DISS. ETH NO. 29264

Characterization and modelling of the epigenetic dynamics during the transition from naïve to primed pluripotency

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

JOÃO PEDRO AGOSTINHO DE SOUSA

MSc Biochemistry, University of Lisbon, Portugal

born on 06.04.1989

accepted on the recommendation of

Prof. Dr. Ferdinand von Meyenn, ETH Zurich, Switzerland

Prof. Dr. Ori Bar-Nur, ETH Zurich, Switzerland

Prof. Dr. Mark Robinson, University of Zurich, Switzerland

2023

[Page intentionally left blank]

Acknowledgements

First, I would like to thank my parents, brother, and grandparents. They were always supportive and encouraging in every stage of my life. Their sacrifice and focus on providing me with an advanced education are why I am finishing my doctoral degree. I am forever grateful. The same extends to all members of my family, friends, and teachers.

I would also like to give a special thank you to my girlfriend, Jocsana. She has been my main support during my doctoral studies and the reason I could push forward and finish the dissertation. She gave advice and encouraged me every step of the way. I could not ask for a better partner in my life.

To my colleagues from the Instituto Superior Técnico, Faculdade de Ciências, Instituto de Medicina Molecular, my master's degree supervisor Ana Rita Grosso, and the principal investigator from my master's internship lab Sérgio Fernandes de Almeida, thank you for your advice and teachings. As I progress in my career, I understand how foundational it is to work with wonderful people.

A special thank you to all members of the von Meyenn lab for their advice and friendship during my doctoral studies. I also thank the members of the Austin Smith lab and Wolf Reik lab for discussions, suggestions, and guidance on the data analysis and manuscript structure.

To my collaborators from the Stegle Group, Steffen Rulands group, and Wolf Reik lab, in particular, Marc Jan Bonder and Maria Rostovskaya, I thank you for your advice during my project on how to improve my bioinformatic analysis; on how to interpret the results; and on how to prepare the manuscripts.

I thank Maike Paramor and Vicki Murray from the Stem Cell Institute Genomics Facility for helping with library preparation from the samples used during my research. Simon Andrews and Felix Krüger from the bioinformatics facility at the Babraham Institute for helping with data mapping and handling and for providing advice on how to analyse multi-omics data; and Jason Ernst from UCLA for giving me helpful tips regarding chromHMM.

I also want to thank Ferdinand von Meyenn and Ori Bar-Nur for their support and supervision during my doctoral research. They showed me how to work and navigate in an academic environment and keep a positive outlook. Valuable lessons for my future career.

Finally, I would like to thank the professors, postdocs, and administration staff from ETH Zurich for making my doctoral studies more enjoyable and easier to manage.

Table of Contents

Summary 5

Résumé 7

List of Figures 9

List of Tables 10

1. Introduction 11

1.1. Motivation 11

1.2. Human embryonic development 12

1.3. Human pluripotent stem cells (hPSCs) 14

1.3.1. Naïve hPSCs 14
1.3.2. Primed hPSCs 15
1.3.3. Advantages and disadvantages of naïve and primed hPSCs 15
1.3.4. Capacitation 16

1.4. Epigenetic characteristics of human embryonic development 17

1.4.1. Epigenetics 17
1.4.2. Histone modifications in mammalian embryo development 19
1.4.3. DNA methylation in early human development 20
1.4.4. Promoter bivalency in early human development 20
1.4.5. Gene imprinting 21

1.5. X chromosome inactivation (XCI) 22

1.5.1. XCI in early mammalian development 22
1.5.2. XCI in human pluripotent stem cells 23
1.5.3. XCI erosion 24

1.6. Sequencing methods 24

1.6.1. RNA sequencing (RNA-seq) 24
1.6.2. Chromatin immunoprecipitation sequencing (ChIP-seq) 25
1.6.3. Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq) 25
1.6.4. Post-bisulfite adaptor tagging (PBAT) 26

1.7. Bioinformatic analysis 26

1.7.1. Multi-omics factor analysis 26
1.7.2. Chromatin-state discovery with ChromHMM 27

1.8. Aims of the dissertation 27

2. Results 29

2.1. Experimental setup 29

2.2. Multi-omics landscape of naïve and primed human pluripotent stem cells (hPSCs) 30

2.2.1. Sample clustering 30

2.2.2. Genome-wide and feature-specific epigenetic modification levels 31

2.2.3. Imprinting control regions 32

2.2.4. Genome-wide epigenetic changes across conditions and association between epigenetic modifications 34

2.3. Gene expression and promoter analysis 35

2.3.1. Characterization of differentially expressed genes between conditions 35

2.3.2. Modelling gene expression based on promoter epigenetic modification levels 37

2.3.3. Promoter bivalency 38

2.4. Epigenetic dynamics between pluripotency states 39

2.4.1. Combinatory role of epigenetic modifications 39

2.4.2. Multi-omics factor analysis (MOFA) 40

2.4.3. Enrichment analysis of MOFA factor 1 over annotated genomic regions 41

2.4.4. Motif enrichment in differentially accessible regions 44

2.5. X chromosome-specific epigenetic landscape 45

2.5.1. Chromosome-wide epigenetic modification levels 45

2.5.2. Epigenetic differences between conditions 46

3. Discussion 48

4. Conclusion 57

5. Outlook 58

6. Material and methods 59

6.1. Cell culture 59

6.1.1. Cell lines 59

6.1.2. Human pluripotent stem cell maintenance 59

6.1.3. Capacitation 59

6.2. Data generation 60

6.2.1. RNA sequencing (RNA-seq) library preparation 60

6.2.2. Chromatin Immunoprecipitation (ChIP) and sequencing (ChIP-seq) 60

6.2.3. Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) library preparation 61

6.2.4. Whole-Genome Bisulfite Sequencing (WGBS) 62

6.2.5. Western blot ... 62

6.3. Data processing ... 62

6.3.1. Genome build and annotation ... 62

6.3.2. Read alignment and quantification ... 63

6.3.3. RNA-seq gene read counts ... 63

6.3.4. ChIP-seq and ATAC-seq read counts in genomic regions ... 64

6.3.5. Methylation percentage in CpGs and genomic regions ... 64

6.3.6. Differentially expressed genes and clusters ... 64

6.3.7. Differential ChIP-seq and ATAC-seq peaks ... 65

6.4. Data analysis ... 65

6.4.1. Principal component analysis and hierarchical clustering ... 65

6.4.2. Multi-omics factor analysis (MOFA) ... 65

6.4.3. Chromatin state discovery ... 66

6.4.4. Enriched genomic regions ... 66

6.4.5. Multivariate regression analysis ... 66

6.4.6. Gene function profiling ... 67

6.4.7. Bivalency and developmental genes ... 67

6.4.8. Imprinting control regions allelic analysis ... 67

6.4.9. Motif enrichment ... 68

6.4.10. Transposable elements expression ... 68

7. Supplemental figures ... 69

8. Tables ... 79

Bibliography ... 98

Appendix A – Description of genomic features ... 115

Appendix B – Abbreviations ... 119

Appendix C – Code and data availability ... 122

Appendix D – Scientific Contributions and Collaborations ... 123

Appendix E – Funding ... 124

Appendix F – Publications ... 125

Appendix G – Curriculum Vitae ... 128

Summary

Human pluripotent stem cells (hPSCs) show great potential for regenerative medicine applications, but several limitations must be addressed before their widespread clinical use. These limitations include epigenetic instability, which can lead to spontaneous differentiation or genomic instability, and heterogeneity among hPSC lines, resulting in cells with different and unexpected properties. In addition, there are safety concerns regarding the potential for teratoma formation, immune rejection, and prolonged culture of hPSCs–which may result in X chromosome-linked erosion, referring to the gradual loss of the silencing of one of the two X chromosomes in female cells. This erosion can cause increased expression of X-linked genes, affecting developmental and physiological consequences that can limit their differentiation potential and, consequently, their therapeutic efficacy. To overcome these limitations, it is necessary to improve our understanding of the genetic and epigenetic mechanisms that regulate hPSCs and develop culture conditions that more closely resemble the *in vivo* environment. Recent advances in human naïve culture conditions offer potential solutions for these challenges. Naïve human pluripotent stem cells maintained in culture resemble pre-implantation epiblast cells and represent an earlier developmental state than conventionally cultured hPSCs, which are in a primed state of pluripotency and resemble post-implantation epiblast cells. Compared to naïve cells, conventional hPSCs have a more restricted developmental potential and are closer to differentiation. We recently developed new culture conditions that allow naïve hPSCs to transition to the primed state of pluripotency within a time frame that mimics early development. We refer to this transition as "capacitation". However, it is uncertain whether capacitated naïve hPSCs can accurately replicate the gene expression and epigenetic patterns observed in the primed state after acquiring multi-lineage differentiation ability. Furthermore, it is unknown whether resetting conventionally cultured primed hPSCs to the naïve state and then allowing them to capacitate in the primed state could potentially address epigenetic anomalies such as X-chromosome inactivation erosion. Aiming to clarify these questions, we employed sequencing assays to characterize and model gene expression, DNA methylation, histone modifications, and chromatin accessibility landscapes of naïve, capacitated, and long-term cultured hPSCs. This dissertation presents and interprets the results of this multi-omics analysis, intended to enhance our understanding of the role of epigenetics in pluripotency, identify new targets for future research, and determine the advantages of naïve hPSCs capacitation. The results show that resetting conventionally cultured primed hPSCs to the naïve state followed by repriming restores the epigenetic characteristics of X chromosome inactivation from an eroded landscape. Additionally, they emphasise the importance of CpG islands, enhancers, and retrotransposons as hotspots of epigenetic dynamics between pluripotency states. Finally, they highlight the putative role of H3K27ac modifications in promoters of genes with naïve-specific expression. This study underscores the potential of improved hPSC culture methods for developing novel regenerative therapies and provides insights into the molecular mechanisms underlying epigenetic changes during capacitation. These findings have implications for generating clinically relevant cell types and contribute to the ongoing progress in regenerative medicine, laying the foundation for future investigations in optimizing human pluripotent stem cell culture methods.

