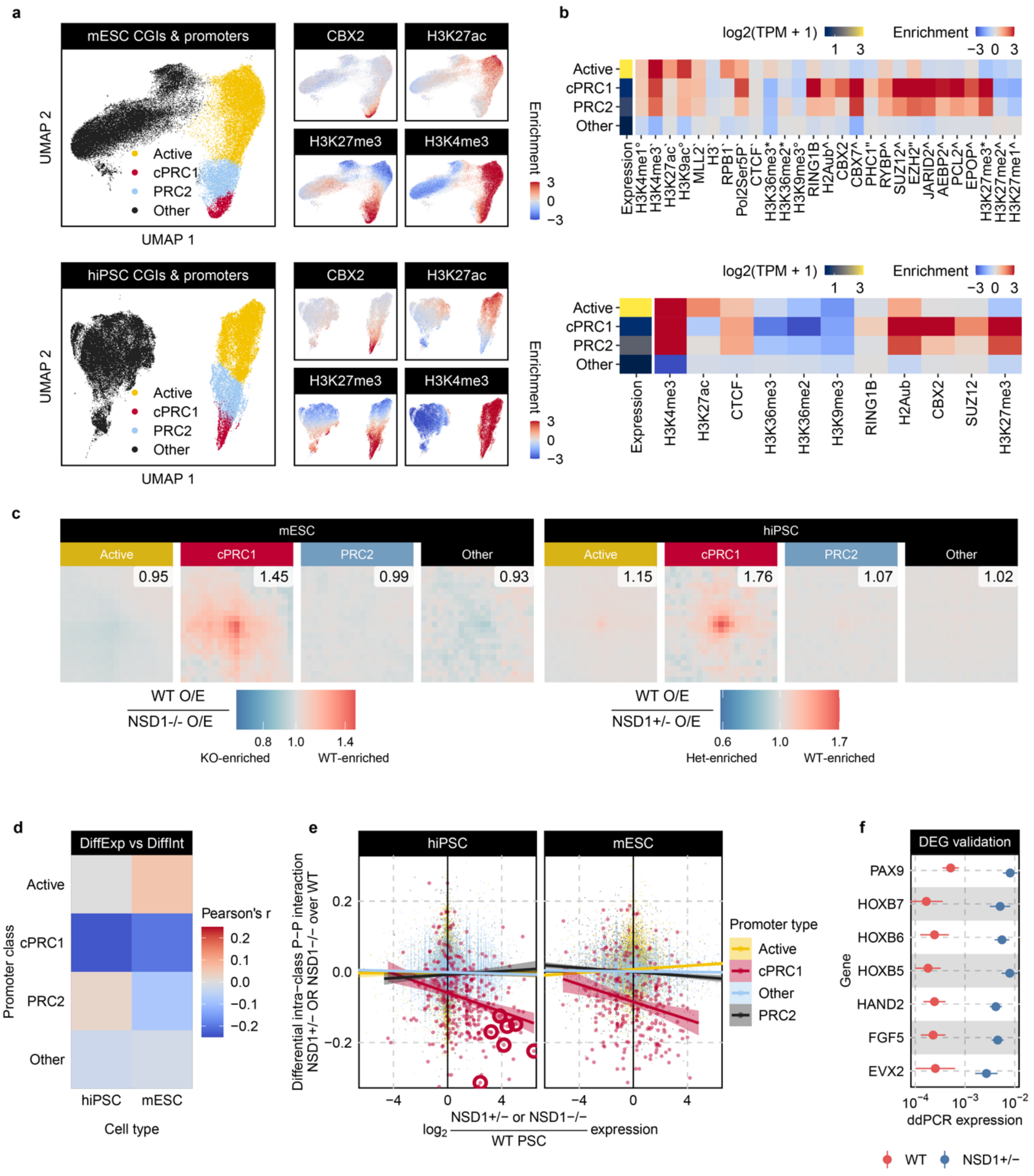
e. (top) Euler diagram of peak call at CGIs for H3K27me3 and PRC1 sub-units CBX2 & RING1B in WT mESCs in serum, confirming CBX2's reader function in localizing cPRC1 to H3K27me3-enriched regions in PSCs; (bottom) differential long-range interaction strength (loop score computed for pairs of regions within 20kb-2mb) for various peak overlap subsets, revealing that sites marked by all three (H3K27me3, RING1B, CBX2) preferentially engage in strong distal interaction in WT mESCs as compared to NSD1-KO cells. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

f. Representative locus of differential interaction between cPRC1 binding sites bridging the promoters of NKX2-1 and FOXA1, along with ChIP-seq profiles of H3K36me2, H3K27me3, RING1B, and CBX2. Whereas H3K27me3 spread accompanies H3K36me2-depletion in NSD1-KO, PRC1 binding also becomes demonstrably more diffuse

g. Metaplot of H3K27me3 and PRC1 aggregate ChIP-seq signal around H3K27me3-enriched CpG islands, in units of log2 enrichment over input, confirming NSD1-KO reduces occupancy of cPRC1 at H3K27me3-enriched CGIs (union set of top 1000 most enriched in both conditions, as defined previously).

h. (left) Differential binding analysis on the union set of CBX2 peaks across both WT and NSD1-KO cells confirms sweeping loss of CBX2 binding upon NSD1-KO. (right) CBX2 loss overwhelmingly takes place in regions with higher H3K27me3 in WT, validating CBX2's role as a reader of H3K27me3 in our system and explaining its subsequent dilution upon H3K27me3 spread.

Figure 3.4. Compaction of Polycomb bodies is associated with the repression of target genes



a. UMAP embedding and HDBSCAN clustering of epigenetic signal at CpG island & promoters (as listed in panel b) in WT hiPSC (NCRM1, from this study) and mESC (a combination of public and data from this study). Individual data points correspond to a genomic interval (promoter or CpG island), and the embedding is based on dimension reduction of features listed in panel b. Four different clusters/classes of regions are discovered: Active (enriched for H3K4me3 and H3K27ac), cPRC1 (with both PRC1 and PRC2 marks), PRC2 (with SUZ12 and H3K27me3, lacking CBX2), and Other (no specific enrichment for particular targets).

b. Average signals of transcription and epigenetic features for CGIs & promoters in each of the four clusters, demonstrating the characteristic chromatin state of each accordingly labelled class. Symbols indicate data sources: * = Chen 2022, \ = Kundu 2017, ^ = Healy 2019, ` = Mas 2018, ° = ENCODE, no symbol = this study.

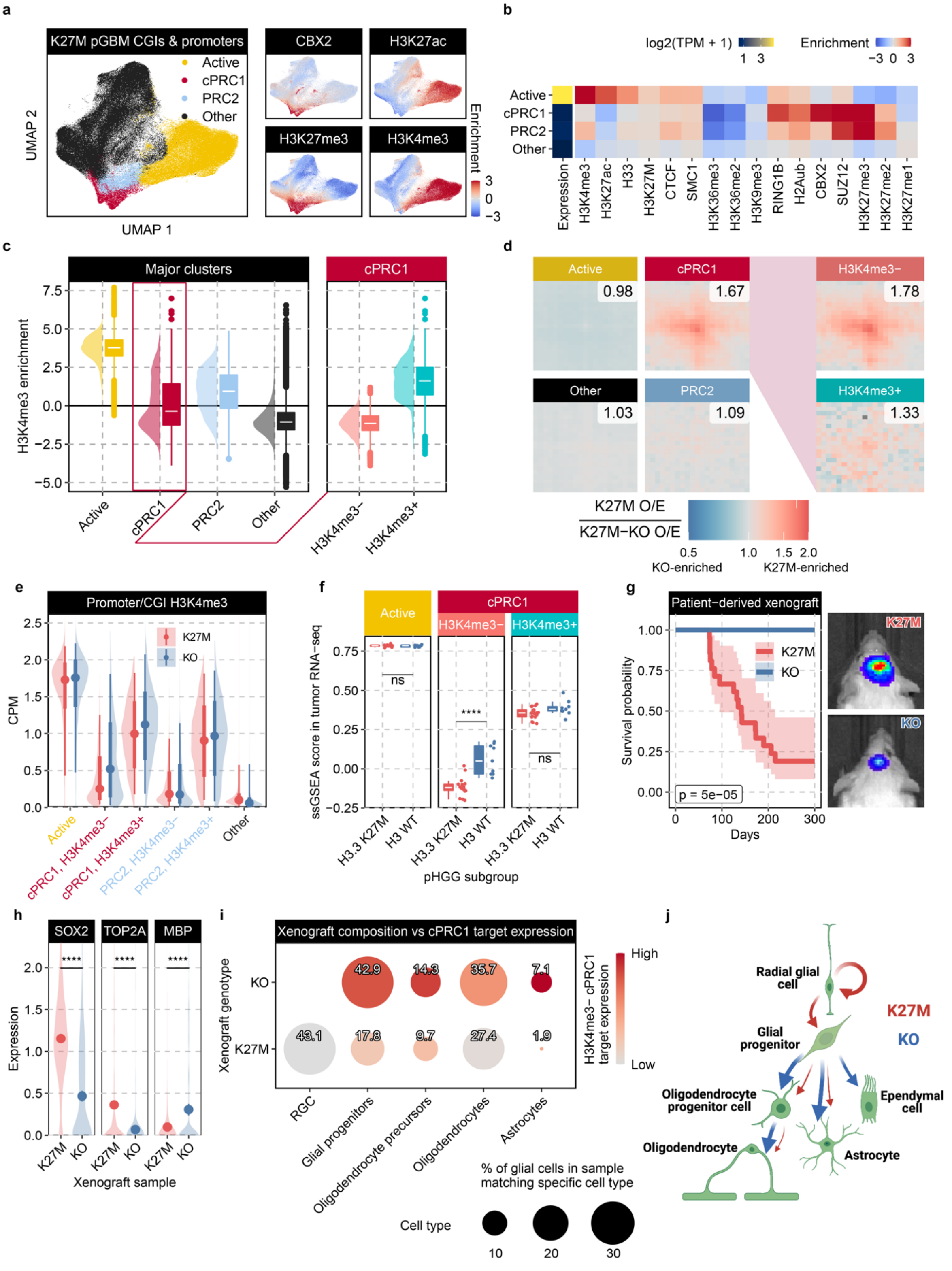
c. Pile-up of Hi-C interaction were computed among pairs of genomic regions belonging to the same cluster (i.e., intra-class looping), after which the ratio of Hi-C contact enrichment between WT and NSD1+/- or NS1-/- cells are displayed. On average, cPRC1 sites demonstrate the strongest signal of differential intra-class looping.

d. Correlation between differential gene expression and differential intra-class promoter-promoter interactions (loop score of each promoter with every other promoter within 2mb is averaged), comparing WT and NSD1+/- or NSD1-/- PSCs. Only cPRC1 target genes demonstrate visible dependency of differential expression on differential looping, with loss of cPRC1 loops driving de-repression.

e. Individual data points underlying panel d along with robust linear regression lines (with 95% confidence intervals in a lighter shade) confirming only cPRC1 targets exhibit negative relationship between differential interaction and expression. Circled genes were selected for validation.

f. ddPCR validation of select differentially expressed cPRC1 targets as marked in panel e. Points describe mean expression and intervals delineate standard error.

Figure 3.5. Canonical PRC1 targets H3K27M-repressed genes in primary brain tumors and patient-derived xenografts



a. UMAP embedding and HDBSCAN clustering of epigenetic signal at CpG island & promoters (as listed in panel b) in three separate K27M pHGG cell lines (BT245, DIPGXIII, and HSJ019, all data from this study). Individual data points correspond to a genomic interval (promoter or CpG island), and the embedding is based on dimension reduction of features listed in panel b. Four different clusters/classes of regions are discovered: Active (enriched for H3K4me3 and H3K27ac), cPRC1 (with both PRC1 and PRC2 marks), PRC2 (with SUZ12 and H3K27me3, lacking CBX2), and Other (no specific enrichment for particular targets).

b. Average signals of transcription and chromatin features for CGIs & promoters in different classes, demonstrating the characteristic chromatin state of each accordingly labelled class.

c. (left) H3K4me3 enrichment distribution is visualized for regions belonging to the four different classes, with a bimodal distribution of H3K4me3 enrichment only found in the cPRC1 cluster; (right) additionally sub-clustering distinguished sites with higher H3K4me3 (H3K4me3+) from those depleted of H3K4me3 (H3K4me3-). Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

d. Pile-up of Hi-C interaction were computed among pairs of genomic regions belonging to the same cluster (i.e., intra-class looping), after which the ratio of Hi-C contact enrichment between K27M and K27M-KO BT245 cells are displayed. On average, cPRC1 sites demonstrate the strongest signal of differential intra-class looping, with H3K4me3- sites showing greater preferential K27M-enrichment as compared to H3K4me3+ sites.

e. Comparison of H3K4me3 signal between K27M pHGG cell line BT245 with its isogenic K27M-KO counterpart, revealing the greatest gain of H3K4me3 upon K27M-KO taking place in cPRC1 H3K4me3- sites. Points indicate median value, with a thicker band describing the 66th percentile, whereas the thin line extends to 95th percentile.

f. Non-bivalent cPRC1 (i.e., H3K4me3-) targets identified in cell line models are consistently more repressed in bulk RNA-seq datasets of primary patient tumors, unlike bivalent cPRC1 (i.e., H3K4me3+) targets.

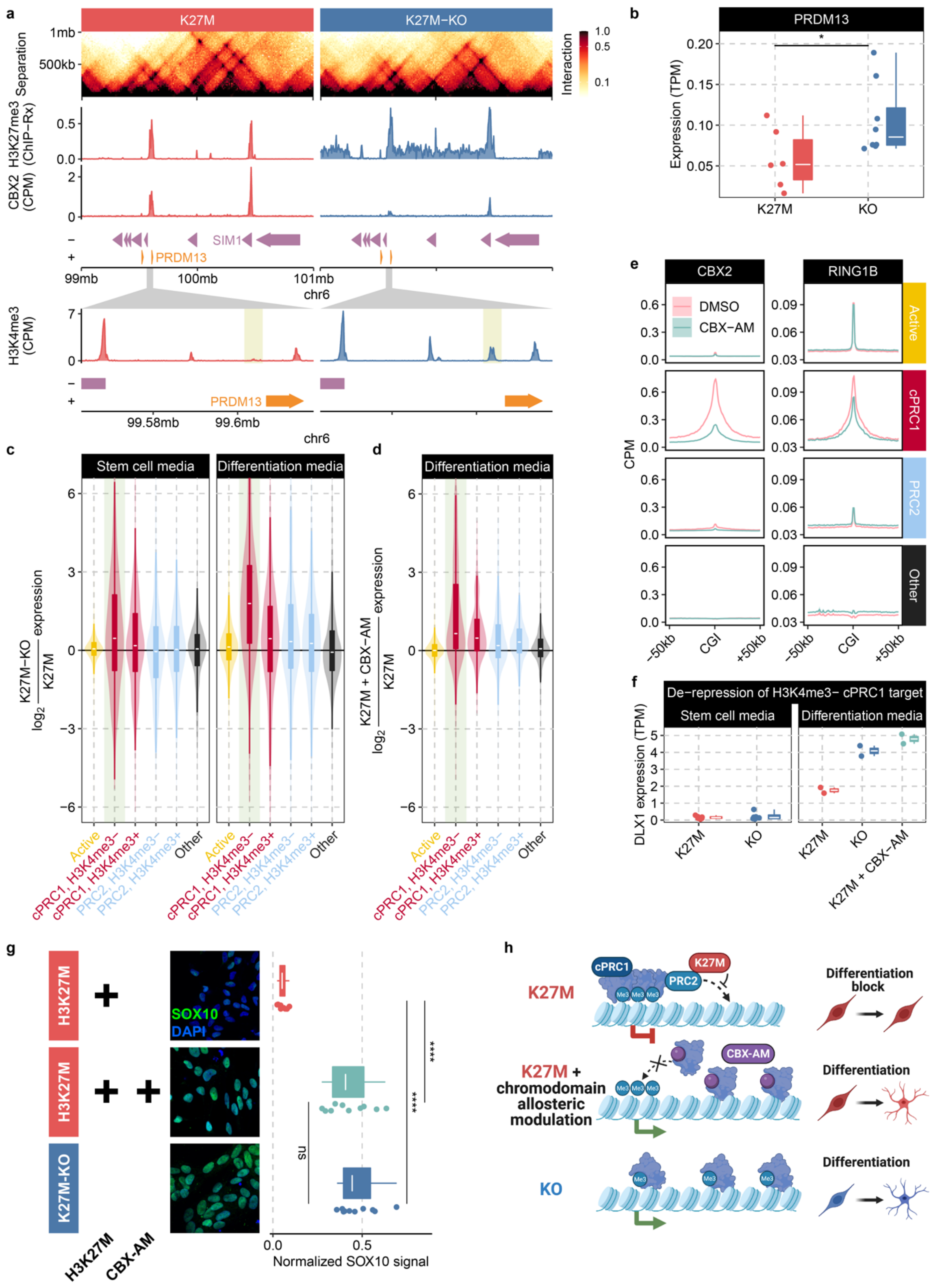
g. Patient-derived orthotopic xenograft model recapitulates tumorigenicity of K27M but not K27M-KO pHGG cells (BT245). (Left) Kaplan-Meier survival analysis (with 95% confidence interval in a lighter shade) for engrafted mice, (right) *in vivo* imaging of tumor cells confirming greater luciferase signal for K27M, localized to the caudate putamen injection site.

h. K27M cells in xenografts consistently show higher expression of stemness markers SRY-box 2 (SOX2) and DNA Topoisomerase II Alpha (TOP2A) and lower expression of oligodendrocyte differentiation marker Myelin Basic Protein (MBP). Dots indicate mean expression across single cells.

i. Comparison of scRNA-seq from K27M and K27M-KO PDOXs confirm that up-regulation of H3K4me3- cPRC1 targets is accompanied by an increase in the proportion of more differentiated cell types in K27M-KO.

j. Developmental trajectory of neural cell types related to gliomagenesis, and the relative cell type transitions characterizing K27M and KO xenografts.

Figure 3.6. Upregulation of cPRC1 target genes upon the removal of cPRC1-mediated chromatin compaction



a. Representative loci of K27M-specific cPRC1-associated loop in K27M pHGG cell line BT245 demonstrating long-range looping between inactive cPRC1 target sites (depleted of H3K4me3) in K27M that becomes de-repressed and gains H3K4me3 at the PR-SET Domain 13 (PRDM13) gene promoter in KO cells (green shading).

b. Modest transcriptomic consequences for the PRDM13 representative cPRC1 target gene shown in a. while cultured in stem cell media. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

c. Differential expression, specifically of targets in cPRC1, -H3K4me3 cluster (green shading), become heightened after differentiation compared to stem cell media.

d. Similar transcriptional de-repression of cPRC1 targets is recapitulated via the application of a CBX allosteric modulator to K27M cells, demonstrating the link between cPRC1-mediated chromatin compaction and repression of cPRC1 targets.

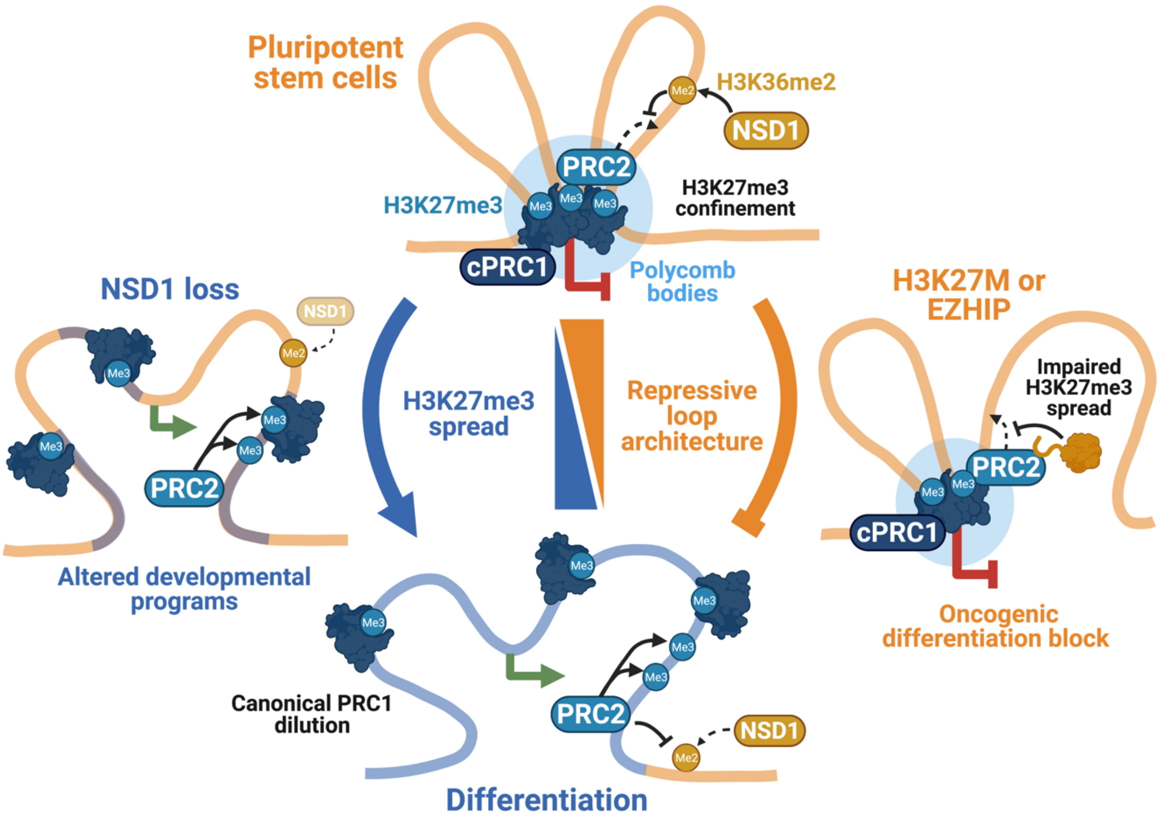
e. CBX-AM treatment specifically attenuates the binding of PRC1 at cPRC1 target sites while leaving RING1B enrichment equivalent at active promoters.

f. Distal-Less Homeobox 1 (DLX1) as a representative example neurodevelopmental cPRC1 target gene that becomes de-repressed upon treatment with CBX-AM or removal of K27M.

g. Up-regulation of differentiation markers observed in both K27M-KO and CBX-AM treated K27M cells in pHGG line BT245.

h. Summary of CBX-AM treatment and K27M-KO's effect on H3K27me3, chromodomain localization and the differentiation of K27M pHGG cells.