Résumé

Les cellules souches pluripotentes humaines (CSPh) présentent un grand potentiel pour les applications de la médecine régénérative, mais plusieurs limites doivent être adressées avant leur utilisation clinique généralisée. Ces limites comprennent une instabilité épigénétique, qui peut conduire à une différenciation spontanée ou une instabilité génomique, ainsi qu'une hétérogénéité parmi les lignées de CSPh, ce qui résulte en des cellules ayant des propriétés différentes et inattendues. De plus, il existe des préoccupations de sécurité concernant le potentiel de formation de tératomes, le rejet immunitaire et la culture prolongée des CSPh - qui peuvent entraîner une érosion liée au chromosome X, se référant à la perte progressive de la silencing de l'un des deux chromosomes X dans les cellules femelles. Cette érosion peut causer une expression accrue de gènes liés au chromosome X, affectant les conséquences développementales et physiologiques qui peuvent limiter leur potentiel de différenciation et, par conséquent, leur efficacité thérapeutique. Pour surmonter ces limitations, il est nécessaire d'améliorer notre compréhension des mécanismes génétiques et épigénétiques qui régulent les CSPh et de développer des conditions de culture qui ressemblent davantage à l'environnement *in vivo*. Les récents progrès dans les conditions de culture naïves offrent des solutions potentielles à ces défis. Les cellules souches pluripotentes humaines naïves maintenues en culture ressemblent aux cellules épiblastiques pré-implantatoires et représentent un état de développement antérieur à celui des CSPh cultivées de manière conventionnelle, qui se trouvent dans un état de pluripotence primée et ressemblent aux cellules épiblastiques post-implantatoires. Comparées aux cellules naïves, les CSPh conventionnelles ont un potentiel de développement plus restreint et sont plus proches de la différenciation. Nous avons récemment développé de nouvelles conditions de culture qui permettent aux CSPh naïves de passer à l'état de pluripotence primée dans un délai qui imite le développement précoce. Nous appelons cette transition "capacitation". Cependant, on ne sait pas si les CSPh naïves capacitées peuvent reproduire avec précision l'expression génique et les motifs épigénétiques observés dans l'état primé après avoir acquis la capacité de différenciation multi-lignée. De plus, on ignore si le retour à l'état naïf des CSPh cultivées de manière conventionnelle et ensuite la capacitation à l'état primé pourrait potentiellement résoudre des anomalies épigénétiques telles que l'érosion de l'inactivation du chromosome X. Pour clarifier ces questions, nous avons utilisé des tests de séquençage pour caractériser et modéliser l'expression génique, la méthylation de l'ADN, les modifications d'histones et les paysages d'accessibilité de la chromatine des CSPh naïves, capacitées et cultivées à long terme. Cette thèse présente et interprète les résultats de cette analyse multi-omique, destinée à améliorer notre compréhension du rôle de l'épigénétique dans la pluripotence, à identifier de nouvelles cibles pour la recherche future et à déterminer les avantages de la capacitation des CSPh naïves. Les résultats montrent que le retour à l'état naïf des CSPh cultivées de manière conventionnelle, suivi d'une repriming, restaure les caractéristiques épigénétiques de l'inactivation du chromosome X à partir d'un paysage érodé. De plus, ils soulignent l'importance des îlots CpG, des enhancers et des rétrotransposons en tant que points chauds de la dynamique épigénétique entre les états de pluripotence. Enfin, ils mettent en évidence le rôle putatif des modifications H3K27ac dans les promoteurs des gènes à expression spécifique des CSPh naïves. Cette étude souligne le potentiel des méthodes améliorées de culture des CSPh pour développer de nouvelles thérapies régénératives et fournit des informations sur les mécanismes moléculaires sous-jacents aux changements épigénétiques lors de la capacitation. Ces résultats ont des implications pour la génération de types cellulaires cliniquement pertinents et contribuent aux progrès en cours en médecine régénérative, posant les bases pour de futures investigations visant à optimiser les méthodes de culture de cellules souches pluripotentes humaines.

List of Figures

Figure 1. Diagram of the experimental design. 29
Figure 2. Principal component analysis for each sequencing assay and dimensionality reduction using all sequencing data. 31
Figure 3. Genome-wide and feature-specific epigenetic modification levels. 32
Figure 4. DNA methylation and epigenetic modification allele-specific level differences in imprinting control regions. 33
Figure 5. Overview of genome-wide epigenetic changes and interplay between epigenetic modifications across conditions. 35
Figure 6. Differentially expressed gene clusters and associated promoter epigenetic modifications and gene ontology terms. 36
Figure 7. Association between gene expression and promoter epigenetic modifications. 38
Figure 8. Comparison of promoters containing H3K4me3 and H3K27me3 marks in H9 cells and associated gene ontology terms. 39
Figure 9. Chromatin state model of genomic distribution of histone modification and chromatin accessibility data across conditions. 40
Figure 10. Analysis of MOFA2 factor 1 that distinguishes naïve and primed cells. 41
Figure 11. Analysis of epigenetic changes and transposable element expression between pluripotency states. 43
Figure 12. Epigenetic profiles and motif enrichment of accessible regions undergoing significant changes between naïve and primed states. 44
Figure 13. Epigenetic modification enrichment and XIST expression in the X chromosome. 46
Figure 14. Differential levels of H3K27me3 and H3K9me3 in the X chromosome across experimental conditions. 47

Figure S1. 71
Figure S2. 72
Figure S3. 74
Figure S4. 76
Figure S5. 78

List of Tables

Table 1. Number of biological replicates for each condition. 80

Table 2. List of human allele-specific differentially methylated regions (DMRs), indicating known, novel, and placental-specific imprinted regions. 83

Table 3. List of genes classified by their association with naïve pluripotency, primed pluripotency, post-implantation epiblast, and lineage-specific expression. 84

Table 4. List of genes and their corresponding protein products that belong to the Polycomb-group proteins (PcG), Trithorax-group proteins (TrxG), and the DNMT, TET, and IDH protein families. 89

Table 5. Gene ontology terms that were significantly enriched from genes belonging to the RNA sequencing clusters. Each row represents a different gene ontology term. Only gene ontology terms with a significant enrichment (adjusted p-value < 0.05) are included in the table. 94

Table 6. List of software packages used in data processing and analysis. 97

1. Introduction

1.1. Motivation

Pluripotency describes a dynamic cellular state, conferring the potential to develop into all embryonic lineages. Pre- and post-implantation epiblast cells exist for a brief period of embryogenesis and are both pluripotent. However, they have distinct characteristics. Human pluripotent stem cells (hPSCs) can be grown in conditions that represent these two stages of embryogenesis, known as naïve and primed pluripotent stem cells, respectively[1,2]. Epigenetic modifications are among the primary mechanisms that cells use to maintain their pluripotency status and identity during development and differentiation, ensuring that specialization and determination occur in a unidirectional manner[3]. However, studying the epigenetic dynamics of pluripotency transitions during human embryonic development has been challenging. Much of our current understanding comes from research on model organisms, such as mice, which can be limiting because their gestational period, metabolism, transcriptional characteristics, and genetic composition may differ from those of humans. Thus, it is crucial to focus on human cells to gain a better understanding of pluripotency transition during human development[4–7]. A recent study revealed that naïve and primed hPSCs possess pluripotent state-specific Polycomb-associated interaction networks and specific enhancer activity and interactivity–reflected by their distinct epigenetic landscapes[8]. However, studies that aim to characterize epigenetic modifications in hPSCs often use long-term cultured conventional cells, and those cells have some drawbacks. For example, some long-term cultured cells show X-chromosome inactivation erosion[9], which may lead to incorrect reproduction of the epigenetic characteristics of post-implantation epiblast cells. My collaborators have recently developed cell culture conditions that mimic key features of the developmental transition of human epiblast cells, specifically the transition from the pre-implantation to the early gastrulation stage[10]. These conditions allow us to observe and study the developmental changes occurring during this critical period of human embryonic development. By understanding the cellular and molecular mechanisms that underlie this transition, we can gain insights into the fundamental processes that drive cell fate decisions and differentiation. In this study, my colleagues and I aimed to investigate the epigenetic dynamics of human pluripotency using this capacitation system that enables naïve hPSCs to transition to a primed state of pluripotency. To achieve this, we employed a range of sequencing assays to characterize and model gene expression, DNA methylation, histone modifications, and chromatin accessibility landscapes of naïve, capacitated, and long-term cultured hPSCs. Our main objective was to generate global epigenetic maps of active and repressive histone modifications, chromatin accessibility, and

DNA methylation from naïve and capacitated human pluripotent stem cells. By correlating these data with our previously reported gene expression datasets from the same system, we aimed to provide an overview of the epigenetic dynamics between the pluripotency states and highlight the differences between conventionally cultured and capacitated hPSCs. By conducting a multi-omic analysis of the datasets, my goal is that the findings presented in this dissertation highlight the mechanisms that underlie the transition from naïve to primed pluripotency in human cells, elucidate the role epigenetic modifications play in this process, and determine whether capacitated cells have improved characteristics compared to conventionally cultured primed hPSCs. I believe that the study and improvement of stem cell culture systems are critical for advancing our understanding of human embryonic development. By characterizing the epigenetic changes that occur during pluripotency transitions, we may be able to develop more efficient and effective methods for differentiating stem cells into specific cell types for use in regenerative medicine.

1.2. Human embryonic development

Human early development refers to the period of growth and development that occurs from conception to birth. This process is divided into several stages, each with its unique characteristics. The germinal stage begins at fertilization and lasts for about two weeks. During this stage, the fertilized egg (or zygote) divides rapidly and forms a blastocyst. This stage marks the beginning of embryonic development and is characterized by rapid cell division[11]. Afterwards, the embryonic stage begins and lasts until the end of the eighth week of gestation. During this stage, the embryo undergoes rapid growth and development. Major organs and systems begin to form, including the heart, lungs, nervous and digestive systems. It is also the stage during which sex differentiation occurs[12]. The final stage is the foetal stage. This stage begins at the end of the embryonic stage and lasts until birth. During this stage, the foetus continues to grow and develop, and major systems become more refined and functional. By the end, the foetus is fully formed and ready for birth[13].

The human pluripotent stem cells studied in this dissertation represent the epiblast cells of both the end of the germinal stage and the beginning of the embryonic stage, divided by implantation. The germinal stage, the first stage of human development, can be divided into three substages: the zygote stage, which lasts from fertilization to the blastocyst formation and takes place in the fallopian tube. After several rounds of division, the fertilized egg forms a cluster of cells called a morula; the morula stage, which lasts from about day 4 to 5 after fertilization, is characterized by the formation of the blastocyst; and the blastocyst stage, composed of an outer layer of cells called the trophoblast, which will give rise to the placenta, and an inner cell mass (ICM), which will give rise to the embryo. This stage lasts from day 5 to 14 after fertilization and results in the differentiation of the ICM into two distinct cell types: the epiblast and the hypoblast. Human pluripotent stem cells, such as our naïve and primed hPSCs, are derived from the ICM cells of blastocysts. The ICM cells are pluripotent, meaning they have the potential to differentiate into any of the three germ layers (ectoderm, mesoderm, and endoderm) and ultimately give rise to all the cells of the human body. In contrast, trophoblast cells give rise to the extraembryonic tissues, such as the placenta and umbilical cord. These cells are not pluripotent as they are restricted in their differentiation potential and can only give rise to specific cell types[14].

As mentioned, primed human pluripotent stem cells are derived from the ICM, just like naïve human pluripotent stem cells[15]. However, primed cells have undergone some degree of commitment towards a specific lineage, resembling cells of the epiblast stage of post-implantation development[16]. The epiblast undergoes a process of lineage commitment and differentiation towards the three germ layers during the early stages of post-implantation development. This process is regulated by various signalling pathways and transcription factors, which activate or repress specific genes to drive differentiation. During this developmental stage, the embryo undergoes gastrulation, in which the epiblast layer invaginates and forms a primitive streak. This primitive streak will give rise to the mesoderm and endoderm layers, while the ectoderm is formed from the remaining epiblast cells. The mesoderm will form the muscles, bones, and circulatory system, while the endoderm will form the gut and associated organs such as the liver and pancreas. In addition to its role in germ layer formation, the epiblast also forms the amniotic sac, which surrounds and protects the developing embryo. The amniotic sac is filled with amniotic fluid, which provides a cushion for the developing embryo and helps to regulate its temperature and protect it from mechanical shocks. Overall, this stage is a critical phase of human embryonic development, as it prepares for the formation of all the body's major tissues and organs[14]. Research into the mechanisms that govern the maintenance and transition from the naïve to the lineage-committed primed pluripotent state represents a significant opportunity to advance our comprehension of early human development. This stage is difficult to study because it occurs during the first few days after fertilization, which is an inaccessible period for research due to ethical and practical constraints. Therefore, advancements in the maintenance and manipulation of pluripotent stem cells *in vitro* offer a promising approach to modelling early human development.

1.3. Human pluripotent stem cells (hPSCs)

1.3.1. Naïve hPSCs

Naïve human pluripotent stem cells (hPSCs) are derived from the inner cell mass of pre-implantation embryos and represent an earlier stage of embryonic development, as mentioned above. They have an immature epigenetic state and are more plastic in their ability to differentiate into various cell types than primed hPSCs–having the potential to differentiate into all three germ layers (ectoderm, endoderm, and mesoderm) and extra-embryonic tissues[17–19]. However, they require a period of adaptation to culture in conventional human pluripotency stem cell media, such as mTESR, FGF/KSR, or E8[20,21]. In contrast, primed hPSCs are more restricted in their differentiation potential. They have undergone some degree of differentiation towards specific lineages and have a more mature epigenetic state[2].

Naïve hPSCs exhibit unique molecular and physical features that distinguish them from primed cells. They have an epigenetic landscape similar to pre-implantation epiblast cells and express high levels of pluripotency markers, including NANOG, OCT4, and SOX2, showing low levels of differentiation markers such as CDX2 and GATA6[22]. Furthermore, they express stage-specific genes, including KLF4, essential for maintaining their pluripotent state[23], and TFCP2L1, believed to regulate their self-renewal[24]. Their maintenance depends on the activation of the JAK/STAT3 pathway[25,26], commonly stimulated by LIF, and the inhibition of the MEK/ERK and GSK3β pathways, which have been demonstrated to promote self-renewal and sustain pluripotency[27,28]. In addition to these molecular features, naïve hPSCs also possess distinct physical characteristics from primed hPSCs. They exhibit a rounder and flatter morphology, while primed hPSCs tend to be more elongated and compact[29].

The discovery of these cells has been a significant achievement in regenerative medicine, offering a new model to study the regulation of pluripotency. While the first human pluripotent stem cells were derived from human embryonic cells in 1998[30], it wasn't until more than a decade later that naïve hPSCs could be derived and maintained in culture[29]. Currently, two primary medium conditions are used: the 2i/L medium supplemented with a protein kinase C (PKC) inhibitor and the 5i/L/A medium. The 2i/L+PKCi medium includes two small molecule inhibitors (2i) that suppress the MEK and GSK3 pathways, which are critical for maintaining pluripotency, along with the cytokine Leukaemia Inhibitory Factor (LIF), which supports pluripotency maintenance, and the PKC inhibitor, which encourages self-renewal and inhibits differentiation[21]. On the other hand, the 5i/L/A medium is a modification of the 2i/LIF medium and features five small molecule inhibitors that obstruct different signalling pathways. These include a combination of five kinase inhibitors targeting MEK, GSK3, BRAF, SRC, and ROCK, as well as LIF and activin A[15].

The naïve hPSCs used in this study were cultured in similar conditions to the 2i/L+PKCi medium. The naïve medium was composed of a basal medium called N2B27, supplemented with four components: PD0325901, a MEK inhibitor; human LIF; Gö6983, a PKC inhibitor; and XAV939, which inhibits tankyrase activity and subsequently reduces β-catenin levels, leading to decreased Wnt signalling activity.

1.3.2. Primed hPSCs

Primed hPSCs and post-implantation epiblasts exhibit similar gene expression patterns and cell signalling pathways, including the Wnt, FGF, and Nodal pathways, that have significant roles in maintaining pluripotency and promoting cell differentiation. Primed hPSCs can differentiate into the three cell lineages also derived from the epiblast during embryonic development, including neural, cardiac, and mesodermal cells. These cells are cultured with specific growth factors that promote self-renewal and differentiation into desired lineages. The commonly used commercial media for this purpose include E8 and mTeSRplus. E8 medium contains a simplified formulation of growth factors, such as bFGF and TGF-β, that are essential for maintaining pluripotency. In contrast, mTeSRplus is a modified version of mTeSR1, optimized for maintaining primed hPSCs. Includes additional growth factors, such as Activin A, and is supplemented with a small molecule inhibitor of the ROCK signalling pathway. In this study, primed hPSCs were cultured using an E8 medium, prepared according to Chen et al. (2011)[31].

1.3.3. Advantages and disadvantages of naïve and primed hPSCs

Primed hPSCs are commonly used in research. They have been extensively characterized and there are established protocols for their culture and differentiation into various cell types. Additionally, naïve hPSCs require specific culture conditions that have only recently been developed[29]. However, while there has been a wide range of research applications, including disease modelling, drug discovery, and regenerative medicine, based on primed hPSCs, there is growing interest in the potential of naïve hPSCs for research and clinical applications.

Recent advances in culture conditions and protocols for deriving and maintaining naïve hPSCs make them increasingly accessible for research. Naïve hPSCs have a broader differentiation potential, which may enable better disease modelling and drug discovery as they can differentiate into a wider range of cell types affected by different diseases. Additionally, naïve hPSCs are less committed to a specific lineage than primed hPSCs, making them more responsive to external signals. As a result, they can be a useful tool for studying how cells respond to different stimuli, such as growth factors, hormones, and drugs.

However, both hPSCs occasionally show epigenetic aberrations–more thoroughly characterized in primed hPSCs. These aberrations include modifications in DNA methylation patterns, parental imprinting, and X chromosome inactivation. They can arise during culture, mainly through promoter hypermethylation, resulting in gene silencing. Additionally, loss of imprinting can occur due to hypomethylation of imprinted DMRs, leading to improper biallelic expression of imprinted genes. There can also be an erosion of X chromosome inactivation, characterized by repression of XIST and partial reactivation of genes from the silent X chromosome. Furthermore, aberrations can also be introduced during hPSC derivation and the transition to naïve pluripotency[32].