Figure 3.7. Model of the link between H3K27me3 spread and cPRC1 chromatin architecture leading to altered developmental programs



Our understanding of the epigenome and its normal and abnormal regulation is going to increase at the present rapid rate. The fundamental knowledge emerging is going to provide ever increasing ramifications for translational science in areas such as cancer. New targeting strategies are going to allow us to build on already exciting indications that epigenetic therapy could provide a tremendous component for cancer therapy and for other diseases as well such as neurodegenerative disorders, diabetes, and so on.

Stephen B Baylin

Chapter 4

Discussion

Panorama of chromatin re-organization in health and disease

The preceding chapters outline molecular portraits of embryogenesis with an emphasis on the germline as well as of diseases associated with abnormal developmental progression. Ascertainment of the very same 3D epigenomic factors and processes as essential components of both healthy development and pediatric afflictions underscores universal regulatory mechanisms whose elucidation will carry far-reaching implications. Although strategies such as genome-wide CRISPR screens can provide tremendous value in narrowing down the scope of disease-relevant genes,[68] elucidation of the mechanistic cascade is no less important for moving towards the ultimate aim of uncovering therapeutically actionable insights. We've shown that a chromatin-centric perspective is especially apt for the integration of multi-omics datasets to unite upstream genetic alterations with downstream phenotypic outcomes as well as to pinpoint not only the crucial determinants of unimpeded and healthy differentiation themselves, but also the means through which they act; and this feat was only made possible by fully capitalizing on the complementary information revealed by orthogonal assays targeting various aspects of cellular activity such as different facets of the epigenome, nucleome, and transcriptome. Through detailed dissection of biological contexts undergoing dramatic 3D epigenome remodelling, whether it be to the benefit or detriment of the system, we unveiled unexpected modes of genome regulation that bear significant relevance to understanding pathogenesis. Our systematic analysis of global 3D epigenomic features, both within and across omics layers, charts an illuminating path in navigating the sea of high-dimensional multi-omics datasets.

Filling in the gap of mammalian germline nucleome trajectory

Germ cells' unique role in facilitating the transmission of genetic information across generations places them squarely at the origin of totipotency. Nonetheless, previous efforts have placed greater emphasis on the completion of meiosis, fertilization, and early zygotic development due to a combination of technical factors as well as greater interest in the genesis of a entirely new life; as a result, while still an intensively investigated period, the mitotic development of germ cells spanning the blastocyst stage until germline stem cells have remained less understood (Fig. 4.1).[69] This particular stretch of germ cell development comprises a number critical cell fate decision known to involve sweeping epigenome remodelling, with perhaps the most drastic one being epigenetic reprogramming in primordial germ cells.

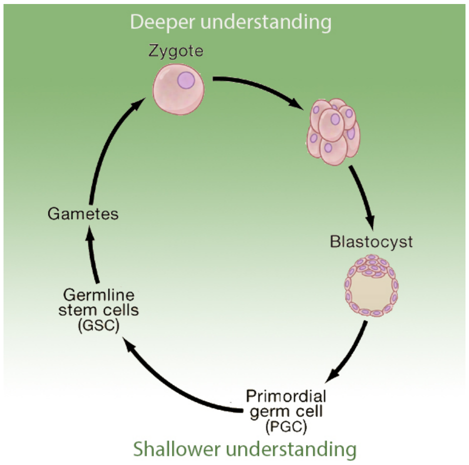


Figure 4.1: Cycle of gametogenesis. Adapted from Seydoux & Braun [69] with permission

Epigenetic reprogramming is known as one of the most unique remodelling developmental events, as the genome-wide de-methylation wipes the slate clean and potentiates the activation for essential germline expression programs via a combination of promoter de-methylation and removal of heterochromatic modifications such as H3K9me3 and H2AK119ub.[70] In this vulnerable stage of genome-wide hypomethylation, H3K27me3's expansion has been previously reported to provide a layer of restraint,[71] and our dataset indeed confirms this phenomenon as well as newly uncover a corresponding shrinkage

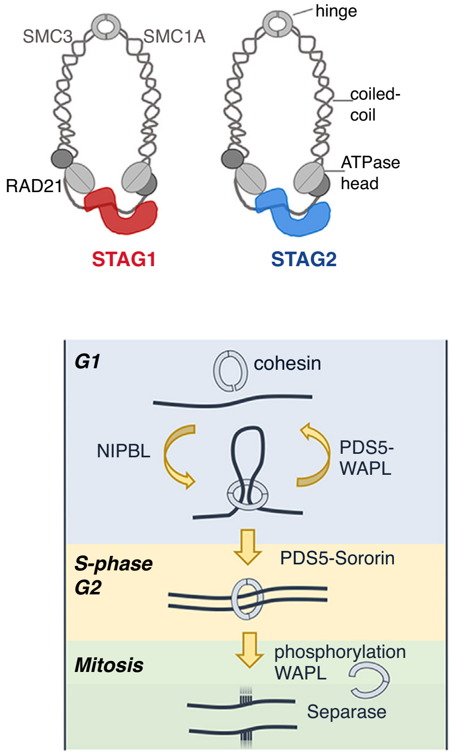


Figure 4.2: Functions of cohesin in 3D genome organization. Reproduced from Cuadrado & Losada [72] with permission

of H3K36me2 – once again highlighting the two chromatin marks' close coupling. Nevertheless, we discovered abundant enhancer signatures such as open chromatin and H3K4me1 pervade across the genome, bringing into question what other mechanisms exist to prevent these elements from driving spurious activation. Looking beyond histone modifications, we were able to identify elevated insulation as a novel hallmark of epigenetic reprogramming, emphasizing the importance of maintaining proper enhancer-promoter wiring for bridling a hypo-methylated genome. In contrast to the conventional view of de-DNA de-methylation promoting CTCF binding,[73] we saw that enhanced insulation in PGCs was not associated with differential CTCF patterns – but rather a shortening of cohesin/CTCF's residence time at boundary elements. As we did not detect strong associations between insulation change and differential signals for any of the myriad epigenetic modifications profiled, we consider this to implicate alternative mechanisms. Beyond epigenetic modifications, the loop extrusion machinery's chromatin association can be governed through diverse processes ranging from varying the subunit composition of cohesin (e.g., SA1- vs SA2-containing cohesin) to fluctuating levels of cohesin loading or unloading proteins including NIPBL and WAPL (Fig. 4.2).[72] Earlier reports have claimed that, though both SA1 and SA2 are associated with NIBPL-independent cohesin loading onto chromatin at R-loops, cohesin-SA1 preferentially contributes to greater insulation at TAD boundaries while cohesin-SA2 instead demonstrates greater association with PRC1-mediated compaction.[74] On the other hand, defects in cohesin loading via mutations NIBPL has been previously associated not only with a weakening of domain structures, but also directly implicated with developmental delay in Cornelia de Lange syndromes.[75] As we found the 3D genome organization of PGCLCs concurrently matches phenotypes observed for alterations in both cohesin-intrinsic and -extrinsic factors (e.g., broader compartments, shorter loops), further work is necessary to resolve the ambiguity. In particular, determining the most important mechanism at play in modulating physiological chromatin insulation will be essential to understanding germline nucleome dynamics at a deeper level as well as bear relevance to both developmental and chromatin biology at large – especially considering the broad purview of loop extrusion in gene regulation.

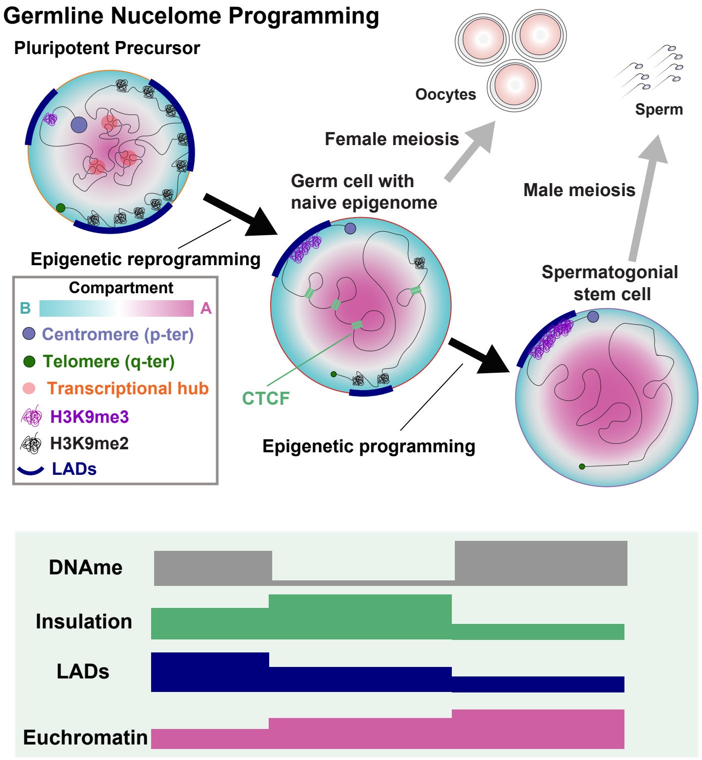


Figure 4.3: Germline nucleome dynamics.

Re-methylation of the entire genome following the sweeping loss of DNA methylation during epigenetic reprogramming has led most to believe that the broader epigenome and nucleome may undergo a similar re-setting. Indeed, our observation of transient changes such as temporary increase of H3K27me3 and H3K4me1 as well as decrease of H3K36me2 specifically at the lowest point of DNA methylation, with a rapidly ensuing rebound, is consistent with this proposition. On the other hand, we noted that nucleome re-organization remarkably progresses in a monotonic manner, with decompaction continuing unaffected throughout epigenetic reprogramming, culminating in a highly euchromatized genome in spermatogonia. The remaining heterochromatic region persisting in the loosened spermatogonial genome was linked by us to the expansion of H3K9me3 into broad domains that engage in both local and distal aggregation, often corresponding to pericentromeric regions. The demonstrated association between H3K9me3 domain expansion with heightened heterochromatic aggregation thus provides a physiological context in which the genomic distribution of histone modifications is tightly coupled with higher-order chromatin organization,[24] adding to the growing body of evidence highlighting liquid-liquid phase separation and nuclear condensates as a pertinent regulatory modality.[29] Through FISH, we were additionally able to visualize that spermatogonial chromosomes show minimal attachment to the peripheral nuclear lamina, save for pericentromeric heterochromatin (PCH) that migrates from the nuclear interior radially outwards. As the assembly of pericentromeres into dense chromocenters have been previously noted as a prominent feature of mouse spermatogenesis, with the structure even persisting after meiosis,[76] our finding thus pinpoints the origin of this process. The drastic radial repositioning of chromosome we witnessed also suggests a decoupling of periphery organization of germline progenitors with terminally differentiated germ cells, which is consistent with the knowledge that lamina associated domains are established in a wholly de novo manner post fertilization.[77] At smaller scales, we determined that the elevated insulation in PGCs is swiftly weakened in spermatogonia stem cells as CTCF became evicted from chromatin due to methylation on H3K9 and H3K36 as well as DNA in spermatogonia stem cells. We subsequently demonstrated that the lost of CTCF binding can promote both transcriptional up- and down-regulation, in line with prior accounts of acute CTCF depletion causing varied transcriptomic effects.[78] But importantly, we found a robust relationship between ectopic enhancer-promoter contacts facilitated by the loss of insulatory boundaries with up-regulation of genes central to meiotic pathways; and in germline stem cells with impaired spermatogenic potential, we saw an incomplete erasure of the very same boundaries. These descriptions thus significantly bolster the molecular portrait of male germ cell development in its entirety from multiple angles (Fig. 4.3)

Enabling studies of 3D epigenomic alterations in a dynamic physiological context

Many proteins serve specialized functions such as the recognition or transmission of particular agonists, but chromatin modifiers (e.g., writers of epigenetic modifications) and architectural proteins (e.g., cohesin, lamin) moulding the global 3D epigenome and therefore can have far-reaching effects, modulating multiple pathways and setting off elaborate cascades.[79] In addition to genetic and transcriptomic information, chromatin profiling has demonstrated that unique biological contexts, whether it be a specific cell type or a particular cancer subtype, could also be defined by salient 3D epigenetic signatures.[26] Given this link, there is a growing need for models that can faithfully reproduce these context-specific features, either at steady-state or even as as a dynamic process, in easily manipulated systems such as cell lines rather than laborious animal models or intricate organoid cultures. The *in vitro* model of murine gametogenesis here illustrates the utility of accurately recapitulating differentiation in a dish.[80] In particular, this system allows for large amounts of very specific cell types along germ cell differentiation to be generated in a reproducible manner, which critically enables the use of resource-intensive 3D epigenome profiling techniques such as Hi-C in a straightforward manner and yield data with excellent signal-to-noise ratios. As the *in vitro* differentiation process was previously demonstrated to closely mirror the *in vivo* developmental trajectory and ultimately yield fully functional gametes, this model also facilitates the temporal assessment of 3D epigenome kinetics as well as direct evaluation of biological-relevant consequences by measuring spermatogenic potential. As the germline is known to undergo substantial 3D epigenome remodelling in a condensed period, it furnishes a rich setting to dissect chromatin dynamics; and the ongoing development of analogous *in vitro* systems in primates and other species also holding significant promise for future cross-species comparisons.[45]

During the acquisition of an androgenic DNA methylome after epigenetic reprogramming as PGCs differentiate into spermatogonia, select areas of the genome fail to become fully methylated, giving rise to broad partially methylated domains (PMDs). Using the *in vitro* system, we demonstrated that PMDs of spermatogonia stem cells can be directly traced back to their progenitor PGCs re-directing H3K9me3 into and H3K36me2 away from regions destined to eventually become PMDs. Therefore, this designates a physiological window for which the likes of super-resolution microscopy and live-cell imaging can be carried out to probe the mechanisms giving rise to hypo-methylated regions with high temporal resolution,[40,81] bearing relevance for both cancer and aging given the association of PMDs with mitotic history.[82] Our finding that PCH repositions from the nuclear interior towards the periphery along differentiation towards spermatogonia also provides an opportunity to determine the requisite co-factors involved in such large-scale chromosomal radial repositioning, for example, through a CRISPR screen. Seeing that PCH is known to yet again re-organize into a singular chromocenter in the nuclear interior after meiosis, this signifies a highly dynamic period in development that can help clarify the consequences of improper radial repositioning of chromosomal territories. Subsequently, the implicated processes can be compared to other cases of similar inversion events such as rod photoreceptors with uniquely inverted chromatin organization.[83] The severely DNA hypo-methylated genome of PGCs, too, represents a chance to closely assess what contributes to proper maintenance of genome integrity in this vulnerable state. For instance, we've already identified enhanced chromatin insulation as an additional layer of restraint. Since aberrant DNA hypo- and hyper-methylation are ubiquitous features of carcinogenesis,[84] a deeper mechanistic understanding of the genome's intrinsic defense mechanisms in a naturally hypomethylated state will aid the design of strategies to either steer malignant cells towards a healthier state or push them over the edge. Altogether, the interplay between different modes and scales of 3D genome organization represents a direly under-explored angle for addressing pathological defects.

Balancing act of euchromatin versus heterochromatin

The vast collection of epigenome profiles across diverse cell types has established that the genome is segregated into discrete domains with chromatin states defined by distinct combinations of epigenetic modifications.[85] As many modifications undergo coordinated changes are can be thought of as redundant from an informational standpoint, several key markers have emerged as useful indicators for prominent high-level domain classes such as H3K36 methylation for euchromatin, H3K27 methylation for facultative heterochromatin, and H3K9 methylation for constitutive heterochromatin.[86] In particular, it's been demonstrated that changes in H3K27 methylation patterns can be both induced by the distribution of H3K36 methylation and through perturbations to PRC1/2 sub-units or co-factors and histone H3 itself (e.g., H3K27M or the H3K27M-mimic – EZHIP).[87] By systematically profiling a host of published as well as newly generated samples, we showed that H3K27me3 distributions can be generally categorized into focused versus diffuse modes of chromatin patterning across developmental and disease-relevant contexts. These include not only H3K27M and EZHIP, but also the loss of H3K36 methyltransferase NSD1 (removal of H3K36me2 leading to H3K27me3 spread), over-expression of H3K36 methyltransferase NSD2 (increase of H3K36me2 leading to H3K27me3 shrinkage), loss of H2AK119 deubiquitinase BAP1 (diffusion of H2AK119ub accompanied by H3K27me3 spread due to PRC2's H2AK119ub-reading activity), loss of linker histone H1 (chromatin decompaction hampering PRC2 spread), among other prominent disease-associated mutations in chromatin modifiers. Additionally, we noted that the confinement of H3K27me3 to CpG islands and promoters was more prevalent among progenitor cell types when compared to more differentiated stages, therefore linking confined H3K27me3 with an early cellular state still possessing considerable differentiation potential. Although previous reports have noted an association between focal H3K27me3 with heightened long-range interaction in mouse embryonic stem cells,[88] we systematically extended this finding to other lineages as well as several epigenetically dysregulated cancers (e.g., PFA ependymoma with EZHIP over-expression, glioblastoma carrying the H3K27M mutation, multiple myeloma with duplicated NSD2).

Pursuing this 3D epigenomic relationship further, we specifically identified the binding of a H3K27me3 reader, canonical PRC1 (cPRC1), to designate regions participating in elevated long-range interactions concomitant with H3K27me3 confinement. This close coupling between epigenetics and 3D genome organization falls in line with cPRC1's known capacity to facilitate liquid-liquid phase separation via mechanisms such as the associative properties of cPRC1 subunit CBX2's intrinsically disordered regions.[89]

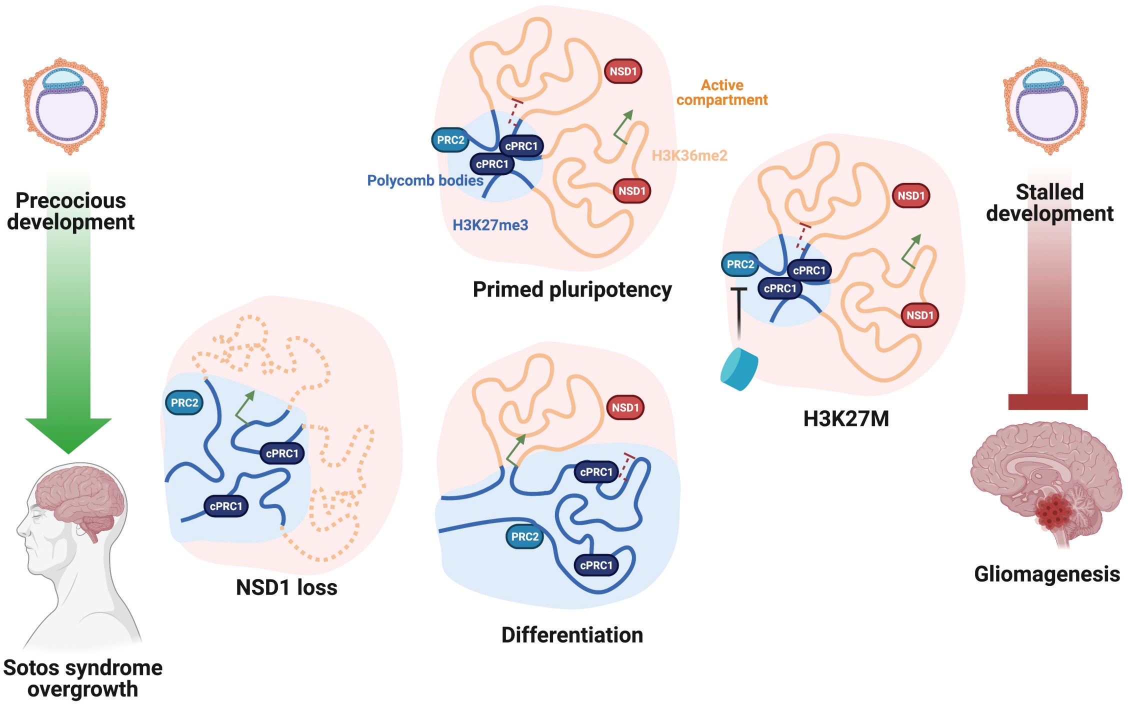


Figure 4.4: Polycomb-associated chromatin architecture.

Although the chromatin states of H3K27me3-enriched regions bound by cPRC1 are largely conserved across cell types and species, we found that cancer systems (H3K27M gliomas in particular) diverge from mammalian stem cells in terms of bivalency, with the former possessing both H3K4me3+ (bivalent) and H3K4me3- (non-bivalent) cPRC1 targets while only bivalent sites were present in the latter cases.[18] Yet the spreading of H3K27me3, and therefore dilution of cPRC1, led to greater differential interaction for non-bivalent promoters than bivalent ones in H3K27M GBMs, further implicating the balance between heterochromatic Polycomb and euchromatic trithorax systems beyond regulation of local chromatin environment to additionally governing higher-order chromatin organization.[90] In stem cells, the obligate expansion of H3K27me3 from nucleation sites as a consequence of decrease in NSD1-mediated H3K36me2 nevertheless results in a loss of long-range interaction between cPRC1 targets and modest increase in gene expression. Considering the over-representation of developmental regulators among Polycomb targets,[65] counting the likes of Hox genes as well as numerous tumor suppressor and oncogenes, our results indicate that the Polycomb system modulates cellular plasticity by means of both proximal and distal processes alike, especially in healthy and malignant stem-like cells (Fig. 4.4).

Therapeutically transforming the 3D epigenome

In view of the firm connection between epigenomic defects and pathogenesis based on evidence such as driver mutations in chromatin modifier genes, therapeutic development targeting the epigenome have been advancing at full steam (Fig. 4.5).[91] Indeed, a slew of inhibitors for epigenetic regulators has received regulatory approval in recent years: azacitidine, an inhibitor of DNA methyltransferases, for the treatment of acute myeloid leukemia, panobinostat, an inhibitor of histone de-acetylases, for the treatment of multiple myeloma, tazemetostat, a histone methyltransferase inhibitor (targeting EZH2, a component of PRC2), for the treatment of follicular lymphoma.[8] While many compounds act through competitive inhibition and prevent binding of the physiological substrate to the affected enzyme in a global fashion, more recent advances in E3 ligase modulators that can enable the recognition of neosubstrates have expanded the druggable space of proteins beyond conventional targets such as enzymes to even transcription factors and other epigenetic players.[92] In parallel, CRISPR-based technologies are enabling precision epigenome edits through the fusion of chromatin modifiers to inactive Cas9 proteins, thereby allowing sequence-specific targeting.[93] To fully capitalize on the growing collection of molecular instruments for biomedical aims, the initial target selec-

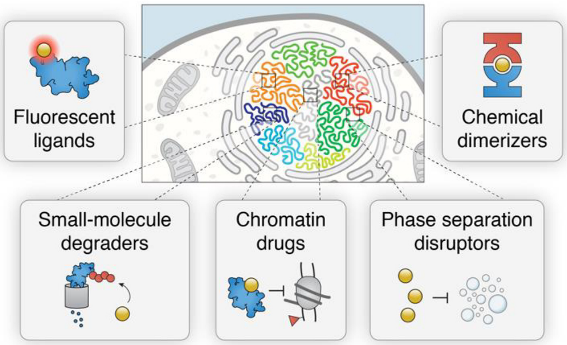


Figure 4.5: Chromatin-targeting drugs and probes. Reproduced from Cermakova & Hodges [91] with permission

tion stands as one of the most critical steps.

Through integration of multi-omics datasets traversing the epigenome, nucleome, and transcriptome, we were able to single out cPRC1-mediated aggregation of distal genomic regions as a key mechanism repressing developmental targets in H3K27M GBMs. As this specific subgroup of pediatric brain tumors has been associated with a stalled development phenotype based on single cell transcriptomic signatures,[58] the natural progression was to identify methods capable of alleviating this blockade through perturbing the hurdles in place. Our findings related to cPRC1 thus presented an excellent target in CBX2, as it is the key factor in both the recognition of H3K27me3 as well as contributing to the self-association of cPRC1 complexes to drive long-range interaction across long genomic separations. On this basis, we were able to quickly dial in a small-molecule treatment regime and observe clear impacts in terms of not only the de-repression of previously silent genes cPRC1 targets, but more important the up-regulation of key marker genes of differentiated cell types to a comparable level to cells with H3K27M fully removed via CRISPR-Cas9 knockout. This process thus highlights the immense potential of rational therapeutic design through first unravelling the epigenomic mechanisms contributing to tumor-intrinsic properties, consequently facilitating efficient translation of early discovery insights to downstream development.

Chapter 5

Conclusion

Genetics' leap to the forefront of biology since the turn of the millennium is in no small part thanks to monumental efforts encyclopedically cataloging sequence variation such as the Human Genome Project, the International HapMap Project, the 1000 Genomes Project, and more recently the UK Biobank; the field of epigenetics, in turn, resoundingly answered with endeavours counting the Roadmap Epigenomics Mapping Consortium and the Encyclopedia of DNA Elements (ENCODE) project. Nonetheless, perhaps more decisive in accentuating genetics in the public conscience has been everyday uses such as prenatal genetic screening and genetic ancestry testing, affording more personal insights. In other words, it's become evident that fundamental biological research must be carried out in lockstep with clinically oriented translational endeavours, as the former can serendipitously bear significant consequence for the latter. Being the immediate layer of organization above DNA strands, chromatin dynamics have emerged as the natural progression as the field looks beyond DNA. With this step up in scale comes along a dramatic escalation in complexity, demanding multi-faceted characterizations of genome regulation. At the same time, the reward for untangling such an intricate network of chromatin remodelling processes can be immense, since it would allow effective use of incredible biotechnological advances and expanding therapeutic arsenal.

As our understanding of germline development improves via the thorough investigation of current systems such as the vitro differentiation model of gametogenesis in mouse, this knowledge could be in turn used to propel the betterment of comparable setups in other species such as human. Simultaneously, the increasing capabilities of single cell technologies will enable analogous efforts for *in vivo* samples, additionally providing complementary insights into key aspects such as the cellular microenvironment during physiological development. Ultimately, the insights obtained from these systems will enable significant progress towards realizing the full potential of stem cell reprogramming, addressing fundamental needs in reproductive medicine. Nevertheless, the pursuit of these efforts must also heed bioethical concerns, as choreographing life itself must be approached sensitively and aim towards bringing equitable benefits to all those in need.

The ballooning literature emphasizing the 3D epigenome's importance thus far has focused on select well-understand model systems such as embryonic stem cells, but as our discovery of distal cis-regulatory process such as cPRC1-associated looping as a pivotal disease mechanism highlight that these processes can have far-reaching impacts. Yet the clarification of these processes often proves non-trivial, as interplay can take place at multiple scales (e.g., individual genomic elements versus expansive domains), modalities (e.g., local chromatin compaction versus long-range spatial aggregation) and mechanisms (e.g., active loop extrusion competing with passive phase separation). Nevertheless, our work demonstrates that it is viable to methodically disentangle these interrelated phenomena through carefully overlaying multi-omics datasets on top of the wealth of existing knowledge, and paints a blueprint for future efforts of understanding different diseases from a chromatin perspective.

Looking beyond the genome, epigenome, and nucleome, there remain other under-explored aspects of the mysterious tiny pearls that are the cells constituting life. These new frontiers range from a quantitative comprehension of the proteome and associated post-translational modifications, clarifying the physiological function of extrachromosomal DNA, as well as understanding the contributions of RNA-DNA and RNA-protein interactions, just to name a few. It is thus imperative to steadily adopt novel vantages and continuously update the axioms of life, pushing us to near frontiers of biomedicine. To thus effectively leverage these massive datasets of growing complexity, ever more sophisticated computational methods for their integration have been and will continue to be indispensable. More importantly, the efforts of computational scientists must be seamlessly united with those of experimental biologists to ensure the smooth and timely march towards a future of personalized medicine for all.

Chapter 6

References

1. Costello, A. & Badran, A. H. Synthetic Biological Circuits within an Orthogonal Central Dogma. *Trends in Biotechnology* 39, 59–71. doi:10.1016/j.tibtech.2020.05.013 (Jan. 2021).
2. McCarty, M. Discovering genes are made of DNA. *Nature* 421, 406–406. doi:10.1038/nature01398 (Jan. 2003).
3. Antonarakis, S. E. History of the methodology of disease gene identification. *American Journal of Medical Genetics Part A* 185, 3266–3275. doi:10.1002/ajmg.a.62400 (June 2021).
4. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nature Reviews Genetics* 7, 277–282. doi:10.1038/nrg1826 (Mar. 2006).
5. Shteinberg, M., Haq, I. J., Polineni, D. & Davies, J. C. Cystic fibrosis. *The Lancet* 397, 2195–2211. doi:10.1016/s0140-6736(20)32542-3 (June 2021).
6. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 20, 467–484. doi:10.1038/s41576-019-0127-1 (May 2019).
7. Greally, J. M. A user's guide to the ambiguous word 'epigenetics'. *Nature Reviews Molecular Cell Biology* 19, 207–208. doi:10.1038/nrm.2017.135 (Jan. 2018).
8. Bates, S. E. Epigenetic Therapies for Cancer. *New England Journal of Medicine* 383 (ed Longo, D. L.) 650–663. doi:10.1056/nejmra1805035 (Aug. 2020).
9. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* 20, 590–607. doi:10.1038/s41580-019-0159-6 (Aug. 2019).
10. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics* 12, 7–18. doi:10.1038/nrg2905 (Nov. 2010).
11. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* 20, 207–220. doi:10.1038/s41576-018-0089-8 (Jan. 2019).
12. Atlasi, Y. & Stunnenberg, H. G. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics* 18, 643–658. doi:10.1038/nrg.2017.57 (Aug. 2017).
13. Slack, J. M. W. Conrad Hal Waddington: the last Renaissance biologist? *Nature Reviews Genetics* 3, 889–895. doi:10.1038/nrg933 (Nov. 2002).

14. Granados, K., Poelchen, J., Novak, D. & Utikal, J. Cellular Reprogramming—A Model for Melanoma Cellular Plasticity. *International Journal of Molecular Sciences* 21, 8274. doi:10.3390/ijms21218274 (Nov. 2020).

15. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 590, 300–307. doi:10.1038/s41586-020-03145-z (Feb. 2021).

16. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* 173, 291–304.e6. doi:10.1016/j.cell.2018.03.022 (Apr. 2018).

17. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* 15, 272–286. doi:10.1038/nrg3682 (Mar. 2014).

18. Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S. & Croce, L. D. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics* 36, 118–131. doi:10.1016/j.tig.2019.11.004 (Feb. 2020).

19. Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* 167, 1188–1200. doi:10.1016/j.cell.2016.10.024 (Nov. 2016).

20. Jones, P. A., Issa, J.-P. J. & Baylin, S. Targeting the cancer epigenome for therapy. *Nature Reviews Genetics* 17, 630–641. doi:10.1038/nrg.2016.93 (Sept. 2016).

21. Jani, K. S. *et al.* Histone H3 tail binds a unique sensing pocket in EZH2 to activate the PRC2 methyltransferase. *Proceedings of the National Academy of Sciences* 116, 8295–8300. doi:10.1073/pnas.1819029116 (Apr. 2019).

22. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* 112. doi:10.1073/pnas.1518552112 (Oct. 2015).

23. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* 20, 535–550. doi:10.1038/s41580-019-0132-4 (June 2019).

24. Wang, L. *et al.* Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism. *Molecular Cell* 76, 646–659.e6. doi:10.1016/j.molcel.2019.08.019 (Nov. 2019).

25. Davidson, I. F. & Peters, J.-M. Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology* 22, 445–464. doi:10.1038/s41580-021-00349-7 (Mar. 2021).

26. Anania, C. & Lupiáñez, D. G. Order and disorder: abnormal 3D chromatin organization in human disease. *Briefings in Functional Genomics* 19, 128–138. doi:10.1093/bfgp/elz028 (Feb. 2020).

27. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences* 115. doi:10.1073/pnas.1717730115 (July 2018).

28. Rhodes, J. D. *et al.* Cohesin Disrupts Polycomb-Dependent Chromosome Interactions in Embryonic Stem Cells. *Cell Reports* 30, 820–835.e10. doi:10.1016/j.celrep.2019.12.057 (Jan. 2020).

29. Sabari, B. R., Dall'Agnese, A. & Young, R. A. Biomolecular Condensates in the Nucleus. *Trends in Biochemical Sciences* 45, 961–977. doi:10.1016/j.tibs.2020.06.007 (Nov. 2020).

30. Norton, H. K. & Phillips-Cremins, J. E. Crossed wires: 3D genome misfolding in human disease. *Journal of Cell Biology* 216, 3441–3452. doi:10.1083/jcb.201611001 (Aug. 2017).

31. Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. *Trends in Biotechnology* 34, 276–290. doi:10.1016/j.tibtech.2015.12.013 (Apr. 2016).

32. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17, 333–351. doi:10.1038/nrg.2016.49 (May 2016).

33. Sigal, Y. M., Zhou, R. & Zhuang, X. Visualizing and discovering cellular structures with super-resolution microscopy. *Science* 361, 880–887. doi:10.1126/science.aau1044 (Aug. 2018).

34. Van Mierlo, G. & Vermeulen, M. Chromatin Proteomics to Study Epigenetics — Challenges and Opportunities. *Molecular & Cellular Proteomics* 20, 100056. doi:10.1074/mcp.r120.002208 (2021).

35. Rivera, C. & Ren, B. Mapping Human Epigenomes. *Cell* 155, 39–55. doi:10.1016/j.cell.2013.09.011 (Sept. 2013).

36. Lu, C., Coradin, M., Porter, E. G. & Garcia, B. A. Accelerating the Field of Epigenetic Histone Modification Through Mass Spectrometry–Based Approaches. *Molecular & Cellular Proteomics* 20, 100006. doi:10.1074/mcp.r120.002257 (2021).

37. Illumina. *For All You Seq – DNA* https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/for-all-you-seq-dna.pdf (2016).

38. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* 21, 207–226. doi:10.1038/s41576-019-0195-2 (Dec. 2019).

39. Fudenberg, G. & Imakaev, M. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nature Methods* 14, 673–678. doi:10.1038/nmeth.4329 (June 2017).

40. Takei, Y. *et al.* Single-cell nuclear architecture across cell types in the mouse brain. *Science* 374, 586–594. doi:10.1126/science.abj1966 (Oct. 2021).

41. Illumina. *For All You Seq – RNA* https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/for-all-you-seq-rna.pdf (2016).

42. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* 20, 631–656. doi:10.1038/s41576-019-0150-2 (July 2019).

43. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nature Genetics* 51, 1369–1379. doi:10.1038/s41588-019-0485-9 (Sept. 2019).

44. Sasaki, H. & Matsui, Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nature Reviews Genetics* 9, 129–140. doi:10.1038/nrg2295 (Feb. 2008).

45. Saitou, M. Mammalian Germ Cell Development: From Mechanism to In Vitro Reconstitution. *Stem Cell Reports* 16, 669–680. doi:10.1016/j.stemcr.2021.01.008 (Apr. 2021).

46. Xia, W. & Xie, W. Rebooting the Epigenomes during Mammalian Early Embryogenesis. *Stem Cell Reports* 15, 1158–1175. doi:10.1016/j.stemcr.2020.09.005 (Dec. 2020).

47. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* 544, 110–114. doi:10.1038/nature21711 (Mar. 2017).

48. Chen, Z., Djekidel, M. N. & Zhang, Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. *Nature Genetics* 53, 551–563. doi:10.1038/s41588-021-00821-2 (Apr. 2021).

49. Du, Z. *et al.* Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Molecular Cell* 77, 825–839.e7. doi:10.1016/j.molcel.2019.11.011 (Feb. 2020).

50. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics* 17, 284–299. doi:10.1038/nrg.2016.13 (Mar. 2016).

51. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi:10.1038/s41586-020-1969-6 (Feb. 2020).

52. Zhao, S., Allis, C. D. & Wang, G. G. The language of chromatin modification in human cancers. *Nature Reviews Cancer* 21, 413–430. doi:10.1038/s41568-021-00357-x (May 2021).

53. Nacev, B. A. *et al.* The expanding landscape of 'oncohistone' mutations in human cancers. *Nature* 567, 473–478. doi:10.1038/s41586-019-1038-1 (Mar. 2019).

54. Jain, S. U. *et al.* H3 K27M and EZHIP Impede H3K27-Methylation Spreading by Inhibiting Allosterically Stimulated PRC2. *Molecular Cell* 80, 726–735.e7. doi:10.1016/j.molcel.2020.09.028 (Nov. 2020).

55. Pollack, I. F., Agnihotri, S. & Broniscer, A. Childhood brain tumors: current management, biological insights, and future directions. *Journal of Neurosurgery: Pediatrics* 23, 261–273. doi:10.3171/2018.10.peds18377 (Mar. 2019).

56. Bayliss, J. *et al.* Lowered H3K27me3 and DNA hypomethylation define poorly prognostic pediatric posterior fossa ependymomas. *Science Translational Medicine* 8. doi:10.1126/scitranslmed.aah6904 (Nov. 2016).

57. Harutyunyan, A. S. *et al.* H3K27M induces defective chromatin spread of PRC2-mediated repressive H3K27me2/me3 and is essential for glioma tumorigenesis. *Nature Communications* 10. doi:10.1038/s41467-019-09140-x (Mar. 2019).

58. Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nature Genetics* 51, 1702–1713. doi:10.1038/s41588-019-0531-7 (Nov. 2019).

59. Soshnev, A., Josefowicz, S. & Allis, C. D. Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome. *Molecular Cell* 62, 681–694. doi:10.1016/j.molcel.2016.05.004 (June 2016).

60. Streubel, G. *et al.* The H3K36me2 Methyltransferase Nsd1 Demarcates PRC2-Mediated H3K27me2 and H3K27me3 Domains in Embryonic Stem Cells. *Molecular Cell* 70, 371–379.e5. doi:10.1016/j.molcel.2018.02.027 (Apr. 2018).

61. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology* 19, 229–244. doi:10.1038/nrm.2017.119 (Dec. 2017).

62. Farhangdoost, N. *et al.* Chromatin dysregulation associated with NSD1 mutation in head and neck squamous cell carcinoma. *Cell Reports* 34, 108769. doi:10.1016/j.celrep.2021.108769 (Feb. 2021).

63. Lhoumaud, P. *et al.* NSD2 overexpression drives clustered chromatin and transcriptional changes in a subset of insulated domains. *Nature Communications* 10. doi:10.1038/s41467-019-12811-4 (Oct. 2019).

64. Lu, C. *et al.* Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science* 352, 844–849. doi:10.1126/science.aac7272 (May 2016).

65. Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nature Reviews Molecular Cell Biology* 22, 815–833. doi:10.1038/s41580-021-00398-y (Aug. 2021).

66. Maezawa, S. *et al.* SCML2 promotes heterochromatin organization in late spermatogenesis. *Journal of Cell Science.* doi:10.1242/jcs.217125 (Jan. 2018).

67. Ragazzini, R. *et al.* EZHIP constrains Polycomb Repressive Complex 2 activity in germ cells. *Nature Communications* 10. doi:10.1038/s41467-019-11800-x (Aug. 2019).

68. Bock, C. *et al.* High-content CRISPR screening. *Nature Reviews Methods Primers* 2. doi:10.1038/s43586-021-00093-4 (Feb. 2022).

69. Seydoux, G. & Braun, R. E. Pathway to Totipotency: Lessons from Germ Cells. *Cell* 127, 891–904. doi:10.1016/j.cell.2006.11.016 (Dec. 2006).

70. Mochizuki, K. *et al.* Repression of germline genes by PRC1.6 and SETDB1 in the early embryo precedes DNA methylation-mediated silencing. *Nature Communications* 12. doi:10.1038/s41467-021-27345-x (Dec. 2021).

71. Kurimoto, K. *et al.* Quantitative Dynamics of Chromatin Remodeling during Germ Cell Specification from Mouse Embryonic Stem Cells. *Cell Stem Cell* 16, 517–532. doi:10.1016/j.stem.2015.03.002 (May 2015).

72. Cuadrado, A. & Losada, A. Specialized functions of cohesins STAG1 and STAG2 in 3D genome architecture. *Current Opinion in Genetics & Development* 61, 9–16. doi:10.1016/j.gde.2020.02.024 (Apr. 2020).