Regardless, the differentiation into specific cell lineages requires that naïve pluripotent stem cells transition through a peri- and early post-implantation committed state. Therefore, despite the recent advancements in reprogramming techniques and the development of naïve pluripotent stem cells, primed hPSCs remain crucial in the field. Primed hPSCs are still necessary for deriving new cell types for regenerative medicine, as they are more efficient at generating certain cell types compared to their naïve counterparts[33]. Therefore, it is still required to advance the study of primed hPSCs to enhance our comprehension of human development and our capacity to utilize stem cells for therapeutic applications. However, there is a need for improved medium conditions to derive primed hPSCs and mimic human development accurately[34].

1.3.4. Capacitation

As mentioned above, previous studies showed that naïve pluripotent stem cells cannot differentiate directly into germ cells or somatic lineages. Instead, they must first go through the peri- and early post-implantation phase of epiblast development in a process called 'capacitation'. This process enables naïve cells to acquire the competence required for lineage induction[35–37]. However, the current capacitation methods require prolonged culture outside of developmentally relevant timing to achieve robust multilineage differentiation[20]. Recently, Rostovskaya et al. (2019) developed a new approach that uses a capacitation medium containing a Wnt inhibitor to induce priming in naïve hPSCs. This new approach aimed to recapitulate *in-utero* progression to the late epiblast stage, where cells become fully competent for germ layer induction at appropriate developmental timing. Additionally, they wanted to improve viability, reduce cellular heterogeneity, and have a higher efficiency of multilineage differentiation without the need for exogenous growth factor stimulation. According to their study, after ten days in the capacitation medium, the initially naïve hPSCs appeared to be fully primed and able to differentiate into neuroectoderm, endoderm, and mesoderm lineage precursors, which can undergo further differentiation into tissue progenitors and post-mitotic cell types[10].

The inhibition of the Wnt signalling pathway played a significant role in the success of this capacitation medium. The Wnt signalling pathway is essential in regulating both the pluripotency and differentiation of stem cells and is active in naïve hPSCs. However, during the priming process, Wnt signalling is downregulated, which is necessary for the cells to become competent and responsive to differentiation cues[38,39]. During embryonic development, the Wnt pathway is involved in a variety of processes, including the formation of the neural tube, the establishment of the body axis[40], and the development of various organs, such as the heart and kidneys. Dysregulation of this pathway has been implicated in some developmental disorders, including neural tube defects, limb malformations, and certain forms of cancer[41,42].

The cells in this dissertation were derived from the same protocol used in the Rostovskaya et al. study[10]. To achieve capacitation, the naïve cells were cultured in an N2B27 basal medium supplemented with a Wnt inhibitor (XAV939, IWP2, or WNT-C59). This medium had a similar composition to the naïve medium but lacked a MEK inhibitor (PD0325901), human LIF, and a PKC inhibitor (Gö6983). After ten days in the capacitation medium, the cells exhibited characteristics of primed hPSCs. To maintain the cells in the primed state, they were cultured in either E8 or N2B27 supplemented with XAV939, activin A, and FGF2. That medium is called the 'XAF medium' in this dissertation.

1.4. Epigenetic characteristics of human embryonic development

1.4.1. Epigenetics

Epigenetics refers to the study of changes in gene expression or cellular phenotype that can be inherited by cells or organisms without changing the underlying DNA sequence[43]. Various factors, including environmental exposures, lifestyle choices, and ageing, can influence epigenetic modifications[44,45], and they help explain how different cells can have distinct functions despite having the same DNA.

DNA methylation is one of the most common epigenetic modifications. It consists of a methyl group added to the five-carbon of the DNA cytosine nucleotide. This modification is catalysed by the DNA methyltransferases (DNMT) family of proteins and, in mammals, typically occurs in a 5'–C–phosphate–G–3' sites (CpG) context. DNA methylation can affect gene expression by regulating the accessibility of the DNA to the transcription machinery. In somatic tissues, DNA methylation is present in approximately 70-80% of CpGs in all tissues, with some variation between tissues[46]. In addition to methylation, cytosine can be oxidized by the ten-eleven translocation (TET) family of proteins, resulting in hydroxymethyl, formyl, and carboxyl cytosine[47]. These oxidation forms of cytosine are less common in somatic cells. For example, the percentage of 5-hydroxymethylcytosine (5hmC) in CpG sites of human cells can vary from 0.3-5% depending on the tissue type and developmental stage of the organism[48]. However, although less common, they are believed to play a role in gene regulation[49]. It's worth noting that while epigenetic changes are heritable, they can also be reversible. It is possible to modify gene expression by altering the chromatin structure through environmental factors such as diet or exposure to toxins. Thus, the study of epigenetics has significant implications for human health, as it provides insight into how environmental factors can affect gene expression and contribute to diseases such as cancer.

Histone modifications are diverse and serve specific purposes in controlling gene expression and chromatin structure. They entail the post-translational modification of histone tails through methylation, acetylation, ubiquitylation, and phosphorylation, among others. However, for this study, we only analysed five histone modifications: H3K4me3, H3K4me1, H3K27ac, H3K27me3, and H3K9me3. H3K4me3 is associated with active gene promoters and plays a critical role in activating gene expression, H3K4me1 with enhancer regions, H3K27ac with active promoters and enhancers and is often found in actively transcribed regions, H3K27me3 with polycomb repression in different genomic contexts, and commonly in repressed and bivalent promoters, and H3K9me3 with heterochromatin and repetitive regions[50]. We also included a chromatin accessibility assay that, although not an epigenetic modification, provides information on how transcription factors interact with DNA. Finally, epigenetics also includes non-coding RNAs, such as microRNAs and long non-coding RNAs (lncRNAs)[51,52], although those were not studied in this dissertation.

Cells can maintain their epigenetic information across generations through passive and active mechanisms. Passive mechanisms include the transmission of epigenetic marks from parental cells to daughter cells during cell division. During DNA replication, parental histones and associated epigenetic marks are transferred to the newly synthesized daughter strands in a process known as histone inheritance[53]. This ensures that epigenetic modifications are faithfully transmitted to progeny cells, helping to maintain stable gene expression patterns. Active mechanisms include the action of epigenetic enzymes and associated proteins that can recognize and maintain specific epigenetic marks. For example, DNA methyltransferases can retain DNA methylation patterns by copying them onto newly synthesized DNA strands. The combination of passive and active mechanisms allows cells to faithfully transmit and preserve their epigenetic information across cell generations, ensuring the proper regulation of gene expression and cellular function[54].

1.4.2. Histone modifications in mammalian embryo development

Histone modifications are associated with gene expression regulation during human embryonic development. These modifications, which include methylation, acetylation, phosphorylation, and ubiquitination, can affect the accessibility of DNA to regulatory proteins, which ultimately impacts gene expression[55]. However, due to the complex and dynamic nature of human development, along with ethical and practical concerns, there is still much to learn about the dynamics of histone modifications in this process. In contrast, mouse development is a well-characterized and widely used model organism in developmental biology research. As such, there is a better understanding of histone modification dynamics in this species. Additionally, mouse development is easier to manipulate and observe in a laboratory setting, which allows researchers to conduct experiments that may be challenging or impossible to carry out in human development.

In early mouse embryonic development, H3K4me3 quickly decreases in the paternal genome after fertilization but is restored during zygotic gene activation (ZGA). Conversely, the maternal genome contains a non-canonical form of H3K4me3 that spans broad domains in promoters and distal regions. During this stage, the level of H3K4me3 marks in promoter regions is highly dynamic and positively correlated with gene expression levels. Furthermore, broad H3K4me3 domains are common and may help sustain the transcription of essential cell-type-specific factors in a stable state[55].

The histone modification H3K27me3, associated with gene repression, quickly decreases in the maternal and paternal alleles during the early stages of mouse embryo development. This decrease is achieved through the global erasure of H3K27me3 in the paternal genome and the depletion of promoter H3K27me3 in the maternal genome. In human pre-implantation embryos, Xia et al. (2019) found that H3K27me3 deposition in human oocytes occurs in the promoters of developmental genes and partially methylated domains, which differs from mouse oocyte patterns. Moreover, the resetting of H3K27me3 during human pre-implantation embryo development is different from that in mice. Specifically, at the ZGA stage of human embryo development (8-cell stage), there is a negligible H3K27me3 signal, suggesting that H3K27me3 is globally erased on both parental genomes. The loss of H3K27me3 in human embryos may be linked to the absence of the core components of polycomb repression complex 2 (PRC2)[56–60].

H3K9me3 histone modification is associated with heterochromatin, repetitive elements, and specific protein-coding genes in pre-implantation mouse embryos. After fertilization, there is significant DNA demethylation and a transition from DNA methylation to other types of epigenetic modifications, including repressive histone modifications, is necessary to control the transcription of long terminal repeats (LTRs). In the post-implantation stage, the accurate regulation of H3K9me3 deposition at lineage-specific genes is critical for proper mammalian embryo development[55,58,61].

1.4.3. DNA methylation in early human development

During human pre-implantation development, the genome of the fertilized egg undergoes demethylation followed by *de novo* methylation. Demethylation occurs shortly after fertilization and continues until the 4- to 8-cell stage. This demethylation event results in the erasure of most of the methylation marks inherited from the gametes, resetting the epigenome to a pluripotent state. After implantation, the *de novo* methylation process begins, resulting in the establishment of tissue-specific methylation patterns. The *de novo* methylation is initiated by DNA methyltransferases DNMT3A and DNMT3B. This process is critical for establishing the cell identity and differentiation potential. At this stage, the methylation patterns become more stable and less dynamic and start to have regulatory regions with lineage-specific methylation[62]. The embryonic cells differentiate into the three germ layers–ectoderm, mesoderm, and endoderm–and acquire specific methylation patterns necessary for their function. Dysregulation of these processes can lead to developmental disorders and diseases[63].