73. Wiehle, L. *et al.* DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Research* 29, 750–761. doi:10.1101/gr.239707.118 (Apr. 2019).

74. Cuadrado, A. *et al.* Specific Contributions of Cohesin-SA1 and Cohesin-SA2 to TADs and Polycomb Domains in Embryonic Stem Cells. *Cell Reports* 27, 3500–3510.e4. doi:10.1016/j.celrep.2019.05.078 (June 2019).

75. Garcia, P. *et al.* Disruption of NIPBL/Scc2 in Cornelia de Lange Syndrome provokes cohesin genome-wide redistribution with an impact in the transcriptome. *Nature Communications* 12. doi:10.1038/s41467-021-24808-z (July 2021).

76. Berrios, S. Nuclear Architecture of Mouse Spermatocytes: Chromosome Topology, Heterochromatin, and Nucleolus. *Cytogenetic and Genome Research* 151, 61–71. doi:10.1159/000460811 (2017).

77. Borsos, M. *et al.* Genome–lamina interactions are established de novo in the early mouse embryo. *Nature* 569, 729–733. doi:10.1038/s41586-019-1233-0 (May 2019).

78. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* 169, 930–944.e22. doi:10.1016/j.cell.2017.05.004 (May 2017).

79. Chen, T. & Dent, S. Y. R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews Genetics* 15, 93–106. doi:10.1038/nrg3607 (Dec. 2013).

80. Ishikura, Y. *et al.* In vitro reconstitution of the whole male germ-cell development from mouse pluripotent stem cells. *Cell Stem Cell* 28, 2167–2179.e9. doi:10.1016/j.stem.2021.08.005 (Dec. 2021).

81. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* 568, 235–239. doi:10.1038/s41586-019-1049-y (Mar. 2019).

82. Decato, B. E. *et al.* Characterization of universal features of partially methylated domains across tissues and species. *Epigenetics & Chromatin* 13. doi:10.1186/s13072-020-00363-7 (Oct. 2020).

83. Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* 570, 395–399. doi:10.1038/s41586-019-1275-3 (June 2019).

84. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* 1, 239–259. doi:10.2217/epi.09.33 (Dec. 2009).

85. Dixon, J., Gorkin, D. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell* 62, 668–680. doi:10.1016/j.molcel.2016.05.018 (June 2016).

86. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Experimental & Molecular Medicine* 49, e324–e324. doi:10.1038/emm.2017.11 (Apr. 2017).

87. Phillips, R. E., Soshnev, A. A. & Allis, C. D. Epigenomic Reprogramming as a Driver of Malignant Glioma. *Cancer Cell* 38, 647–660. doi:10.1016/j.ccell.2020.08.008 (Nov. 2020).

88. McLaughlin, K. *et al.* DNA Methylation Directs Polycomb-Dependent 3D Genome Reorganization in Naive Pluripotency. *Cell Reports* 29, 1974–1985.e6. doi:10.1016/j.celrep.2019.10.031 (Nov. 2019).

89. Plys, A. J. *et al.* Phase separation of Polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. *Genes & Development* 33, 799–813. doi:10.1101/gad.326488.119 (June 2019).

90. Schuettengruber, B., Bourbon, H.-M., Croce, L. D. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* 171, 34–57. doi:10.1016/j.cell.2017.08.002 (Sept. 2017).

91. Cermakova, K. & Hodges, H. Next-Generation Drugs and Probes for Chromatin Biology: From Targeted Protein Degradation to Phase Separation. *Molecules* 23, 1958. doi:10.3390/molecules23081958 (Aug. 2018).

92. Chamberlain, P. P. & Hamann, L. G. Development of targeted protein degradation therapeutics. *Nature Chemical Biology* 15, 937–944. doi:10.1038/s41589-019-0362-y (Sept. 2019).

93. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR technologies for precise epigenome editing. *Nature Cell Biology* 23, 11–22. doi:10.1038/s41556-020-00620-7 (Jan. 2021).

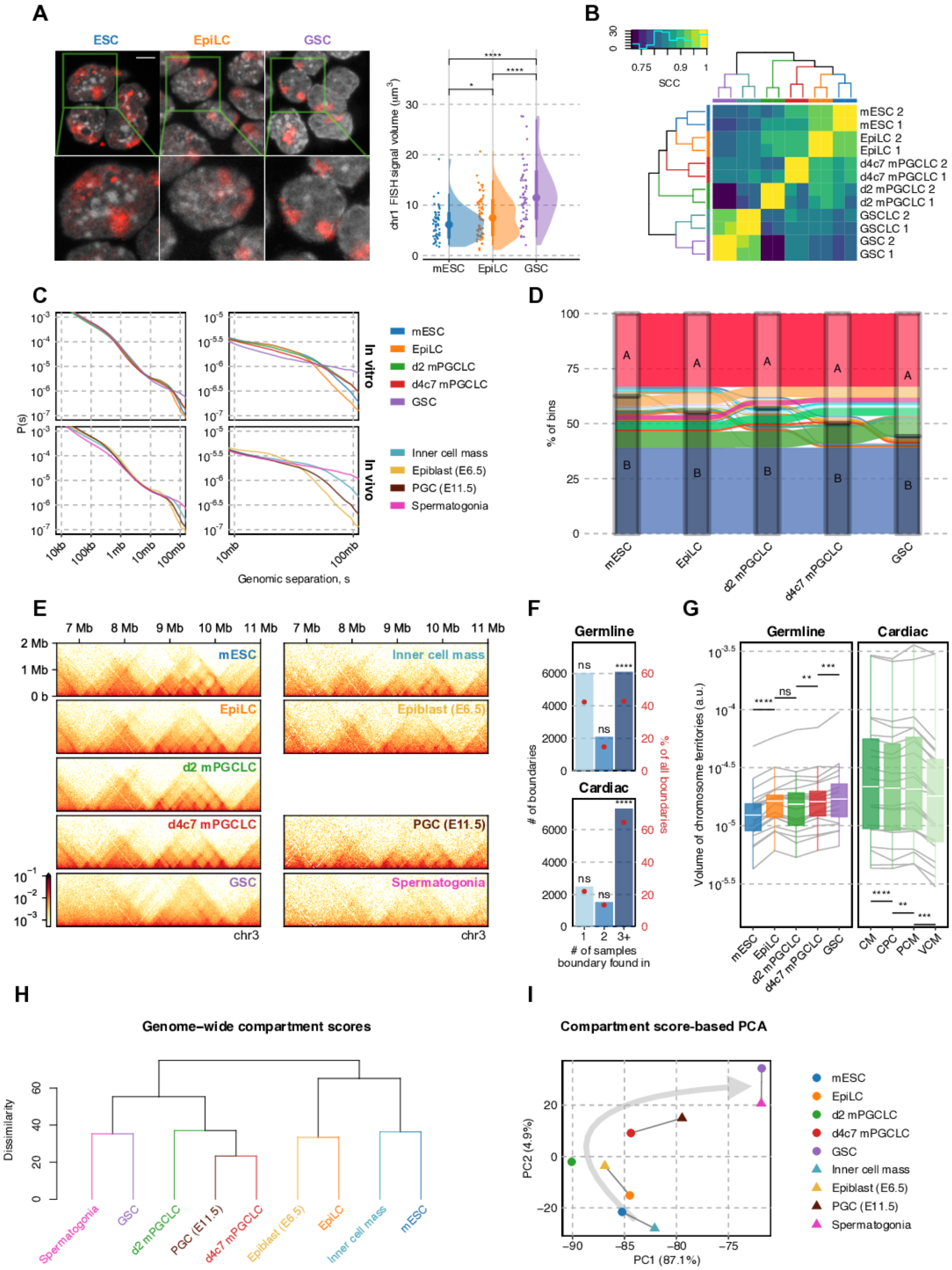
Appendix A

Supplementary information for chapter 2

The following pages contain supplementary information for chapter 2: Nucleome programming for the foundation of totipotency in mammalian germline development.

FIGURES

Figure A.1. Investigation of global nuclear architecture dynamics through Hi-C and FISH.



(A) Fluorescence in situ hybridization (FISH) against chromosome 1 (red) with DAPI counterstaining (grey). (Left) Z-stacked representative images (top left) are paired with magnified views (bottom left). (Right) The distribution of "surface" volumes for chr1, as seen for chr16, validates chromosomal decondensation in GSCs. Number of cells = 51/68/53 for mESC/EpiLC/GSC. Wilcoxon rank-sum test p-values (left to right): 4.16e-2, 4.33e-6, 8.68e-9. P-value symbol brackets: **** = [0, 0.0001); *** = [0.0001, 0.001]; ** = [0.001, 0.01); * = [0.01, 0.05); ns = [0.05, 1].

(B) Hierarchical clustering of stratum-adjusted correlation coefficients (SCC) between samples validating the reproducibility of biological replicates.

(C) Contact probability decay across different inter-loci separation distances for various cell types throughout *in vivo* and in vitro germ cell differentiation, demonstrating a gain of distal interactions along differentiation, especially at distances >50 Mb.

(D) Sankey diagram of compartment identities in 50 kb bins across cell types. Compartment A regions newly acquired by GSCs are formed through a unidirectional switch of B-A with relatively little reversal.

(E) 25 kb-resolution balanced contact maps spanning chr3:5–12.5 mb.

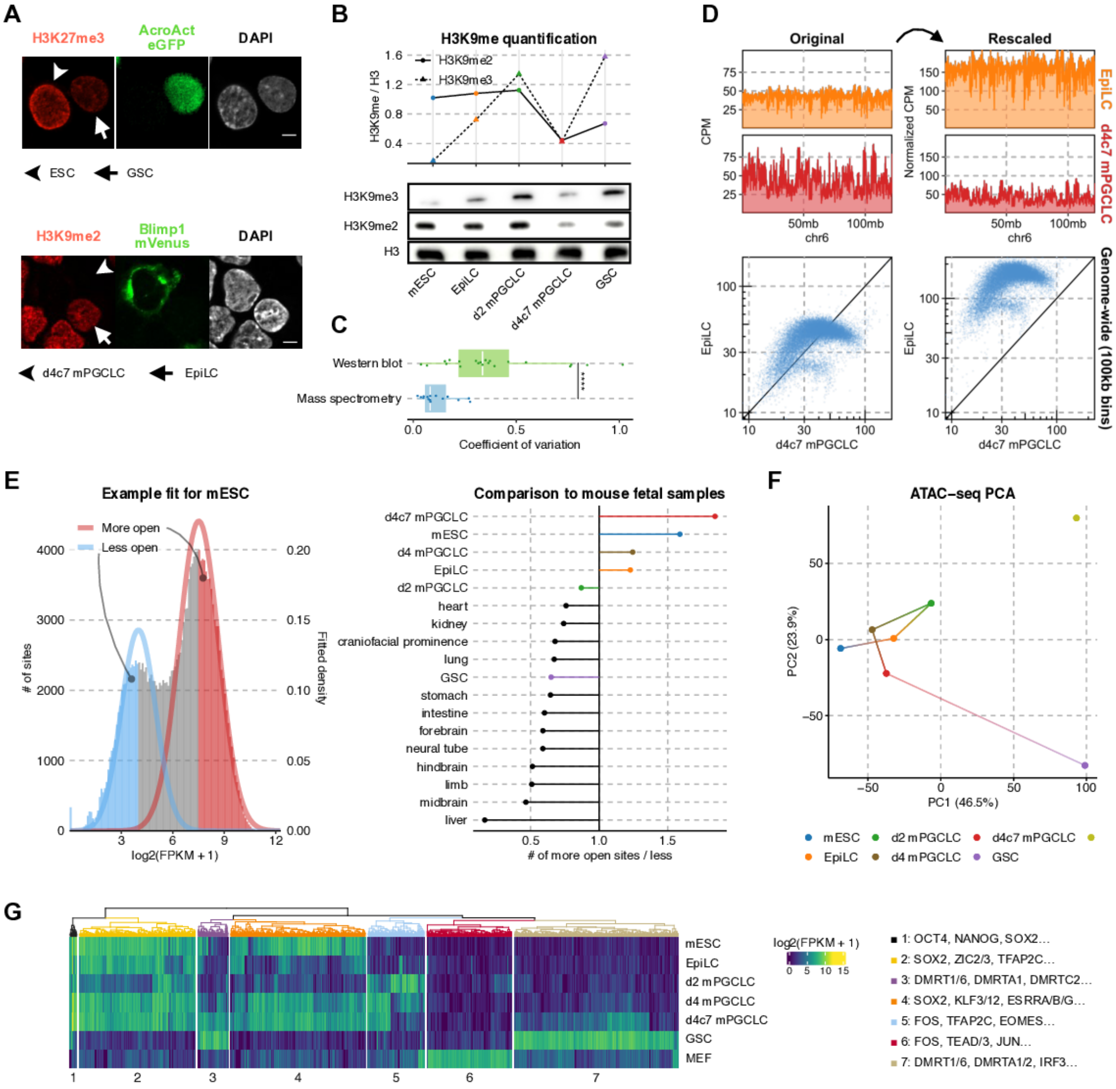
(F) Degree of TAD boundary conservation in different lineages. Consistent across different lineages, more than 40% TAD boundaries are significantly conserved across differentiation. One-sided permutation tests were carried out by shuffling sample labels 100000 times, with p-values (left to right, top to bottom): 1, 1, 1e-5, 1, 1, 1e-5.

(G) Convex hull volumes of CSynth-produced chromosome 3D models during the development of different lineages, after normalization to unit backbone length. n = 22/19 for cardiac/germline. Wilcoxon signed-rank test p-values (left to right): 1.91e-6, 1.89e-1, 1.69e-3, 2.61e-4, 4.77e-6, 1.86e-3, 2.93e-4.

(H) UHC based on Euclidean distance between 100 kb compartment score tracks for cell types from in vitro and *in vivo* germ cell differentiation, with comparable stages consistently grouped together.

(I) PCA of compartment scores at 100 kb resolution for various cell types throughout *in vivo* and in vitro germ cell differentiation, with comparable stages consistently grouped together.

Figure A.2. Quantitative epigenome analysis by mass spectrometry and chromatin accessibility analysis by ATAC-seq.



(A) (Top) Immunofluorescence against H3K27me3 in mESCs and GSCs; the shaftless arrow marks a GFP+ GSCs and the shaftless arrowhead indicates mESCs. (Bottom) Immunofluorescence against H3K9me2 in EpiLCs and d4c7 mPGCLCs; the shaftless arrowhead marks a Blimp1-mVenus+ d4c7 mPGCLCs and the arrow indicates EpiLCs. Scale bars = 10 μm.

(B) Western blot against H3K9me3, H3K9me2, and histone H3 in each cell type (bottom) and H3-normalized quantification (top).

(C) Coefficients of variation across replicates of histone modification abundance as measured by quantitative histone mass spectrometry versus western blot for H3K9me2, H3K9me3, and H3K27me3. Mass spectrometry measurements consistently exhibit higher reproducibility. Number of biological replicates = 15/21 for mass spectrometry/western blot. Wilcoxon rank-sum test p-value: 5.34e-5.

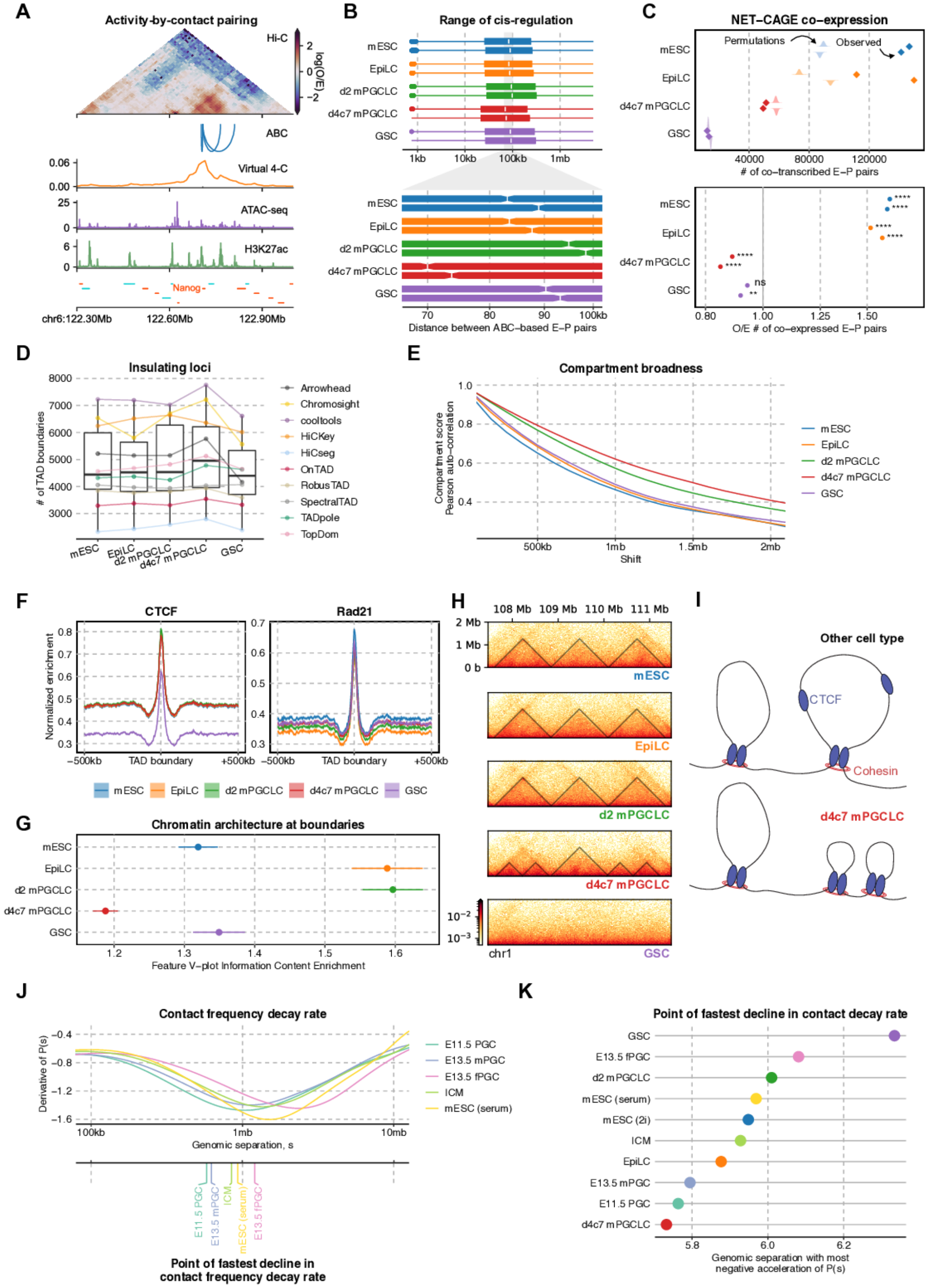
(D) Schematic of normalizing histone modification ChIP-seq via mass spectrometry-derived coefficients. With only depth-normalization (left), EpiLCs and d4c7 mPGCLCs appear to have comparable H3K9me2 profiles both in terms of coverage tracks (top) and in a pairwise scatter plot comparing the two cell types (bottom); after multiplication of their relative abundances based on mass spectrometry, the comparatively lower levels of H3K9me2 in d4C7GCLCs become apparent (right).

(E) Comparison of regions with greater ("more open") and reduced ("less open") accessibility in the union peak set of germline samples and E14.5 mouse fetal tissues [41] (left). Through fitting two-component gaussian mixture models, d4c7 mPGCLCs stand out as possessing the most permissive genome (right).

(F) PCA of ATAC-seq signals in the top 10,000 most variable peaks from the union peak set including MEFs [151].

(G) UHC of the top 2,000 most variable ATAC-seq peaks in the union peak set including MEFs. (left) Clustered ATAC-seq enrichment heatmap; (right) overrepresented TF-binding motifs in each cluster.

Figure A.3. Exploration of cis-regulatory element by NET-CAGE combined with Hi-C and comparison against public Hi-C datasets.



(A) An example of enhancer-promoter interactions for Nanog in mESCs as predicted by ABC, all of which correspond to known associations including super-enhancers.

(B) (Top) Distribution of distances separating ABC-predicted enhancer-promoter pairs in each replicate. The central band of boxplots indicate median values, while the lower and upper hinge correspond to the first and third quartile, and the upper whiskers extend to the largest value % 1.5 * IQR and vice versa for the lower whiskers. Notches correspond to 1.58 * interquartile range of distances / (# of E-P pairs)$^{1/2}$, comparable to 95% confidence intervals around the median. d4c7 mPGCLCs' E-P pairs are significantly shorter in range than those of other cell types. (Bottom) Magnified view from 60 kb to 100 kb. Number of ABC E-P pairs from left to right: 60535, 59312, 59116, 59075, 58702, 58704, 53092, 52074, 60867, 58858.

(C) (Top) Co-transcription of enhancer-promoter pairs with correlated NET-CAGE expression. The observed number of correlated E-P pairs involving tag clusters transcribed (TPM > 1) in a given cell type (points) are compared against a permuted background in which tag clusters are sampled from the union tag cluster set. (Bottom) Observed / expected number of E-P pairs with correlated NET-CAGE expression and co-expressed (>1 TPM) in a given cell type. Two-sided permutation tests were carried out by sampling 100000 times from the set of elements expressed in at least 1 cell type, with p-values (left to right): 2e-5, 2e-5, 2e-5, 2e-5, 2e-5, 2e-5, 6.44e-3, 7.64e-2. Two biological replicates in each cell type were analyzed.

(D) Number of TAD boundaries in each cell type across 10 different algorithms. Dots correspond to values produced by a specific algorithm for a given cell type and are grouped into lines by algorithm.

(E) Auto-correlation of compartment scores (25 kb bins), with a slower decay indicative of broader compartments.

(F) Aggregate plots of S3V2-normalized ChIP-seq profiles for CTCF and Rad21 around the union set of TAD boundaries.

(G) Mean f-VICE across replicates (error bars indicate standard errors) for CTCF motifs overlapping both Rad21 and CTCF peaks within the union set of TAD boundaries. Two biological replicates per cell type were analyzed.

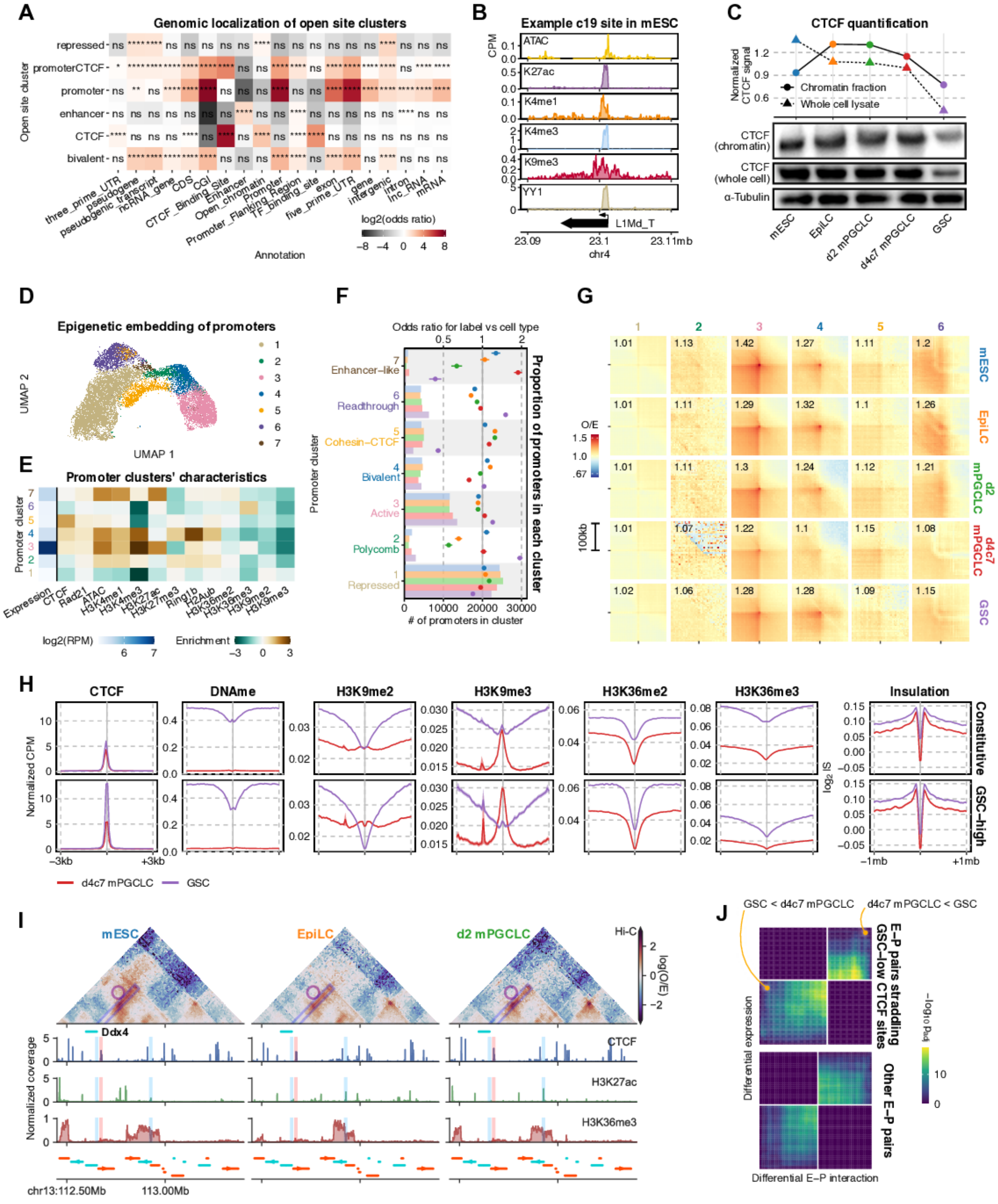
(H) representative locus demonstrating the emergence of smaller insulated domains in d4c7 mPGCLCs within otherwise homogeneous wider TADs observed in earlier stages.

(I) Proposed mechanism for elevated insulation via the reduction of loop extrusion factor's residence time, leading to shorter loops and domains.

(J) (Top) Slope of contact decay (P(s)) curves as a function of genomic separation in log-log space for *in vivo* germline development [4,38]; (bottom) genomic separation with the most negative second derivative of P(s) in log-log space, corresponding to distance of fastest decline in contact frequency.

(K) Genomic separation with fastest decline in contact frequency for cell types across *in vivo* and *in vitro* germ cell differentiation.

Figure A.4. Open site chromatin state dynamics and differential CTCF binding throughout germ cell differentiation.



(A) Overlap enrichment analysis of consolidated open site clusters against annotations from the Ensembl Regulatory build. P-values computed using Fisher's exact tests.

(B) Select ChIP-seq coverage tracks around a representative cluster 2 loci.

(C) Western blot against CTCF in the chromatin-bound fraction (top row) and whole cell lysate (middle row) as well as α-tubulin (bottom row) in each cell type. The signals of CTCF from whole cell lysates were normalized by α-Tubulin, while those of the chromatin-bound fraction were normalized by the mean across all cell types (top panel).

(D) 2D UMAP embedding based on epigenetic signals in promoters for each cell type, with labels derived from semi-supervised HDBSCAN.

(E) Enrichment of epigenetic signals in each promoter cluster and expression of the cognate gene.

(F) Association between promoter clusters and cell types. Number of open sites per cell type in each cluster (top axis: bars) and their enrichment as odds ratios (bottom axis: dots). Error bars indicate 95% confidence intervals.

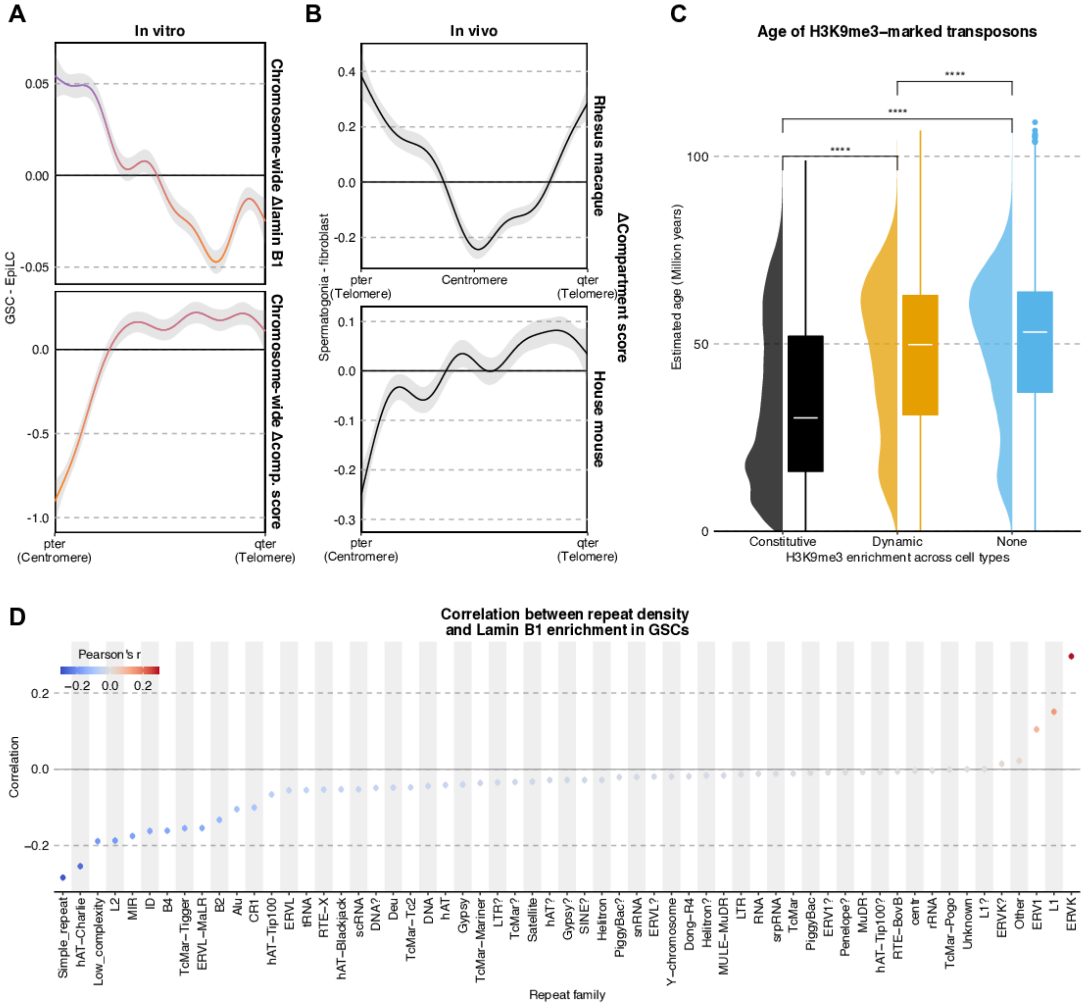
(G) Pile-up plots of intra-class promoter-promoter interactions.

(H) Contributors of differential CTCF binding. The aggregate plot of various ChIP-seq enrichment signals (left) as well as the insulation score (right) near CTCF-binding sites found both in cell types ("constitutive") or only GSCs but not in d4c7 mPGCLCs ("GSC-high") appear largely identical in their chromatin state yet distinct from those lost in GSCs. n = 35692/13364 for constitutive/GSC-high peaks.

(I) 3D epigenetic landscape rewiring near *Ddx4*. Observed/expected contact maps at 10 kb resolution for mESCs, EpiLCs and d2 mPGCLCs are shown alongside select ChIP-seq coverage tracks. A strongly insulating CTCF peak (highlighted in red) upstream of *Ddx4*'s TSS is found in all earlier stages and prevents spurious activation.

(J) Coordinated differential expression and E-P looping between d4c7 mPGCLCs and GSCs. Strong correlation was observed when applying stratified rank-rank hypergeometric overlap to genes ranked by differential expression versus differential E-P interactions straddling sites depleted of CTCF binding in GSCs. While increased E-P looping is correlated with elevated expression regardless of whether the interaction spans differential CTCF-bound sites, the degree of coordination is stronger (i.e., more significant / brighter) for those that do straddle GSC-depleted sites.

Figure A.5. Inter-species comparison of germ-cell specific chromatin structure and characterization of H3K9me3-enriched repeats.



(A) Average distributions of differential (GSC – EpiLC) lamin B1 enrichment (top) or compartment score (bottom) across all chromosomes (1–19, X). Ribbons correspond to 95% confidence intervals of fitted GAMs.

(B) Average distributions of compartment score (spermatogonia – fibroblast) across all chromosomes (excluding Y) for *Macaca mulatta* (top) and *Mus musculus* (bottom).

(C) Estimated age of families overlapping H3K9me3 domains based on age = divergence/substitution rate with 4.5×10^{-9} as the rate and milliDiv from RepeatMasker as the divergence [152]. Wilcoxon rank-sum tests p-values, from left to right: 0, 0, 0. Number of TE instances, from left to right: 227732, 982369, 2671107.

(D) Correlation between lamin B1 enrichment and density for different repeat families.

TABLES AND MOVIES

Table A.1. Sequencing summary.

Table A.2. Histone modification abundances.

Table A.3. Motif enrichment results.

Table A.4. Cluster annotations.

Table A.5. Annotation overlap results.

Table A.6. Pathway association results.

Movie A.1. 3D re-organization of chromosome 16 during germ cell development.

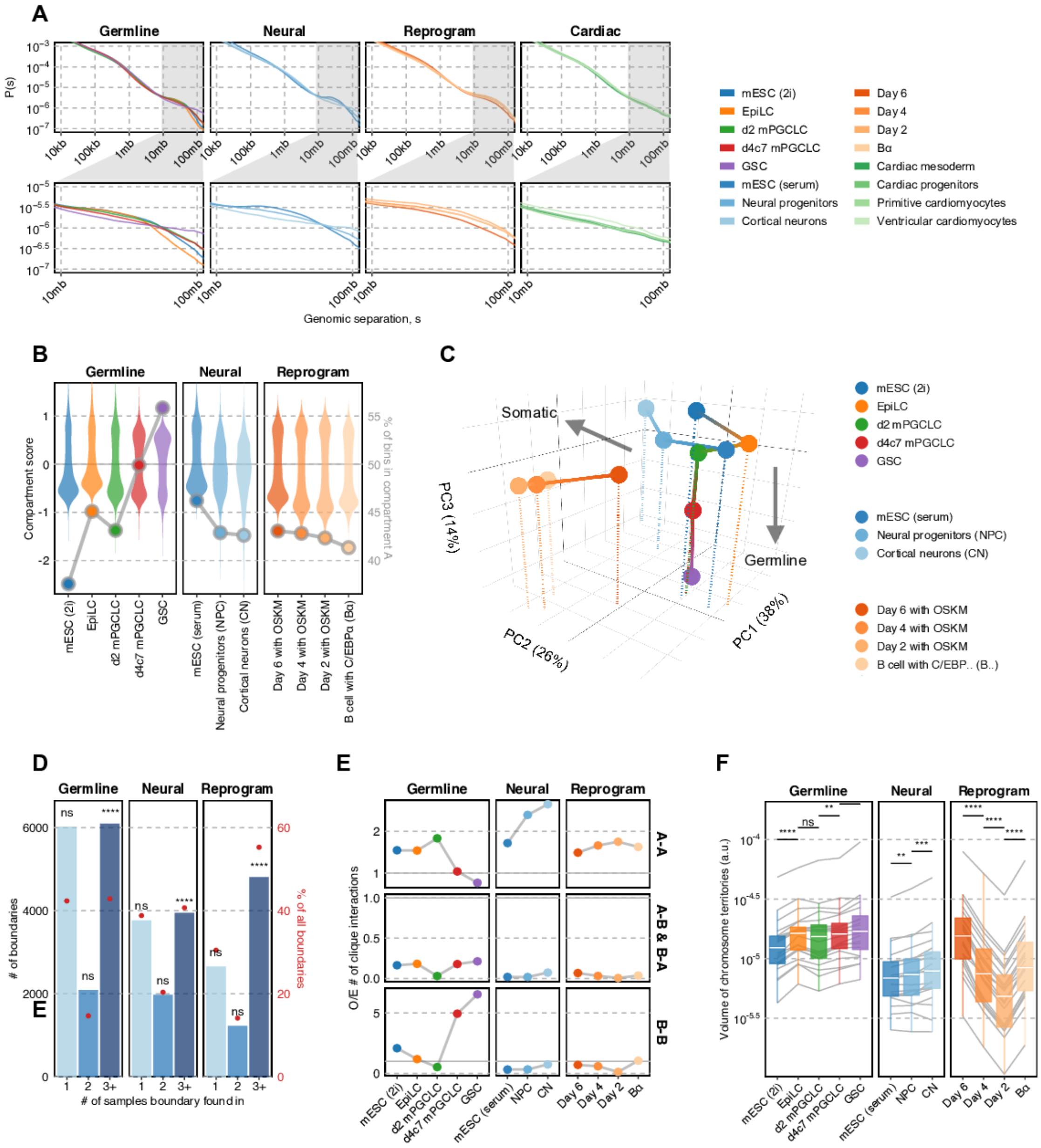
Data listed above are available from the published article linked below:
https://doi.org/10.15252/embj.2022110600

Appendix B

Additional supplementary information for chapter 2

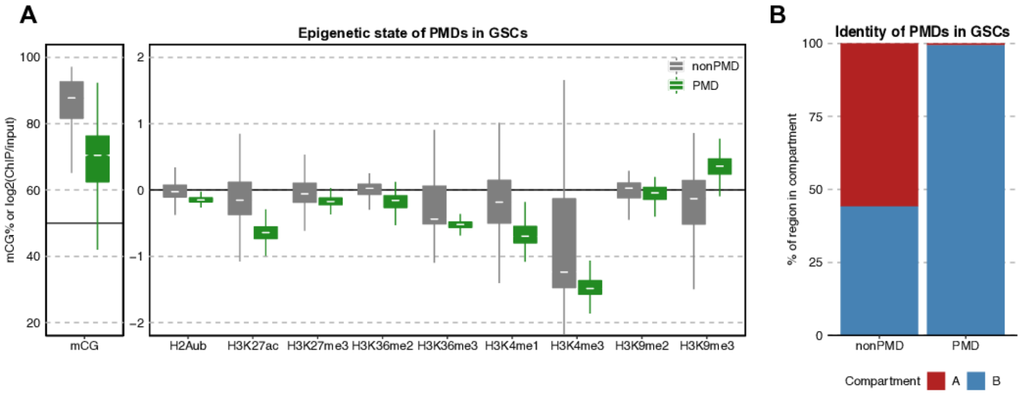
The following pages contain additional supplementary figures for chapter 2: NUCLEOME PROGRAMMING FOR THE FOUNDATION OF TOTIPOTENCY IN MAMMALIAN GERMLINE DEVELOPMENT.

Figure B.1. Hi-C analysis with other lineages.



(A) Contact probability decay across different inter-loci separation distances for different lineages (neural induction (Bonev et al, 2017); B cell reprogramming (Stadhouders et al, 2018)), demonstrating a gain of distal interactions along differentiation, especially at distances >50 Mb.
(B) Transitions in euchromatin-vs-heterochromatin bias during the course of different lineages at 100 kb resolution. (left axis: violin plots) Distribution of compartment scores; (right axis: dots) ratio of A:B compartment bins.
(C) PCA of compartment scores at 100 kb resolution comparing different lineages. Somatic differentiation is mostly reflected in PCs 1 & 2, while germ cell differentiation manifests in PC3.
(D) Degree of TAD boundary conservation in different lineages. Consistent across different lineages, more than 40% TAD boundaries are significantly conserved across differentiation. One-sided permutation tests were carried out by shuffling sample labels 100000 times, with p-values (left to right, top to bottom): 1, 1, 1e-5, 1, 1, 1e-5, 1, 1, 1e- 5.
(E) Enrichment of TAD-TAD interactions involved in max cliques (size ⩾3) during the development of different lineages.
(F) Convex hull volumes of CSynth-produced chromosome 3D models during the development of different lineages, after normalization to unit backbone length. n = 19. P- values are computed using Wilcoxon rank-sum tests. Wilcoxon signed-rank test p-values (left to right): 1.91e-6, 1.89e-1, 1.69e-3, 2.61e-4, 1.69e-3, 1.68e-4, 1.91e-6, 1.91e-6, 1.91e-6.

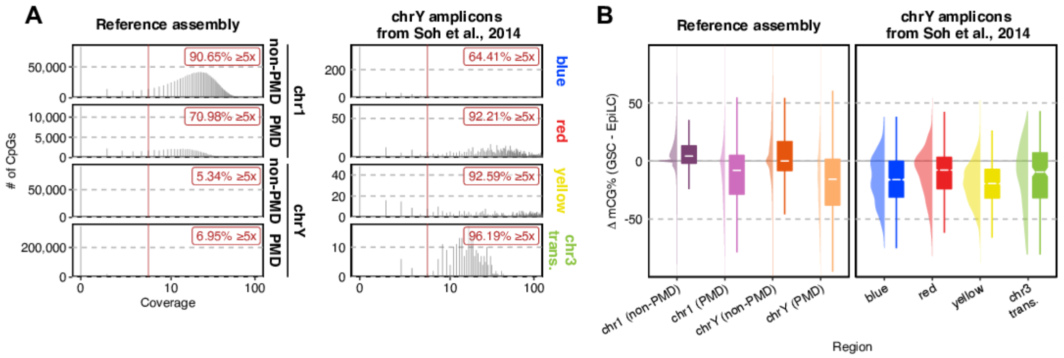
Figure B.2. PMD analysis with epigenome and higher-order chromatin structure.



(A) %mCG and enrichment of various histone modifications in PMDs and non-PMDs.

(B) Proportion of A/B compartment in PMDs and non-PMDs.