Some genes, known as imprinted genes, are shielded from demethylation during early development. Imprinted genes are a specific set of genes whose expression differs based on their parent of origin and are, therefore, identified by their differential allele DNA methylation patterns. These patterns are established during gametogenesis and maintained throughout development in imprinting control regions (ICRs). These regions maintain this distinct DNA methylation pattern by being safeguarded from the active demethylation in the zygote, ensuring that the differential methylation pattern is accurately preserved throughout the developmental process[64].

1.4.4. Promoter bivalency in early human development

Promoter bivalency refers to the epigenetic state of a gene promoter that carries both activating and repressive histone modifications simultaneously. It refers to the coexistence of the histone modifications H3K4me3 and H3K27me3 on gene promoters. H3K4me3 is associated with active gene transcription, while H3K27me3 is associated with gene silencing. This epigenetic state as a role in expression regulation during early human development and in human pluripotent stem cells[65]. This is particularly important for genes involved in cell fate determination and differentiation[66]. Promoter bivalency appears necessary for the timely activation of genes during development, and dysregulation of promoter bivalency may result in developmental disorders and diseases[67–69]. Bernstein et al. (2006) showed that promoter bivalency is a hallmark of genes poised for expression in mouse embryonic stem cells and that bivalency is essential for their timely activation during development. This study also suggested that promoter bivalency is established by the coordinated action of polycomb and trithorax group proteins, which are known to play essential roles in gene regulation during development[70]. Another study by Mikkelsen et al. (2007) also showed that promoter bivalency is present in mouse embryonic stem cells and is associated with genes involved in embryonic development and differentiation. The study further demonstrated that promoter bivalency is dynamic and is resolved to monovalency during lineage commitment, allowing for the precise regulation of gene expression during development[71]. In human pluripotent stem cells, Singh et al. (2015) demonstrated that bivalent domains are unstable, dynamic, and regulated during the cell cycle. These domains are established to regulate developmental gene activation in a cell-cycle-dependent manner, ensuring that differentiation begins in the G1 phase.

1.4.5. Gene imprinting

Gene imprinting is an epigenetic phenomenon where specific genes are expressed only from the maternal or paternal chromosome. This process is essential for normal embryonic development and is regulated by epigenetic marks, such as DNA methylation, on the chromatin region that controls gene expression[72,73]. The human genome contains a small number of imprinted genes, such as the H19/IGF2 locus on the human chromosome band 11p15.5, the SNRPN locus on chromosome band 15q11-q13, and the KCNQ1OT1 locus on chromosome band 11p15.5. Each of these regions controls a cluster of genes that are expressed only from one of the two parental chromosomes. For example, in the H19/IGF2 locus, the H19 gene is expressed only from the maternal chromosome, while the IGF2 gene is expressed only from the paternal chromosome. An imbalance in the gene expression of these imprinted genes can have significant consequences for growth and development. It is thought that it contributes to diseases such as Beckwith-Wiedemann syndrome and Silver-Russell syndrome[74].

The precise mechanisms that regulate gene imprinting are still not fully understood, but the process may be sensitive to environmental factors such as diet and stress[75]. In addition, mutations in imprinted genes or ICRs can lead to developmental disorders such as Angelman syndrome and Prader-Willi syndrome[76]. Prader-Willi syndrome (PWS) occurs when the paternal copy of imprinting control regions on chromosome 15 is deleted or silenced. This results in developmental abnormalities, cognitive impairment, and hyperphagia.

1.5. X chromosome inactivation (XCI)

1.5.1. XCI in early mammalian development

In female mammals, including humans, one of the two X chromosomes in each cell is randomly inactivated during early development. This process is called X chromosome inactivation (XCI) and ensures that females do not have twice as much X chromosome gene expression as males, who only have one X chromosome. X chromosome inactivation occurs in all somatic cells of the female embryo and, once inactivated, the same X chromosome remains inactive in all subsequent cell divisions[77].

XCI is initiated by the expression of the XIST gene on one of the two X chromosomes. XIST is a long non-coding RNA (lncRNA) that spreads along the chromosome, coating it and silencing its gene expression. The inactivation of the X chromosome is associated with several other factors, including histone modifications, DNA methylation, and the formation of a specific chromatin structure called the Barr body[78]. XCI is a crucial mechanism for normal development and function in female mammals[79,80]. However, it can also lead to diseases when it is not properly regulated. For example, mutations or abnormalities in the XIST gene or other factors involved in X chromosome inactivation can cause X-linked disorders that affect females[81].

X-active coating transcript (XACT) is a primate-specific lncRNA encoded on the X-chromosome, which plays a role in X-linked gene dosage regulation during early human development. Although it was first discovered in human pluripotent stem cells, its exact mechanism of action and specific role in human embryogenesis are still unknown[82]. XACT is transcribed from the active X chromosome and coats it, thereby preventing its inactivation. It recruits a complex of proteins to the active X chromosome to maintain its active state. Its expression is exclusive to the active X chromosome, functioning as a safeguard from inactivation during XCI. Additionally, it has been suggested that XACT may play a role in maintaining the pluripotent state[83].

The inactive X chromosome (Xi) in humans has epigenetic characteristics that distinguish it from the active X chromosome (Xa). The most prominent epigenetic modification associated with Xi is the histone modification H3K27me3. This histone modification is catalysed by the Polycomb Repressive Complex 2 (PRC2) and is essential for gene silencing on Xi[84]. Another epigenetic hallmark of Xi is DNA methylation. In general, the CpG islands (regions of DNA with high CpG content) on Xi are more heavily methylated than those on Xa. However, some genes on Xi escape XCI and remain active, and their promoters are usually associated with lower levels of DNA methylation[85]. Finally, the expression of the XIST lncRNA, transcribed exclusively from Xi, also influences the epigenetic characteristics of the inactive X chromosome. XIST RNA coats the entire length of Xi, recruiting epigenetic modifiers that lead to the accumulation of repressive histone modifications and DNA methylation.

1.5.2. XCI in human pluripotent stem cells

In female preimplantation mouse embryos, the paternal X chromosome is imprinted and inactivated in all cells. However, at the peri-implantation stage, pluripotent epiblast progenitor cells reactivate the inactivated paternal X chromosome. Cultured mouse embryonic stem cells (mESCs) represent this temporary population of pluripotent cells possessing two active X chromosomes. During differentiation, mouse embryonic epiblast cells and mESCs inactivate the maternal or the paternal X chromosome in individual cells[86].

Unlike in mice, human female preimplantation embryos do not experience imprinted inactivation of the paternal X chromosome. Instead, both X chromosomes in female human preimplantation embryos initiate some degree of silencing. The X chromosome-associated long noncoding XIST RNA, which is necessary for stable X-inactivation in mice and is a characteristic of the inactive X chromosome, is expressed from and covers both X chromosomes in most cells of female human blastocyst-stage embryos in cis. XIST RNA coating then attracts a range of proteins to the X chromosome resulting in gene expression silencing.

Although X-inactivation during human embryonic development is a critical process, our understanding of it remains limited. Research has mainly relied on the use of human pluripotent stem cells and comparison with mouse pluripotent stem cells and embryos. It is essential to note that the diverse patterns of X-inactivation observed in mouse and human pluripotent stem cells may be due to variations in derivation methods and culture conditions. Therefore, further research is necessary to deepen our knowledge of this biological process in humans[80]. In naïve hPSCs, some cells have XIST RNA coating on both X chromosomes but do not seem to deactivate the XIST RNA-covered X chromosomes transcriptionally, as in the cells of early female human embryos. When these naïve hPSCs differentiate into primed hPSCs, most cells undergo random X-inactivation and display XIST RNA coating on a single X chromosome that is transcriptionally inactive. Primed hPSCs are similar to mouse epiblast stem cells (mEpiSCs), although with some differences. While female mEpiSCs have one inactivated XIST RNA-coated X chromosome, primed female hPSCs may exhibit three X-inactivation patterns: no inactive X chromosome, one inactive X, or an eroded inactive X. Over time, many primed female hPSC lines lose XIST RNA coating, and some previously silenced genes from the inactive X chromosome become expressed. This phenomenon is referred to as X-inactivation erosion and is also a feature of human induced pluripotent stem cells (hiPSCs). The instability of the epigenetically inactivated X chromosome in hiPSCs and primed female hPSCs raises concerns about using pluripotent female human cells in regenerative medicine and disease modelling because increased expression of X-linked genes due to X-inactivation erosion or failure can have harmful effects on development and differentiation[87,88].

1.5.3. XCI erosion

Human pluripotent stem cells exhibit X-chromosome inactivation erosion where the inactive X chromosome becomes reactivated in some cells, leading to a loss of XCI and an imbalanced gene expression. This process can compromise the stability and functionality of hPSCs and affect their downstream applications. The eroded X chromosome is characterized by the loss of repressive polycomb-associated histone modifications, specifically H3K27me3 and XIST RNA expression, resulting in the reactivation of genes on the inactive X chromosome, leading to improper gene dosage compensation that affects cell fate decisions during differentiation[89].

Vallot et al. (2015) conducted a study on X chromosome inactivation (XCI) in hPSCs and showed that the erosion of XCI is not a widespread process throughout the chromosome. Instead, it is a targeted phenomenon that occurs in regions with high levels of H3K27me3, leading to changes in chromatin and transcription. The authors used a single-cell RNA fluorescence in situ hybridization (FISH) method to investigate XCI instability and found a correlation between the erosion of XCI and pluripotency. They also found that loss of XIST expression is not responsible for X inactivation instability and that gene reactivation from Xi occurs before the loss of XIST coating. In particular, the authors highlighted the importance of XACT expression and coating of the X chromosome in XCI erosion, suggesting a potential role for XACT in the epigenetic instability of hPSCs. Additionally, the study revealed that heterochromatin remodelling and gene reactivation occur in specific H3K27me3-enriched domains, while H3K9me3-marked regions remain unaffected[9].

1.6. Sequencing methods

1.6.1. RNA sequencing (RNA-seq)

RNA sequencing (RNA-seq) is a widely used method for studying the transcriptome, which refers to the complete set of RNA molecules in a cell, tissue, or organism. This method enables quantification of gene expression and the detection of alternative splicing events, RNA editing, and other post-transcriptional modifications. The RNA-seq protocol involves several steps: RNA extraction and purification, which entails extracting total RNA from the sample of interest and eliminating any remaining DNA; RNA fragmentation and cDNA synthesis, in which the RNA is fragmented and complementary DNA (cDNA) is synthesized using reverse transcriptase and random primers; and library preparation and sequencing, during which the cDNA fragments are purified, sequencing adapters are added, and the resulting library is sequenced using high-throughput sequencing technology. The resulting reads are then aligned to the reference genome and quantified based on the number of reads that map to each gene[90].

1.6.2. Chromatin immunoprecipitation sequencing (ChIP-seq)

Chromatin immunoprecipitation sequencing (ChIP-seq) is a method used to investigate protein-DNA interactions. Specifically, it helps to identify genomic regions where a particular protein binds to DNA. The ChIP-seq process includes four primary steps: crosslinking, chromatin fragmentation, immunoprecipitation, and sequencing. In the first step, a crosslinking agent is used to preserve the protein-DNA interactions in the cells. Afterwards, the DNA-protein complexes are sonicated or enzymatically digested to generate DNA fragments of the desired size, typically between 100-500 bp. The third step is immunoprecipitation, which involves using a specific antibody to pull down the protein-DNA complexes. The antibody is bound to a solid support, such as protein A/G beads, and the DNA-protein complex is selectively captured on the beads. In the final step, the DNA is released from the protein by reversing the crosslinking and DNA fragments are purified and sequenced using high-throughput sequencing technology. The resulting sequencing reads are then mapped to the reference genome, and the signal peaks are used to identify genomic regions where the protein of interest is bound. These peaks can be analysed to understand the functional roles of the protein in gene expression, chromatin structure, and other cellular processes[91].

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

Assay for Transposase Accessible Chromatin with High-Throughput Sequencing (ATAC-seq) is a method for profiling chromatin accessibility on a genome-wide scale. This method allows the identification of accessible regions to DNA-binding proteins, such as transcription factors. The ATAC-seq protocol has the following steps: first, intact cells are lysed and the chromatin is released. The transposase enzyme is then added to the chromatin solution, which cuts and inserts sequencing adapters into accessible chromatin regions while leaving inaccessible regions intact. Then, the chromatin-DNA fragments are purified, and the sequencing adapters are amplified by PCR, generating a library of DNA fragments. Lastly, the resulting DNA library is sequenced using high-throughput sequencing technology. The sequencing reads are aligned to a reference genome, and the signal peaks are analysed to identify genomic regions accessible to the transposase enzyme and other DNA-binding proteins, providing insights into the mechanisms of gene regulation[92].

1.6.4. Post-bisulfite adaptor tagging (PBAT)

PBAT, Post-Bisulfite Adaptor Tagging, is a method used for genome-wide analysis of DNA methylation at single-base resolution. The PBAT method combines bisulfite treatment of DNA with high-throughput sequencing to identify the location and frequency of cytosine methylation in the genome. Bisulfite treatment is a chemical process that converts unmethylated cytosine residues to uracil, leaving methylated cytosine residues unchanged. The method involves the following steps: bisulfite treatment, in which the DNA is treated with sodium bisulfite to convert unmethylated cytosines to uracil, while methylated cytosines are protected from bisulfite-induced conversion and remain unchanged; adaptor tagging, in which the adaptor tags are added to the ends of the bisulfite-converted DNA fragments and are used to amplify the fragments by PCR; and sequencing, in which the resulting DNA library is sequenced using high-throughput sequencing technology. Finally, the sequencing reads are aligned to a reference genome, and the locations of cytosine methylation are determined based on the presence or absence of cytosine residues at specific positions. The PBAT method has been used to study DNA methylation patterns in various cell types and tissues and to investigate changes in DNA methylation in response to environmental factors, ageing, and disease[93].

1.7. Bioinformatic analysis

1.7.1. Multi-omics factor analysis

The Multi-Omics Factor Analysis (MOFA) is a bioinformatics package for analysing multi-omics datasets. It is a multivariate dimension reduction technique that identifies patterns of variation shared across multiple datasets and identifies common factors that underlie the variation in the data. The MOFA package provides several functions for data pre-processing, factor analysis, and visualization. It can perform unsupervised factor analysis, where the underlying factors are not known a priori, or supervised factor analysis, where the analysis is guided by a specific variable of interest. Importantly, it can be used to identify the most significant factors and explore the relationships between them.

The package combines two main statistical techniques: Probabilistic Matrix Factorization (PMF). MOFA uses a PMF approach to decompose each dataset into a lower-dimensional representation of unique factors. PMF is a statistical technique that decomposes a matrix into a product of two lower-dimensional matrices, one of which represents latent factors that explain the variability in the data; MOFA also uses a Bayesian approach to integrate the lower-dimensional representations across the different datasets by identifying the shared patterns of variation across all datasets. Bayesian Integration is a statistical technique that combines information from multiple sources to make probabilistic inferences.

Together, these techniques enable MOFA to identify patterns of variation shared across multiple datasets, even when the datasets have different scales, dimensions, and noise levels. The package also incorporates several other statistical techniques to account for missing data, handle overfitting, and perform dimensionality reduction. These include imputation of missing values, regularization, and selection of the optimal number of factors[94–96].

1.7.2. Chromatin-state discovery with ChromHMM

ChromHMM is an R package used to identify the chromatin state based on genome-wide profiling methods such as ChIP-seq, DNase-seq, and ATAC-seq. The Chromatin state refers to the functional organization of DNA in the nucleus, which is reflected in the different patterns of histone modifications, DNA methylation, and chromatin accessibility across the genome. The algorithm models the chromatin state as a sequence of discrete states, each defined by a combination of chromatin marks. The model is trained using an unsupervised learning approach, where it automatically learns the chromatin states and their associated chromatin marks based on the input data. Once the model is trained, ChromHMM can be used to annotate the genome with the predicted chromatin states and visualize the chromatin state patterns in different genomic regions. ChromHMM also provides tools for comparing chromatin state patterns across different cell types or conditions and identifying genomic regions that are differentially regulated between them[97].

ChromHMM uses Hidden Markov Models (HMMs). HMMs are used to model the underlying chromatin state transitions and to infer the most likely sequence of states given the observed data. It also uses Multinomial logistic regression to model the association between chromatin marks and chromatin states, Bayesian Information Criterion (BIC) to select the optimal number of chromatin states to model, and random initialization of parameters to avoid local optima and to increase the likelihood of finding the global optimum. Finally, it uses a bootstrap analysis to estimate the stability and reliability of the identified chromatin states.

1.8. Aims of the dissertation

The goal of this dissertation is to provide a comprehensive understanding of the epigenetic, gene expression, and chromatin accessibility differences that exist between naïve and primed pluripotent states in human pluripotent stem cells. By characterizing these differences, the aim is to shed light on the mechanisms that drive the transition from naïve to primed pluripotency in human cells and to demonstrate the role epigenetic modifications play in this process. Additionally, the dissertation aims to identify molecular differences between capacitated and conventionally cultured hPSCs, intending to help the research community decide on the pros and cons of using each primed hPSCs. Finally, it aims to identify genomic features, genes, and transcription factors that contribute to establishing different pluripotency states, potentially serving as targets for future research. The findings of this research will hopefully help advance our understanding of human embryonic development and aid the development of novel regenerative medicine approaches. Understanding the epigenetic changes occurring during pluripotency transitions will potentially lead to more efficient and effective methods for differentiating stem cells into specific cell types for use in regenerative medicine.

2. Results

2.1. Experimental setup

To study epigenetic dynamics during early human embryonic development, we performed global chromatin profiling of human pluripotent stem cells (hPSCs) in the naïve state, after capacitation, and of conventionally cultured primed hPSCs. Naïve pluripotent stem cells can be derived either directly from embryos[98] or by resetting primed conventionally cultured hPSCs[20,21,99]. Here we included both types: HNES1 hPSCs, a male cell line derived in the naïve state directly from human embryo inner cell mass (ICM)[98], and cR-H9 hPSCs, a female hPSC line derived by chemical resetting of primed H9 hPSCs[20]. These cells were cultured for ten days in a capacitation medium ('XAV medium') and then split into two culture conditions suitable for long-term maintenance: 'XAF medium' containing tankyrase inhibitor XAV939, Activin A and FGF2; and E8 medium. Using these cells, we profiled histone modifications with chromatin immunoprecipitation sequencing (ChIP-seq: H3K4me3, H3K4me1, H3K27me3, H3K27ac, and H3K9me3), chromatin accessibility with an assay for transposase-accessible chromatin using sequencing (ATAC-seq), DNA methylation with post-bisulfite adaptor tagging (PBAT), and we included in the analysis prior published transcription dataset (GSE123055; Rostovskaya et al., 2019[10]). We sequenced both HNES1 and cR-H9 hPSC lines in naïve conditions ('d0'), after ten days of capacitation ('d10'), after prolonged culture (for at least ten additional days after capacitation; only cR-H9 cells were sequenced in this condition) in the two alternative conditions suitable for long-term maintenance ('d20X': XAF medium containing tankyrase inhibitor XAV939, Activin A and FGF2; 'd20E': E8 medium), and from H9 hPSCs in standard culture conditions ('H9 Primed': E8 medium) before resetting (Figure 1). Table 1 lists the number of biological replicates for each assay. After having gathered the epigenetic, accessibility and expression data, I proceeded to characterize and model the results.