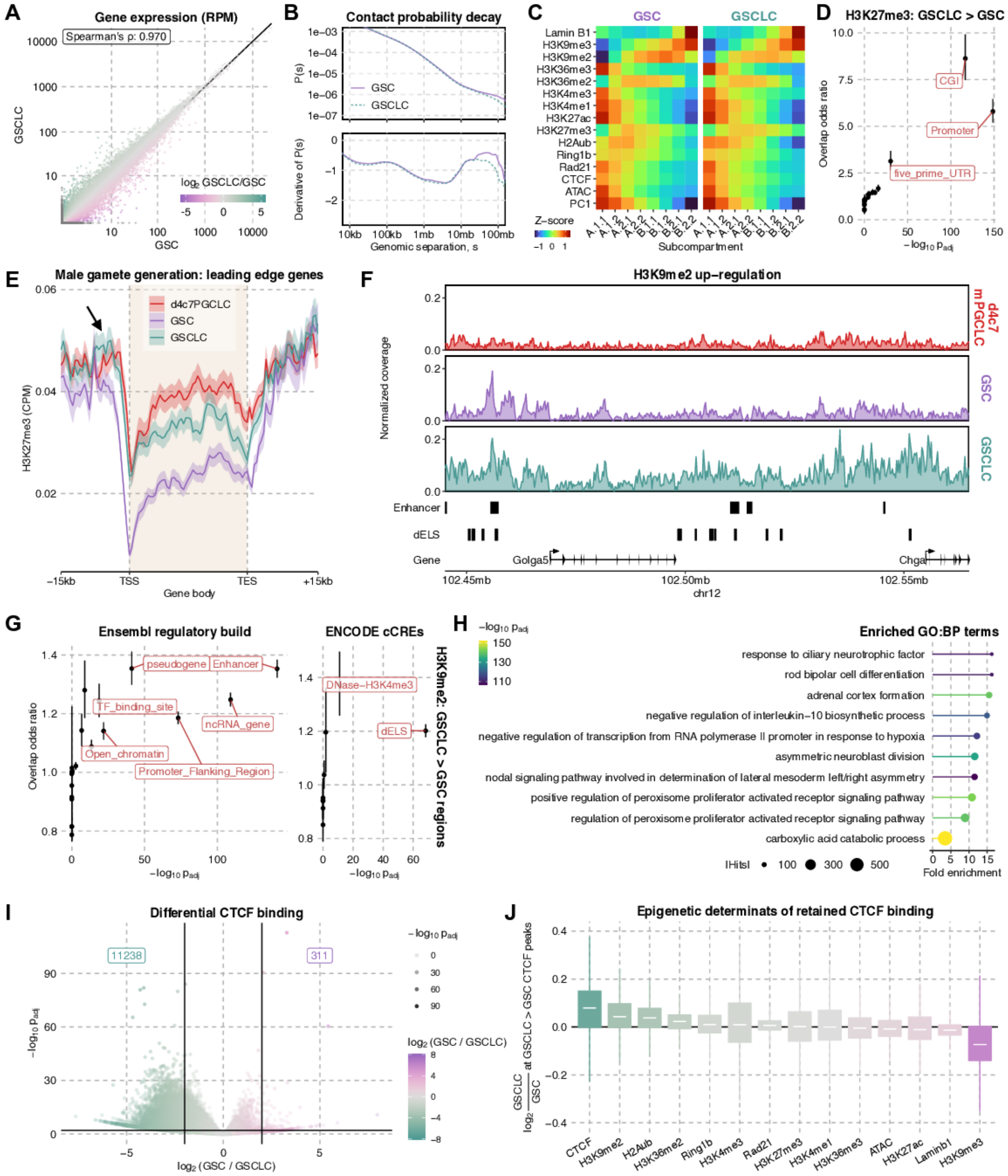
Figure B.3. Methylome analysis on Y chromosome using alternative mapping method.



(A) Histogram of the PBAT coverage in different genomic regions using two alignment methods. (Left) In the conventional mapping approach of alignment against the reference genome, CpGs on chromosome Y exhibit significantly lower coverage than those on autosomes. (Right) Through direct mapping to the ampliconic sequences covering more than 80% of chromosome Y (Soh et al, 2014), most CpGs are now well-covered.

(B) Differential methylation within and outside GSCs' PMDs. Whereas GSCs are methylated at a level comparable to EpiLCs outside of PMDs, chromosome Y (most of which are PMDs) is found to be substantially hypomethylated using both conventional (left) and direct mapping (right).

Figure B.4. Nucleome differences between functional GSCs and spermatogenically-impaired GSCLCs.



(A) Scatter plot of gene expression values between GSCLCs and GSCs based on 3'-seq. (B) Contact probability decay across different inter-loci separation distances. In agreement with the observations from chromosome-wide contact maps, GSCLCs have reduced distal interaction frequencies as compared to GSCs.
(C) Enrichment patterns of epigenetic signals in an 8-state model yield comparable types of states between cell types.
(D) Regions with higher H3K27me3 in GSCLCs than GSCs predominantly correspond to CpG islands and promoters based on overlap enrichment analysis against the Ensembl Regulatory build. The point marks the mean while error bars indicate standard errors. (E) Metagene plots of H3K27me3 for leading edge genes of "male gamete generation" in d4c7 mPGCLCs, GSCs and GSCLCs. The thick line marks the mean while the upper and lower limits indicate standard errors.
(F) Example locus of GSCLC-specific enrichment of H3K9me2 marking intergenic cis- regulatory elements.
(G) Overlap enrichment analysis of regions with higher H3K9me2 in GSCLCs than GSCs against the Ensembl Regulatory build (left) and ENCODE cCREs (right). Elevated H3K9me2 mostly affects distal enhancer elements. The point marks the mean while error bars indicate standard errors.
(H) Pathway enrichment analysis of distal elements with enhancer-like signatures ("dELS") from the ENCODE cCRE database overlapping regions with higher H3K9me2 in GSCLCs than GSCs.
(I) Volcano plot of differential CTCF binding sites. Scatter plot of CTCF enrichment across all peaks in GSCs and GSCLCs, with 11238 peaks substantially higher levels of CTCF in GSCLCs than GSCs.
(J) Correlation between differential CTCF binding and enrichment of various epigenetic marks.

SUPPLEMENTARY REFERENCES

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A et al (2017) Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171: 557-572.e524

Soh S, Y.Q., Alföldi J, Pyntikova T, Brown G, Laura, Graves T, Minx J, Patrick, Fulton S, Robert, Kremitzki C, Koutseva N, Mueller L, Jacob et al (2014) Sequencing the Mouse Y Chromosome Reveals Convergent Gene Acquisition and Amplification on Both Sex Chromosomes. Cell 159: 800-813

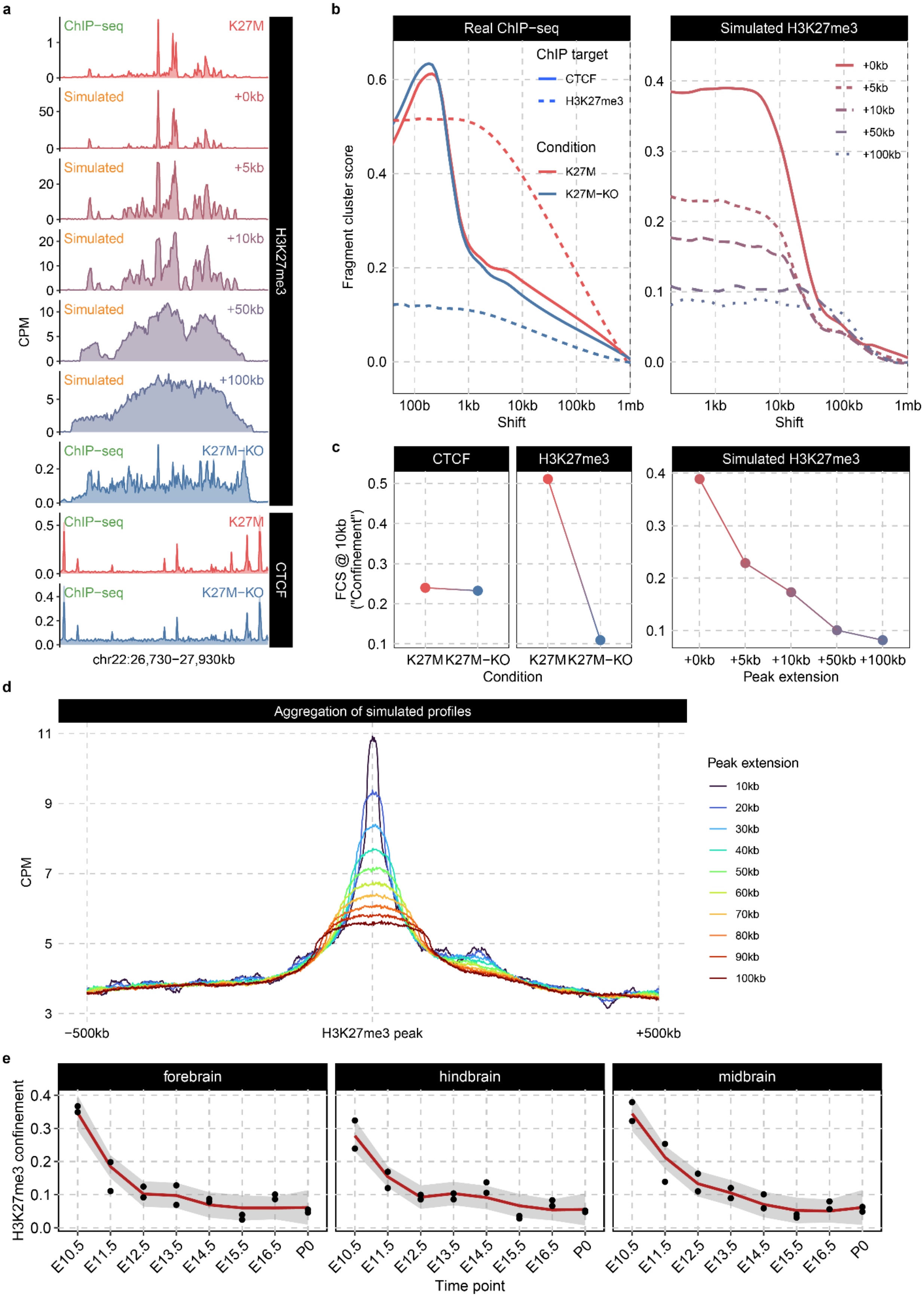
Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer C, Cuartero Y et al (2018) Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nature Genetics 50: 238-249

Appendix C

Supplementary information for chapter 3

The following pages contain supplementary figures for chapter 3: H3K27ME3 SPREADING ORGANIZES CANONICAL PRC1 CHROMATIN ARCHITECTURE TO REGULATE DEVELOPMENTAL TRANSCRIPTIONAL PROGRAM.

Figure C.1. Validation of framework for ChIP-seq signal breadth quantification



a. Experimental ChIP-seq coverage tracks of H3K27me3 and CTCF around a representative locus, where H3K27me3 is either confined or diffuse while CTCF remains consistently confined. Simulated H3K27me3 datasets with varying degrees of confinement (as generated by "ChIPs", see methods) are also shown

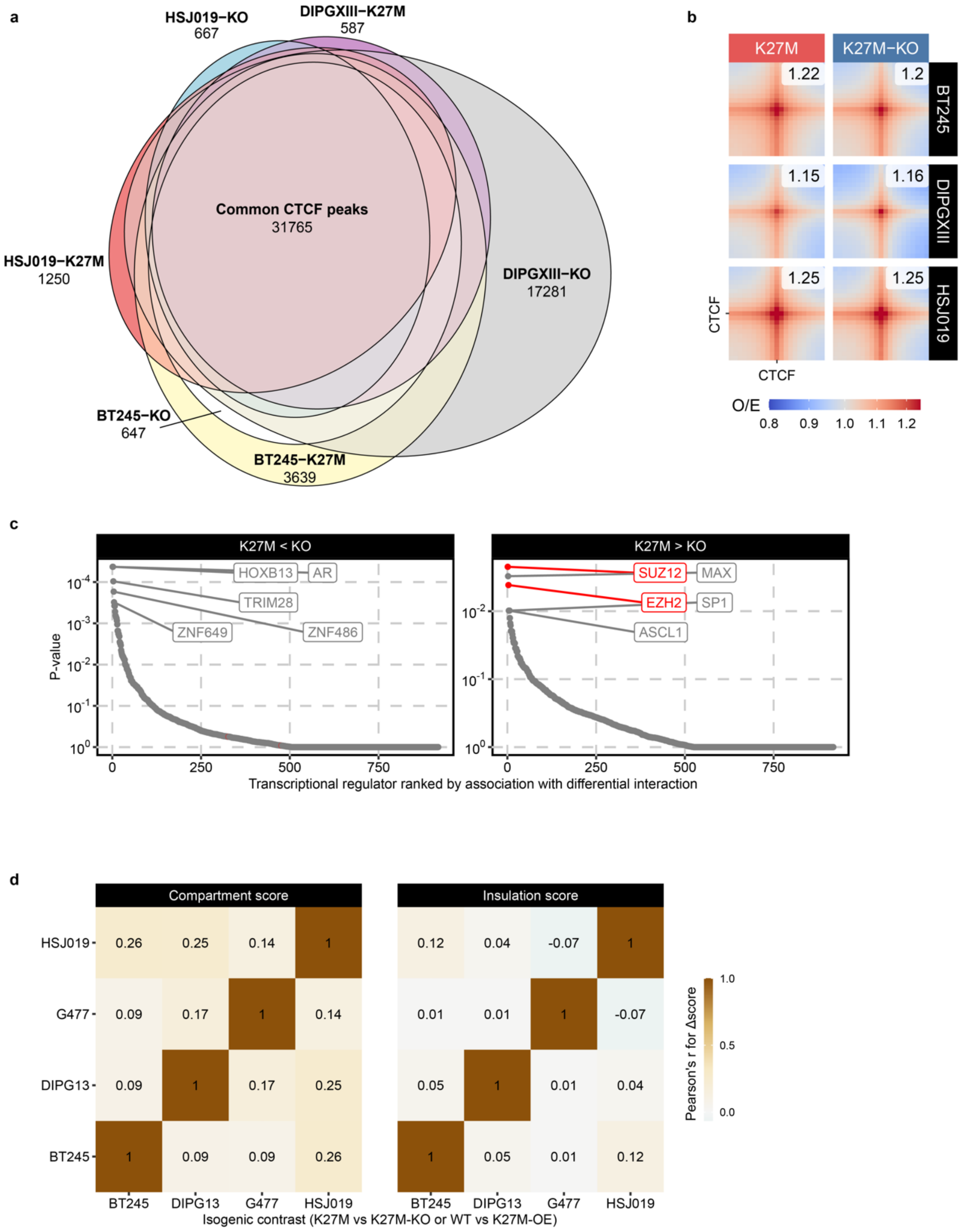
b. Genome-wide fragment cluster score computed at various shift distances for experimental and simulated ChIP-seq datasets, demonstrating the distinction between confined versus diffuse profiles of H3K27me3; note that values for CTCF appear largely unchanged between K27M and KO

c. Fragment cluster score specifically at 10kb shift distance, our choice for measuring "confinement", can quantitative distinguish confined versus diffuse ChIP-seq profiles

d. Metaplots showing aggregate depth-normalized H3K27me3 signals from simulated datasets with varying degrees of confinement, with hypothetically no difference in true modification levels at the very center. This reinforces that depth-normalization (e.g., CPM) of a more diffuse profile will yield the impression of a lower peak as compared to confined profile, despite no difference in the absolute value at the center (i.e., a by-product of ChIP-seq depth-normalization). This phenomenon can be important to consider when assessing normalized metaplots.

e. Confinement scores of H3K27me3 (fragment cluster score at 10kb, see methods) for *in vivo* samples from the developing mouse brain. Diminishing scores indicate the spread of H3K27me3 generally accompanies early brain development.

Figure C.2. Commonalities and differences between K27M and K27M-KO cells



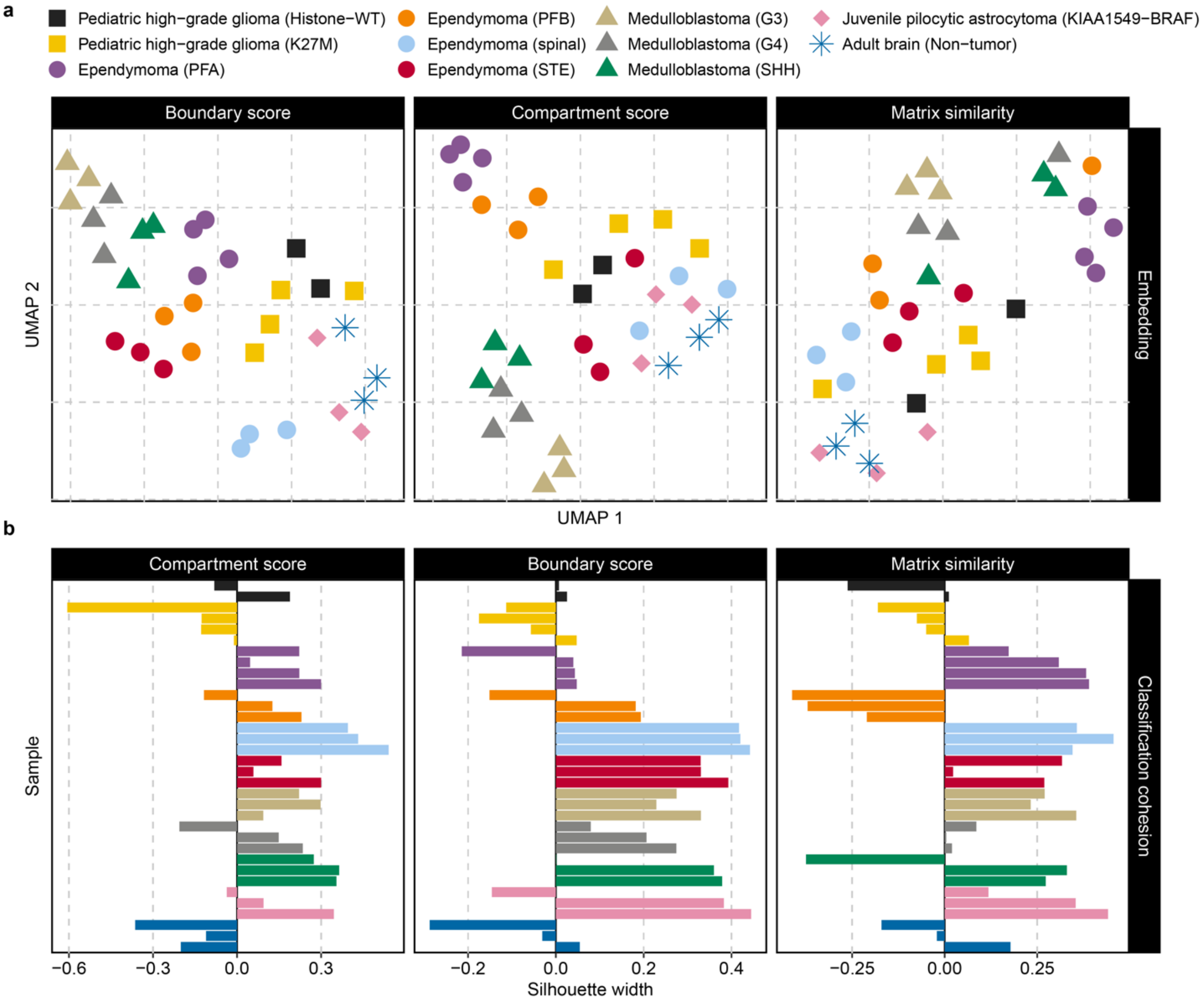
a. Euler diagram of CTCF peak calls for isogenic comparisons of K27M pHGG cell lines and their K27M-KO counterparts, demonstrating substantial overlap.

b. Pile-up of pairwise Hi-C interaction among the union CTCF peak set across all K27M and KO samples; only pairs of sites with convergent motif orientations were considered. This revealed a lack of global differences in CTCF interaction strength between isogenic K27M and K27M-KO pHGG cells.

c. BART3D analysis investigating transcriptional regulators whose binding sites are enriched in regions with differential interactions between isogenic K27M and K27M-KO comparisons. P-values are computed from robust rank aggregation of significances from three different cell lines, with lower p-value indicating consistently high-ranking transcriptional regulator (i.e., stronger consistent association with differential interaction across cell lines). Polycomb-related factors (e.g., EZH2, SUZ12) are among the most predictive of interactions preferentially enriched in K27M cells.

d. Correlation of compartment/insulation score differences (K27M versus KO/WT) between isogenic comparisons; the weak correlation coefficients demonstrate lack of consistent changes in compartment/domain structures upon the removal or overexpression of K27M.