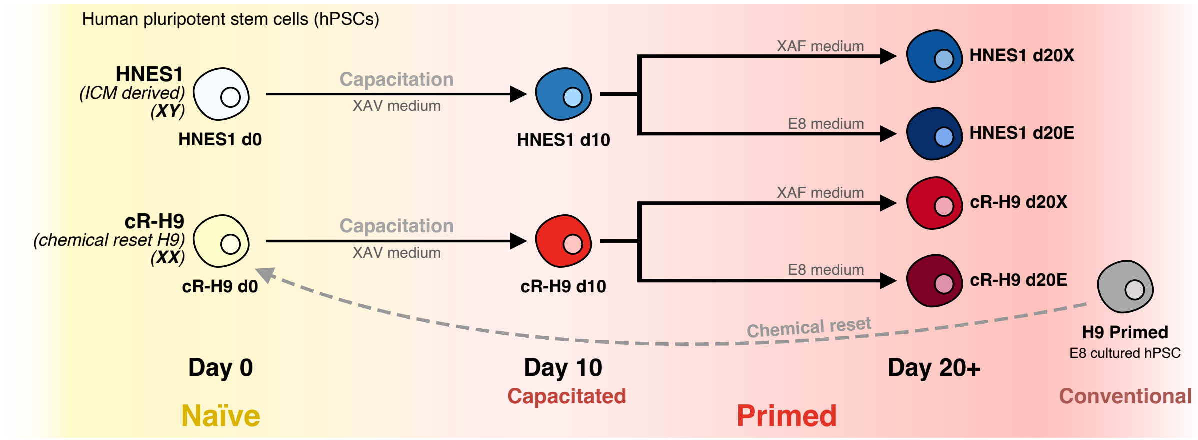

Figure 1. Diagram of the experimental design.

The cells in shades of blue represent the HNES1 cells derived directly from human embryo ICM, while the cells in shades of red represent the cR-H9 cells derived by chemical resetting of primed H9 hPSCs. The conventionally cultured H9 primed cells are represented in grey. The sample names indicate the number of days the cells were kept in culture: day 0 (‘d0’), day 10 (‘d10’), day 20 in XAF medium (‘d20X’) and day 20 in E8 medium (‘d20E’). The colour scheme and sample naming are maintained throughout all figures.

2.2. Multi-omics landscape of naïve and primed human pluripotent stem cells (hPSCs)

2.2.1. Sample clustering

I began the analysis by assessing the similarity between samples. To accomplish this, I performed a principal component analysis (PCA) and hierarchical clustering using differentially expressed genes, differential ATAC-seq and ChIP-seq peaks between all conditions, and the top 5% most variable 200 CpG-containing genome-wide windows based on their methylation average. The clustering showed that each epigenetic modification separated naïve cells (day 0) from primed cells (capacitated and long-term primed cells), highlighting that epigenetic remodelling during the naïve-to-primed transition affects all epigenetic modifications, irrespective of their activating or repressive role (Figures 2A and S1A-B). I obtained a similar result when I integrated the same data subset into a single dimensionality reduction method, specifically t-SNE. The samples clustered separately into the naïve and primed state of pluripotency, although day 10 (capacitated) cells clustered apart from the long-term primed cells (Figure 2B). In the PCA of enhancer-associated histone modification, H3K27ac and H3K4me1, the day 10 samples clustered together with long-term primed cells, while in the PCA of the remaining epigenetic modifications, gene expression, and accessibility there was a progression from day 10 to long-term capacitated (day 20+) and conventionally cultured hPSCs (H9 Primed). This suggests that enhancers may be remodelled during capacitation and remain stable once the primed state is established, even after long-term culture (Figure 2A).

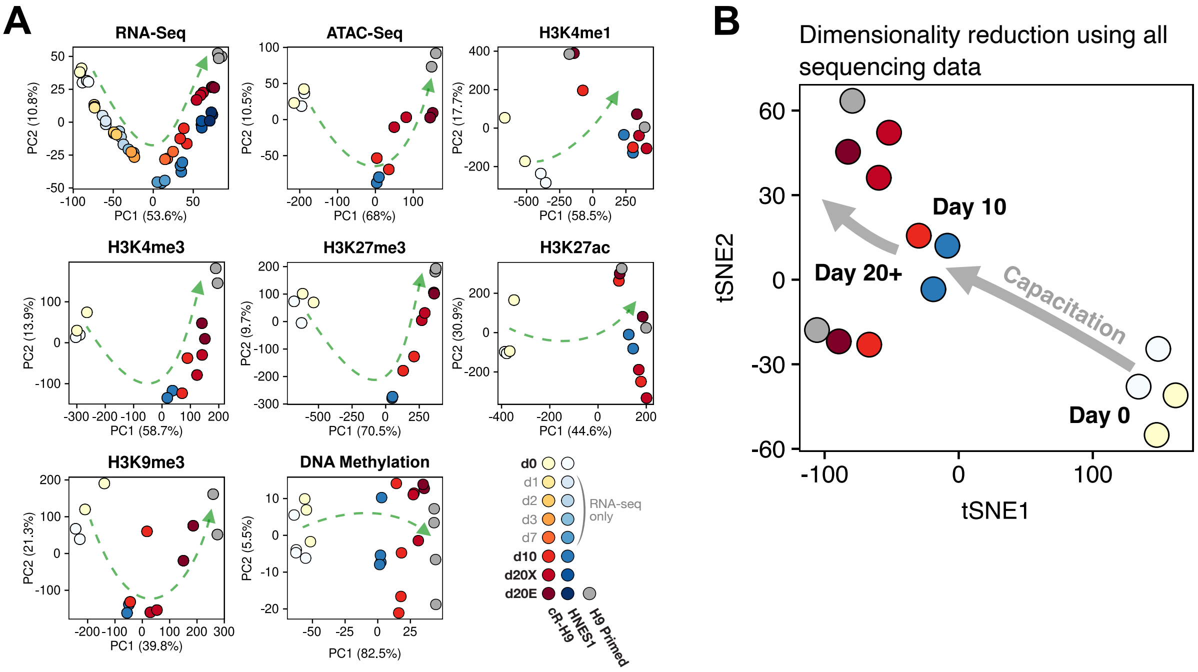

Figure 2. Principal component analysis for each sequencing assay and dimensionality reduction using all sequencing data.

A) Principal component analysis of the top 500 most variable differentially expressed genes, histone modifications ChIP-seq differential peaks, ATAC-seq differential peaks, and DNA methylation average in 200 CpG-containing genomic windows between all conditions. **B)** t-SNE plot generated with the R package MOFA2[95]. The plot was generated with the same data input as the principal component analysis. The PCA of the RNA-seq data includes the samples from day 1, 2, 3, and 7 since those were available in the prior published transcription dataset (GSE123055; Rostovskaya et al., 2019[10]).

2.2.2. Genome-wide and feature-specific epigenetic modification levels

To analyse the epigenetic landscape associated with each condition, I classified the ChIP-seq and ATAC-seq peaks based on their genomic features. The results indicated that the histone modification H3K27me3 peaks had a predominant overlap with CGI (CpG island)-containing promoters and non-promoter CpG islands in the primed state (Figure 3A). Concurrently, the number of H3K27me3 peaks and western blot results showed a significant decline in the H3K27me3 protein levels during the naïve-to-primed transition, suggesting that H3K27me3 levels overlapping CpG islands may increase or remain stable during the pluripotency transition while decreasing noticeably outside of CpG islands. In addition, epigenetic modifications associated with enhancers (H3K4me1 and H3K27ac) and active promoters (H3K4me3) had higher protein and peak levels in the primed state, with H3K4me3 exhibiting a slight enrichment increase in CpG islands and CGI promoters during capacitation (Figures 3A-B and S1C-F). DNA methylation levels also globally increased during the naïve-to-primed transition but remained low in CpG islands (Figures 3C-D, S1G-H). Interestingly, TET2, a gene encoding for a methylcytosine dioxygenase that catalyses the conversion of methylcytosine to

5-hydroxymethylcytosine, showed to be highly expressed in the naïve state, lowly expressed during capacitation, and having only moderate expression in the H9 primed cells, potentially to prevent demethylation during capacitation. Finally, DNMT3L, a gene encoding a protein that stimulates *de novo* methylation by DNA cytosine DNMT3A, was highly expressed until day 10 in cell culture (when the cells reached the capacitated state), however, in the day 10 and long-term cultured cells, its expression was significantly reduced (Figure S2), possibly to increase the rate of *de novo* methylation during the transition.