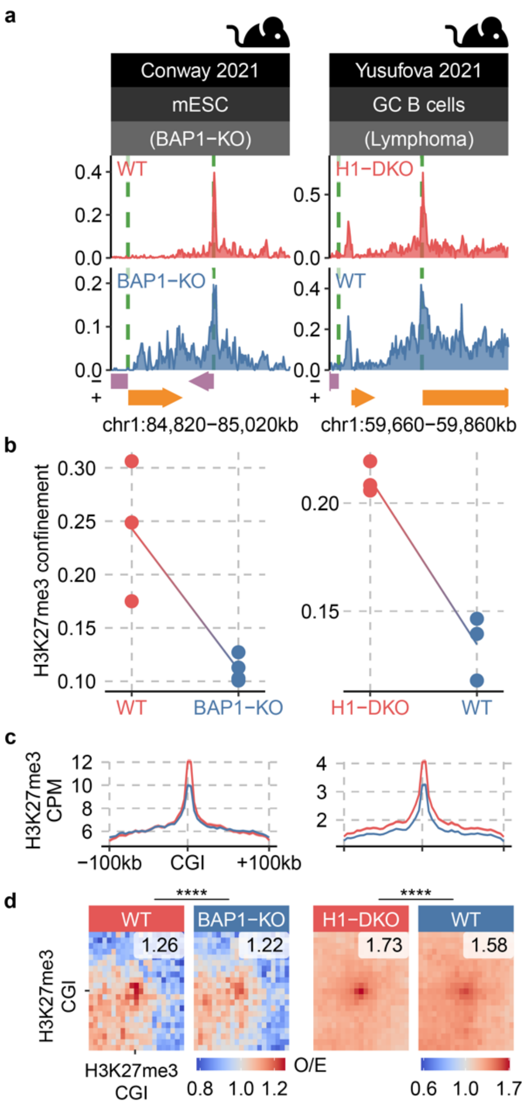
Figure C.3. Subgroup-specific 3D genome features



a. UMAP embedding based on genome-wide comparison across primary tissue (tumor and normal brain) Hi-C contact matrices at three different scales: compartmentalization (first principal component / compartment score), topologically associating domain organization (RobusTAD boundary score), and matrix similarity (HiCRep coefficient). While many tumor samples exhibit substantial clustering by known molecular clinical subtypes, K27M pHGGs appear more heterogenous and does not constitute a tight cluster in any of the three modalities examined.

b. Silhouette width based on inter-sample similarity in terms of three different modalities, with more positive values indicating that a sample is closer to other samples belonging to the same class whereas more negative samples indicating lack of cohesion (i.e., class label is not reflected by high inter-sample similarity for those belonging to the same class). K27M pHGGs emerge as the only tumor subtype demonstrating lack of distinct signatures across all three scales considered, generally showing negative silhouette scores (i.e., less similar to other K27M pHGGs than to tumours of another type). This indicates that K27M does not leave a specific signature on large-scale genome organization

Figure C.4. Relationship between restricted H3K27me3 and long-range inter-CGI interaction in additional contexts



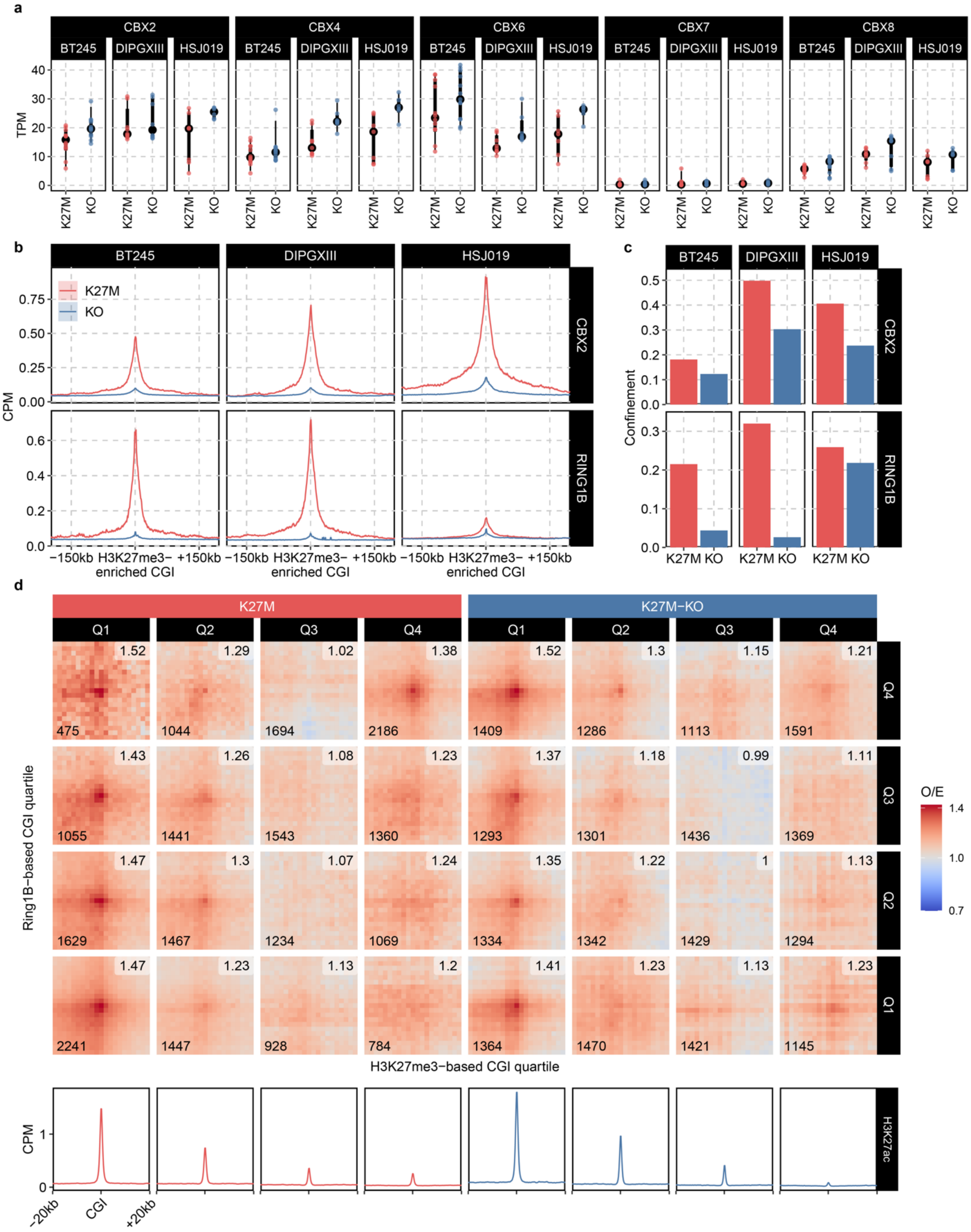
a. Genomic distribution of H3K27me3 (ChIP-seq coverage tracks in units of counts-per-million-alignments) at representative loci in mESC and germinal center B cells, demonstrating distinction of confined versus diffuse profiles. Focal H3K27me3 enrichment preferentially occurs near regulatory regions such as promoters and CpG islands.

b. Measure of H3K27me3 ChIP-seq signal confinement (fragment cluster score at 1kb separation, computed using the tool “ssp”, see methods) in diverse contexts confirming genome-wide distinction of confined versus diffuse profiles. Individual data points correspond to a replicate, with connected points indicating replicates from the same batch; connections not linking points indicate that multiple replicates were sequenced in a batch, and so the links are drawn between the average value per condition.

c. Metaplots of H3K27me3 aggregate ChIP-seq signals around H3K27me3-enriched CpG islands, normalized by total read depth. H3K27me3-enriched is defined as the union set of top 1000 CpG islands with the most H3K27me3 alignments in either condition.

d. Pile-up of Hi-C interaction among H3K27me3-enriched CpG islands, as defined in c., portraying average pairwise contact strength between such regions (in units of enrichment, i.e., observed / expected). Punctate enrichment signal in the center indicates elevated long-range interaction anchored at H3K27me3-enriched CGIs in cells with confined H3K27me3

Figure C.5. Characterizing determinants of long-range polycomb-mediated interaction



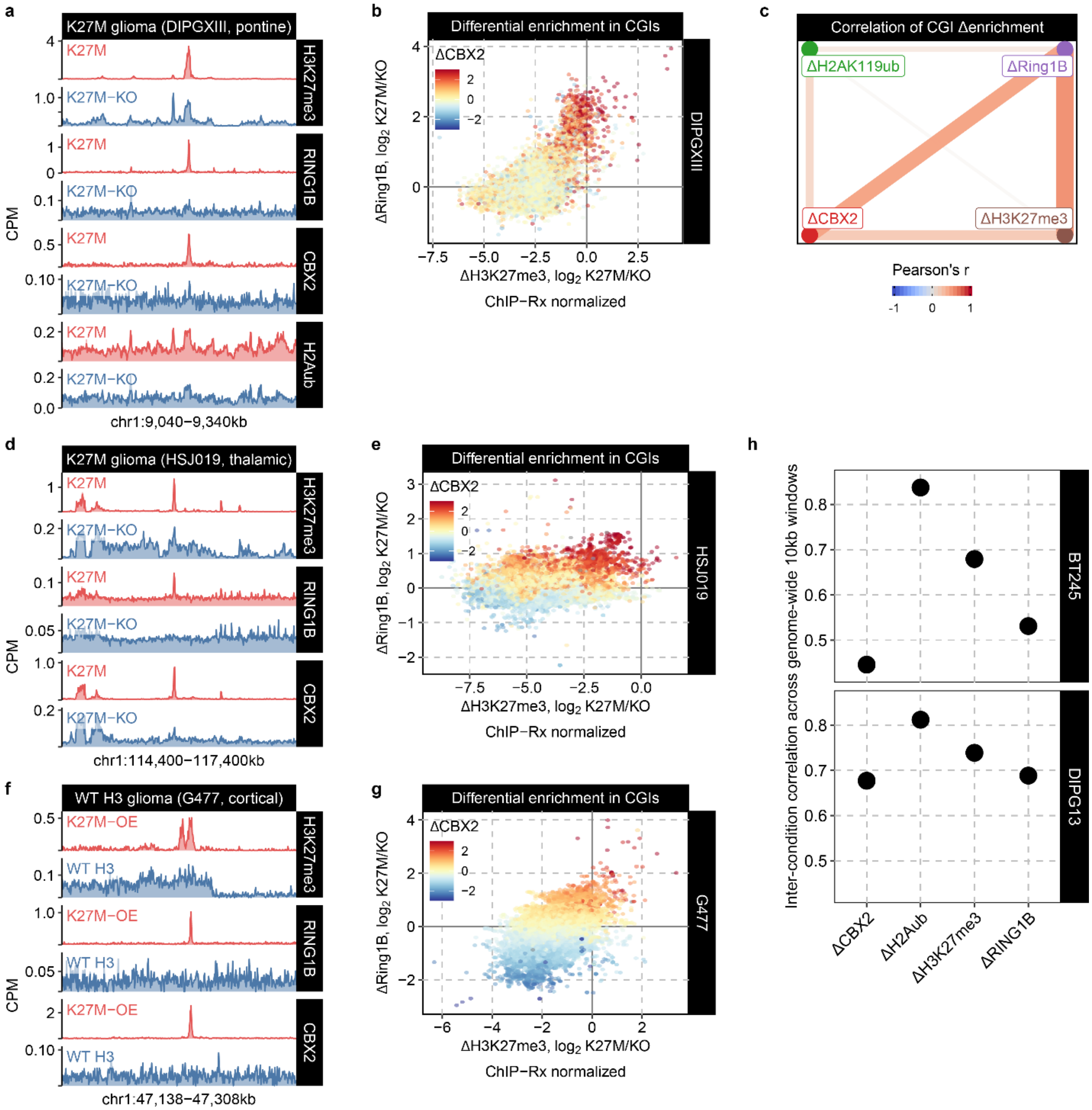
a. Expression of PRC1-subunit CBX proteins in pHGG K27M cell lines based on bulk RNA-seq. CBX2 is consistently expressed at a high level across all three lines.

b. Metaplot of PRC1 aggregate ChIP-seq signal around H3K27me3-enriched CpG islands (union set of top 1000 most enriched in both conditions per cell line, as defined previously), normalized by read depth. PRC1 occupancy from PRC2 target sites are consistently diluted.

c. Metric of RING1B/CBX2 ChIP-seq signal confinement (fragment cluster score at 10kb, see methods) in all three cell lines confirm global dilution of PRC1 signals upon removal of K27M.

d. Pile-up of Hi-C interaction for pairs of CpG islands belonging to the different pairs of quartiles of H3K27me3 and RING1B enrichment in K27M pHGG cell line BT245. Greatest K27M-specific contact enrichment of distal looping involves CGIs with both highest (i.e. Q4) H3K27me3 and RING1B. CGIs lacking H3K27me3 (Q1) and enriched for H3K27ac also engage in strong pairwise interactions, with or without RING1B binding, corresponding to active cis-regulatory elements.

Figure C.6. Consistent H3K27me3 and cPRC1 confinement across diverse backgrounds



a. Representative locus of cPRC1 dilution following H3K27me3 spread, reproducible in another pHGG K27M cell line, DIPGXIII.

b. Correlation of signal enrichment differences in CpG islands between K27M and K27M-KO, showing that CBX2 is most enriched in regions with K27M-specific enrichment of both H3K27me3 and RING1B, confirming strong association between H3K27me3 confinement with enhanced cPRC1 recruitment in another pHGG K27M cell line, DIPGXIII.

c. Correlation network of differential enrichment of H3K27me3, RINGB1, CBX2 and H2AK119ub, showing the weak correlation between H2AK119ub and the rest three, implicating cPRC1 rather than ncPRC1 determines the difference between K27M and KO in another pHGG K27M cell line, DIPGXIII.

d. As (a), except for an additional pHGG K27M cell line, HSJ019.

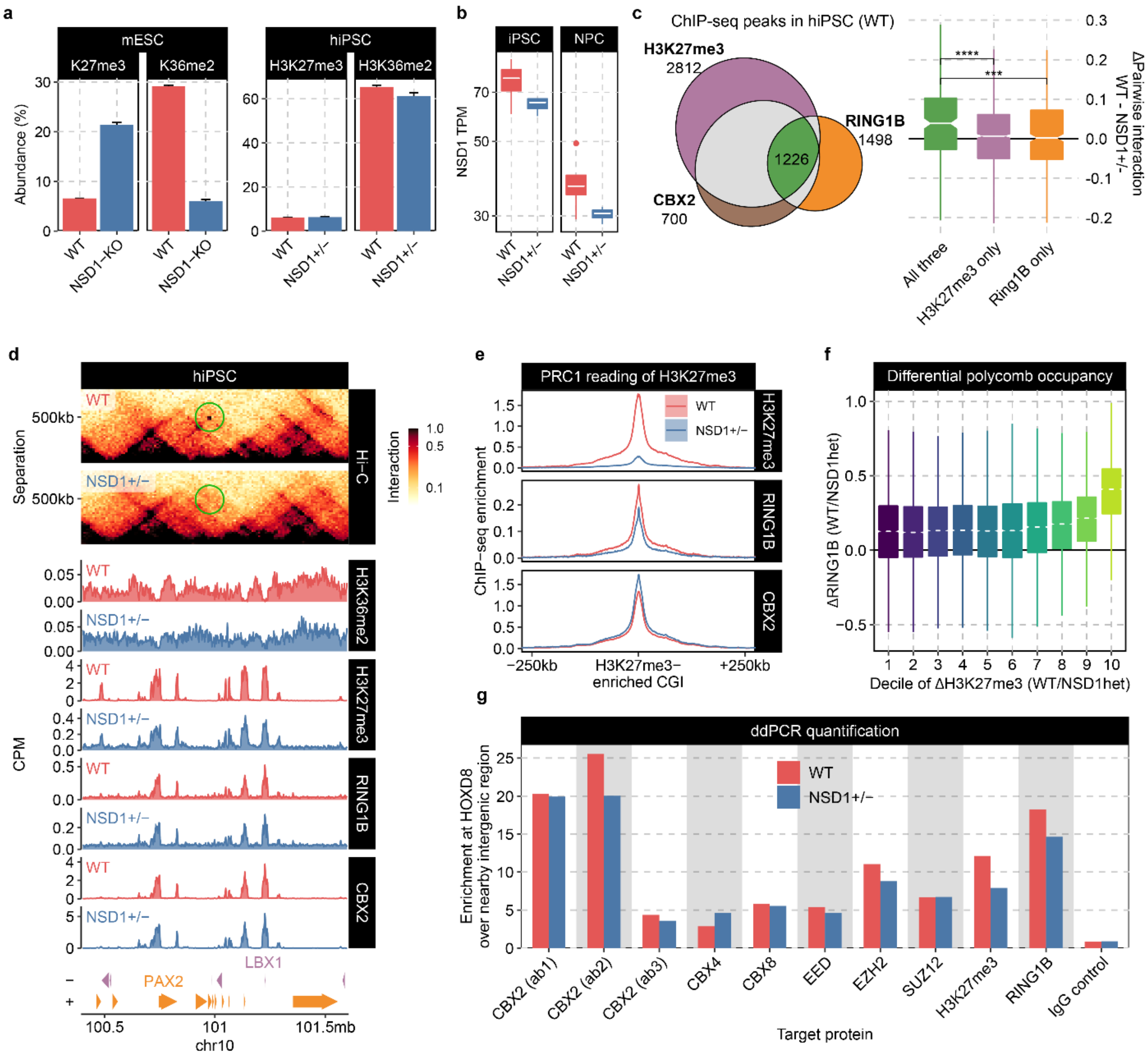
e. As (b), except for an additional pHGG K27M cell line, HSJ019.

f. As (a), except for a pHGG WT H3 cell line, G477.

g. As (b), except for a pHGG WT H3 cell line, G477.

h. Global correlation of signal enrichment confirming H2AK119ub as being the most correlated (i.e., least different) between K27M and K27M-KO cells.

Figure C.7. Weakening polycomb-mediated chromatin architecture in pluripotent stem cells through loss of NSD1



a. Global abundance of H3K27me3 and H3K36me2 measured by quantitative histone mass spectrometry in PSCs, confirming H3K36me2 depletion upon full or partial loss of NSD1 and corresponding elevation of H3K27me3 in mESCs and no change in abundance in NSD1+/- hiPSCs.

b. Expression of NSD1 in NCRM1 hiPSCs and NPC, validating the down-regulation of NSD1 transcripts in NSD1+/- cells. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

c. (left) Euler diagram of peak call at CGIs for H3K27me3 and PRC1 sub-units CBX2 & RING1B in WT hiPSCs, confirming CBX2's reader function in localizing cPRC1 to H3K27me3-enriched regions in PSCs; (right) differential long-range interaction strength (loop score computed for pairs of regions within 20kb-2mb) for various peak overlap subsets, revealing that sites marked by all three (H3K27me3, RING1B, CBX2) preferentially engage in strong distal interaction in WT hiPSCs as compared to NSD1+/- cells.

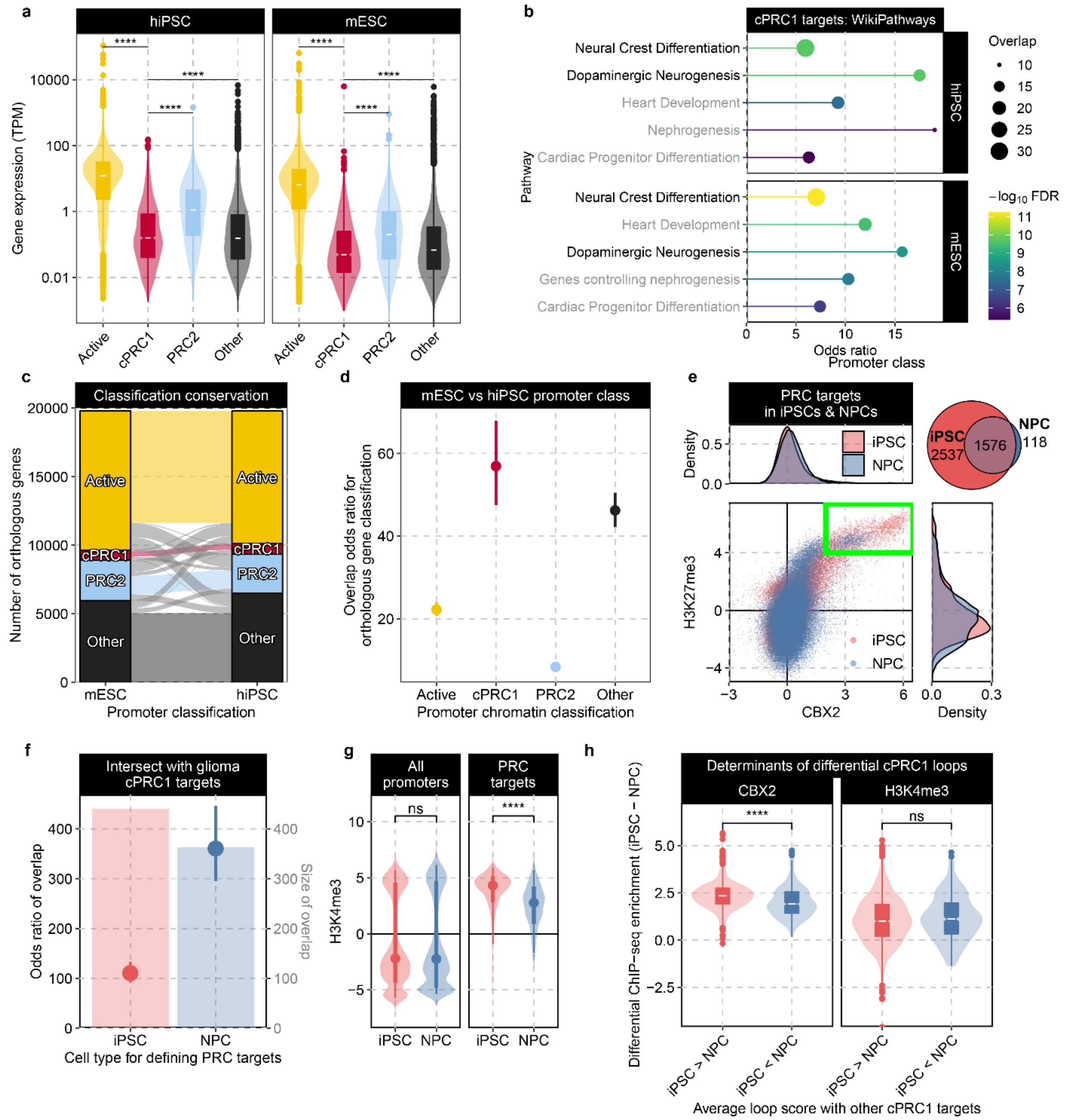
d. Representative locus of differential interaction between cPRC1 binding sites bridging the promoters of PAX2 and LBX1, along with ChIP-seq profiles of H3K36me2, H3K27me3, RING1B, and CBX2. Whereas moderate spread of H3K27me3 accompanies modest depletion of H3K36me2 in NSD1+/- cells, PRC1 binding is also demonstrably more diffuse.

e. Metaplot of H3K27me3 and PRC1 aggregate ChIP-seq signal around H3K27me3-enriched CpG islands, in units of log2 enrichment over input, at H3K27me3-enriched CGIs (union set of top 1000 most enriched in both conditions, as defined previously). Limited differences in PRC1 ChIP-seq were observed, further investigated with greater quantitative sensitivity in panel g.

f. Differential RING1B binding in CpG islands stratified by differential H3K27me3 between WT and NSD1+/- hiPSCs. Coordinated depletion of RING1B and H3K27me3 is observed, especially for CGIs with the greatest loss of H3K27me3.

g. Droplet digital PCR measurement of CUT&RUN library enrichment at HOXD8 (cPRC1 target gene) versus intergenic region to quantify the degree of epitope confinement versus dilution. Antibodies with low enrichment indicate poor target recognition and are deemed less reliable. Strong enrichment confirms H3K27me3, RING1B, EZH2 dilution from cPRC1 target in NSD1+/- compared to WT iPSCs. Result for CBX2 was sensitive to antibody choice (1: CST 18687 – modest dilution, 2: CST E3N6A – substantial dilution, 3: Novus NBP247524 – low quality).

Figure C.8. Properties cPRC1 sites as compared to other chromatin states in stem cells



a. Expression of genes associated with the promoters from the four clusters, demonstrating the lower expression levels of cPRC1 targets as compared to PRC2 targets. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

a. Enrichr pathway over-representation analysis of cPRC1 targets in PSCs, consistently identifying similarly enriched pathways: development and neuron differentiation.

b. Alluvial plot confirming substantial overlap of promoter chromatin state across WT hiPSCs and mESCs for orthologous genes.

c. All classes exhibit greater-than-expected inter-species conservation, but cPRC1 is especially conserved as compared to other classes. Intervals correspond to 95% confidence intervals around odds ratio estimates.

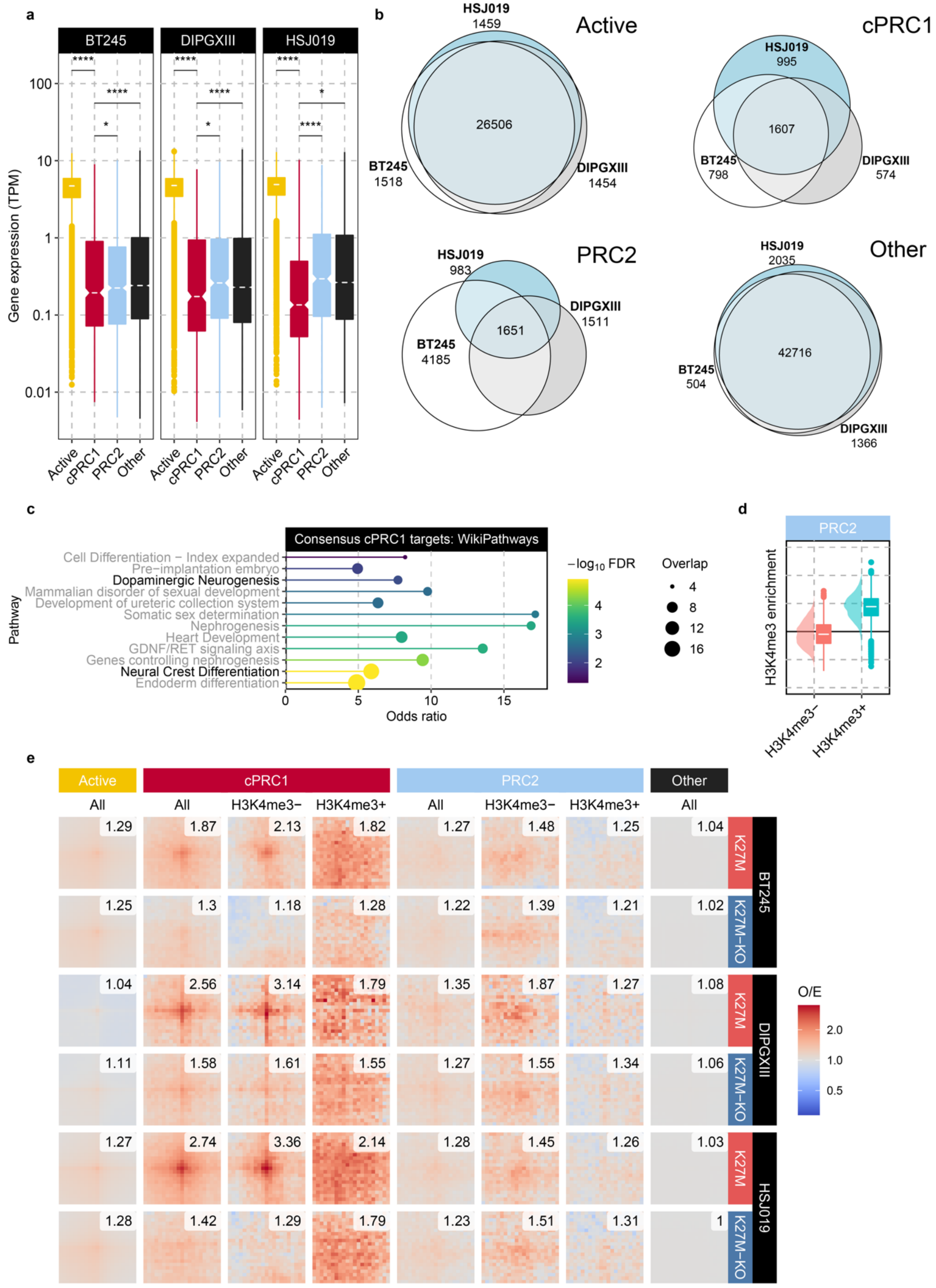
d. Joint plot depicts definition of PRC targets in NCRM1 hiPSCs and derived NPCs based on enrichment of both H3K27me3 and CBX2 using fixed thresholds. Input-normalization alone was deemed sufficient seeing that the overall enrichment distributions shown the margins are not drastically difference across cell types. Venn diagram indicates that PRC targets in NPCs are generally a subset of those identified in iPSCs.

e. Despite fewer in number, PRC target genes in NPCs more strongly overlap cPRC1 targets identified in K27M pHGG cells, consistent with the importance of neurodevelopmental genes.

f. PRC target genes in NPCs do not display the same bimodal H3K4me3 pattern as K27M pHGG cPRC1 targets, indicating lack of H3K4me3 as a distinct signature of genes sensitive to H3K27M-dependent looping. Points indicate median value, with a thicker band describing the 66th percentile, whereas the thin line extends to 95th percentile.

g. The average interaction score between every iPSC cPRC1 targets all neighbouring cPRC1 targets were calculated in both iPSCs and NPCs, after which the difference in average cPRC1 interaction score is taken; cPRC1 targets with greater average loop score in iPSCs as compared to NPCs were separated from those with weaker loop scores. Regions involved in a decrease of cPRC1 looping along differentiation simultaneously demonstrated greater loss of CBX2, consistent with the involvement of cPRC1 in the differential looping. In contrast, differential H3K4me3 was not predictive of differential cPRC1 looping, a departure from the K27M pHGGs.

Figure C.9. Properties cPRC1 sites as compared to other chromatin states in glioma cells



a. Expression of genes associated with the promoters from the four clusters, demonstrating the lowest transcription levels in the cPRC1 cluster. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

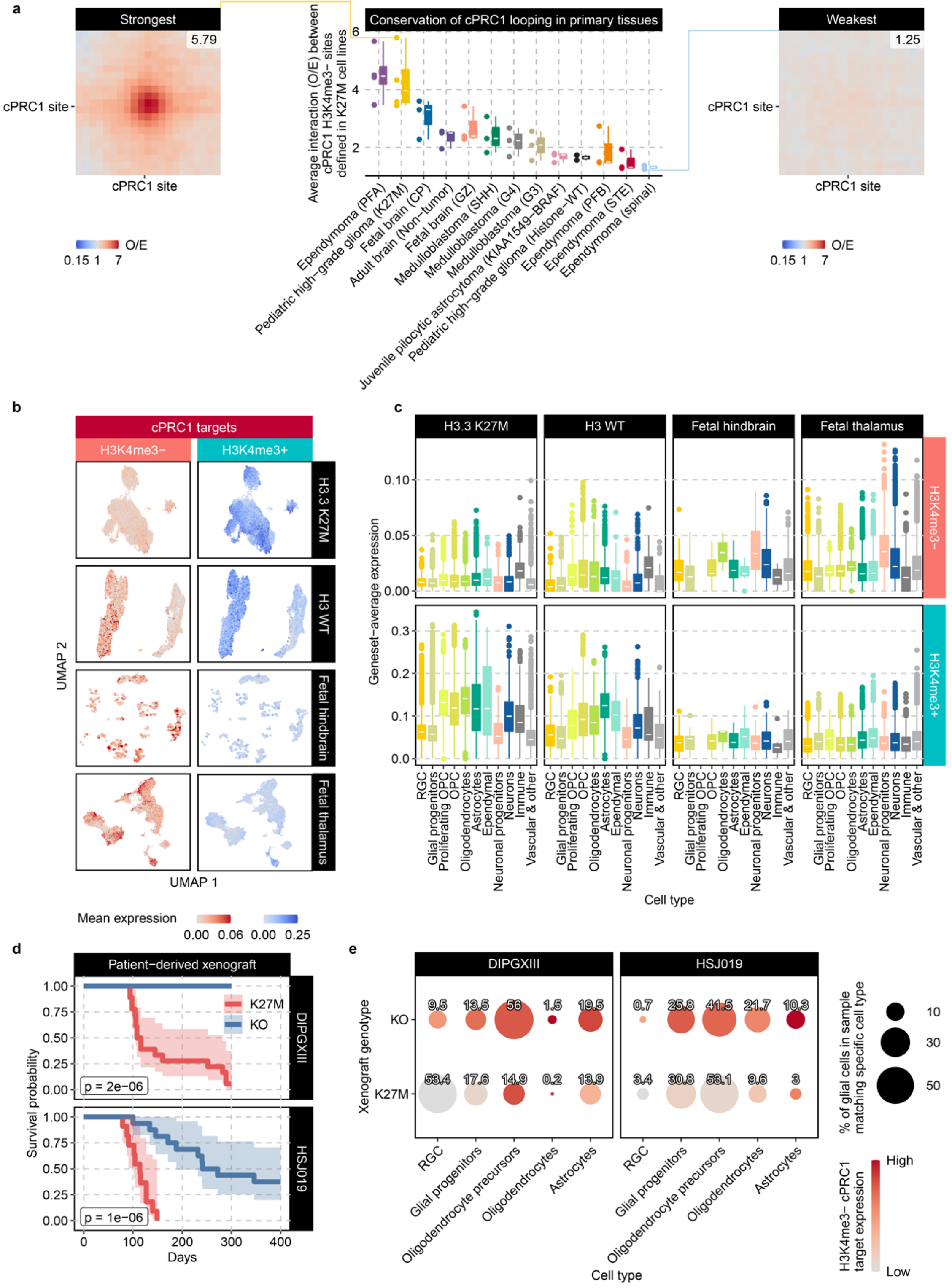
b. Euler diagram of sites from the four clusters showing concordance of "Active" and "Other" sites among the three K27M pHGG cell lines, and less for "cPRC1" and "PRC2" sites.

c. Enrichr pathway over-representation analysis of consensus cPRC1 targets among three K27M pHGG cell lines, demonstrating the enrichment in development and neuron differentiation.

d. Distribution of H3K4me3 in H3K4me3+ and H3K4me3- sub-clusters among PRC2 target sites.

e. Pile-up of Hi-C interaction were computed among pairs of genomic regions belonging to the same cluster (i.e., intra-class looping) for three different isogenic pHGG K27M cell lines. H3K4me3- cPRC1 targets consistently emerge as forming the strongest loops across all three cell lines.

Figure C.10. Validation of cPRC1 looping and target repression in primary tumours



a. Central enrichment (i.e., observed/expected value of the central 3x3 set of pixels) for pile-up of pairwise Hi-C interactions in primary tissues between regions consistently labelled as H3K4me3- cPRC1 targets among three K27M pHGG cell lines. Strongest interaction is found in K27M pHGGs and PFA EPNs, followed by the fetal brain. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

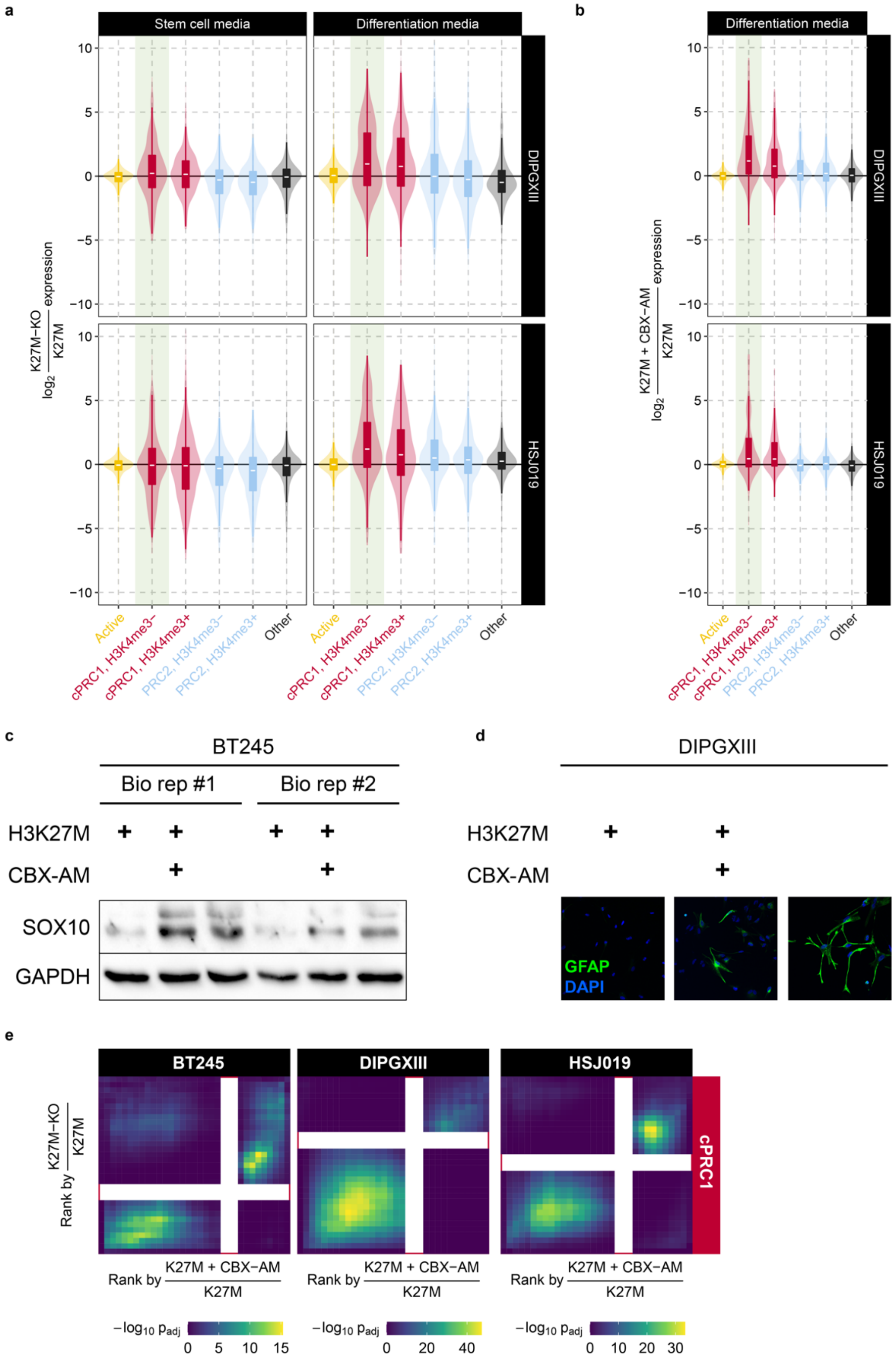
b. cPRC1 H3K4me3- targets are repressed in a homogenous manner across various cell types revealed by scRNA-seq in K27M pHGG, whereas select subpopulations appear to show elevated expression of those genes in WT pHGG and fetal brains, demonstrating the association between repression of cPRC1 H3K4me3- target genes and polycomb body compaction.

c. Mean expression of cPRC1 target genes per cell, averaged over all cells with the same cell type label, from scRNA-seq of primary tissue. H3.3K27M uniquely demonstrates repression of H3K4me3- cPRC1 targets.

d. Kaplan-Meier survival analysis (with 95% confidence interval in a lighter shade) for xenograft mice using two other pHGG K27M cell lines, displaying loss of tumour formation by K27M-KO cells in DIPGXIII, and substantially greater latency and decreased penetrance of tumours by K27M-KO cells in HSJ019.

e. Comparison of scRNA-seq from K27M and K27M-KO PDOXs confirm increased H3K4me3- cPRC1 target gene expression is accompanied by an increase in the proportion of more differentiated cell types.

Figure C.11. Decreased cPRC1 compaction reliably alleviates developmental blockade



a. Differential expression, specifically of cPRC1, H3K4me3- targets (green shading), become heightened after differentiation, recapitulated in additional K27M pHGG cell lines DIPGXIII and HSJ019. Boxplots' hinges correspond to the 25th and 75th percentiles, with whiskers extending to the most extreme value within 1.5 × interquartile range from the hinges, whereas the central band mark the median value.

b. Up-regulation of cPRC1, H3K4me3- targets (green shading) from CBX-AM treatment of K27M cell cultures, in additional K27M pHGG cell lines DIPGXIII and HSJ019.

c. Up-regulation of differentiation marker SOX10 by western blot observed in both K27M-KO and CBX-AM treated K27M cells in pHGG line BT245.

d. Up-regulation of differentiation marker GFAP observed in both K27M-KO and CBX-AM treated K27M cells in pHGG line DIPGXIII, by immunofluorescence.

e. Rank-rank hypergeometric overlap of differential expression for cPRC1 target genes between parental K27M pHGG cells versus K27M-KO or CBX-AM treated cells. Expressional changes induced by CBX-AM and K27M-KO were found to be significantly correlated across three different cell lines.

Appendix D

Copyright permissions

Copyright permissions have been obtained for all figures adapted from previous publications.

Select figures are licensed under the Creative Commons Attribution 4.0 International License or the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of these licenses, please visit the linked webpages or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Additional permissions have been obtained from Springer Nature, Oxford University Press, and Elsevier with the following license numbers:

- 5282171050176
- 5282171015352
- 5282170979067
- 5282170759703
- 5282170661528
- 5282170617699
- 5282170580839
- 5282170547432
- 5282170513172
- 5282170466735
- 5282170421049
- 5282170376648

- 5282160566933
- 5282160529465
- 5282160483752
- 5282160424658
- 5282160345119
- 5282160255634
- 5282160122252
- 5282160013943
- 5282141232740

Full license agreements are available upon request.