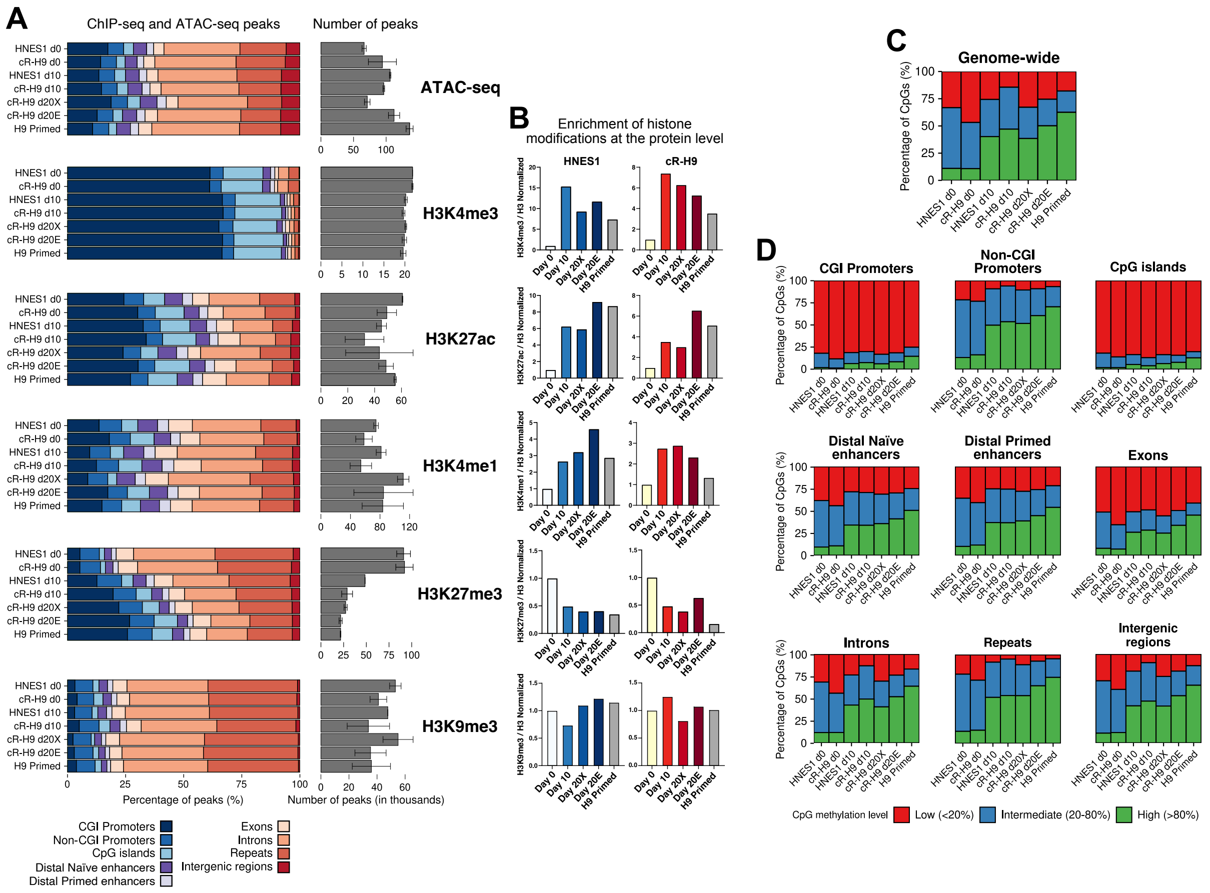

Figure 3. Genome-wide and feature-specific epigenetic modification levels.

A) Percentage of ChIP-seq (H3K4me3, H3K27me3, H3K4me1, H3K27me3, and H3K9me3 marks) and ATAC-seq peaks overlapping annotated genomic regions and the total number of MACS2 peaks for each condition. **B)** Western blot normalized results for each histone modification and condition. **C)** Genome-wide DNA methylation percentage distribution calculated at a CpG resolution. **D)** DNA methylation percentage distribution over annotated genomic regions at a CpG resolution. The CpG methylation levels were divided into three categories: In red, 'Low', with a methylation percentage below 20%; in blue, 'Intermediate', with a methylation percentage between 20 and 80%; and in green, 'High', with a methylation percentage above 80%.

2.2.3. Imprinting control regions

Genome-wide DNA methylation differences between naïve and primed cells are known to be characteristic of these pluripotency states[15]. However, chemical reset cells commonly show a loss of allele-specific methylation in imprinting control regions (ICRs)[15,100]. Possibly due to the long-term naïve culture conditions that can lead to significant demethylation and erasure of CpG methylation in ICRs. Such a process may not occur during the limited cell divisions in the cleavage embryos. Thus, to confirm whether our reset cells showed a loss of imprinting, I characterized the allele-specific epigenetic levels in ICRs of H9 cells by integrating single nucleotide polymorphism (SNP) data (Figure 4). The results revealed that conventionally cultured H9 primed cells had 13 ICRs with a significant methylation percentage difference between alleles, whereas naïve and capacitated cells had no regions with a significant methylation percentage difference between alleles. Interestingly, cells in long-term culture in E8 medium, i.e., 'cR-H9 d20E' cells, had 5 ICRs with significant methylation percentage differences. However, further experiments are necessary to confirm if this result is reproducible and indicative of imprinting recovery. Additionally, H3K4me3 and H3K9me3 histone modifications showed significant differences between alleles in several ICRs in conventionally cultured H9 cells. H3K4me3 was enriched in the opposite allele from DNA methylation, while H3K9me3 was associated with the allele with higher DNA methylation levels. Recent research on mouse cells has shown that paternal H3K4me3 marks escape inactivation during sperm maturation and are present in embryos from early zygotic stages up to implantation[101]. Our results in conventionally cultured H9 cells may reflect these findings, although more research is necessary to understand the mechanisms responsible for allele-specific histone modification levels in ICRs.

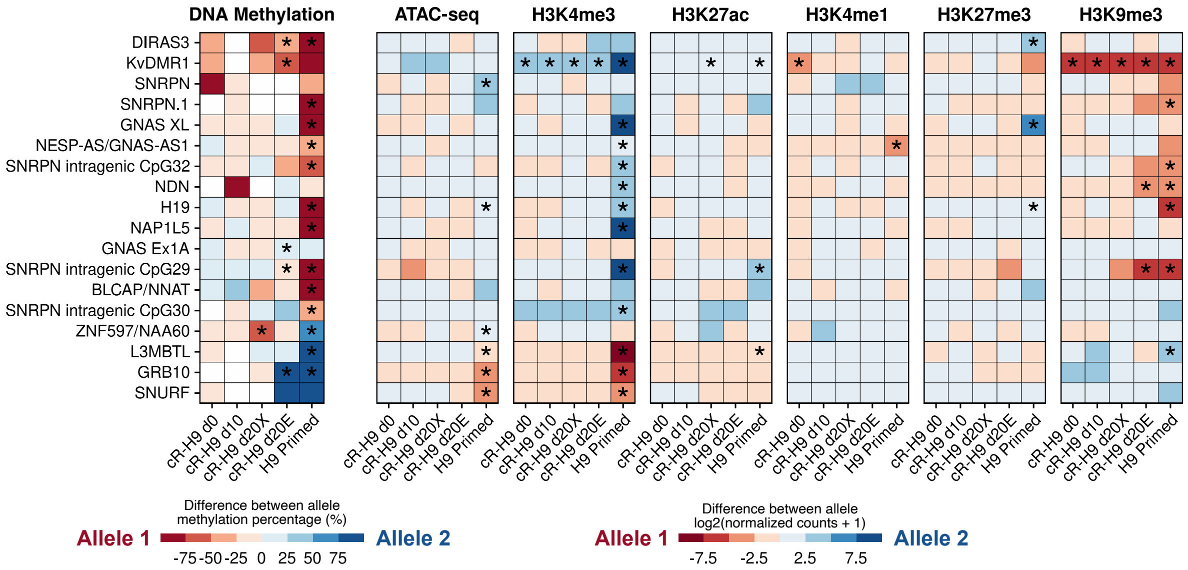

Figure 4. DNA methylation and epigenetic modification allele-specific level differences in imprinting control regions.

Only the imprinting control regions (ICRs) with significant results are shown in the figure. A complete list of ICRs can be found in Table 2. The red-coloured boxes indicate that allele 1 had higher levels than allele 2, whereas the blue-coloured boxes indicate the opposite. The significance threshold was set to * FDR < 0.05, which was calculated using edgeR[102].

2.2.4. Genome-wide epigenetic changes across conditions and association between epigenetic modifications

To gain an overview of how epigenetic modifications and chromatin accessibility change in our experiment, I conducted a genome-wide correlation analysis of 2-kilobase (kb)-sized genomic regions between conditions for each sequencing assay, excluding RNA-seq. I compared the individual correlation between a single condition and the remaining ones, such as the correlation between ‘cR-H9 d0’ cells and all the other H9 cells for each sequencing assay (Figure 5A). I also included the median correlation between all sequencing assays. The results showed that the repressive chromatin-associated epigenetic marks, including H3K27me3, H3K9me3, and methylation, changed the most between the naïve and primed states (Figure 5A; panel ‘cR-H9 d0’). H3K9me3 exhibited a pronounced difference between long-term primed cells cultured in XAF medium and E8 medium, emphasizing how different culture systems can alter the epigenetic landscape (Figure 5A; panel ‘H9 Primed (cR-H9)’).

A correlation analysis between different sequencing assays in each condition, using the same 2 kb-sized genomic regions, showed a higher correlation between H3K4me1 marks and chromatin accessibility in the primed state, suggesting the formation of primed-specific enhancers after capacitation (Figure 5B; panel ‘H3K4me1’), which is in concordance with the principal component analysis (PCA) results (Figure 2A). Concurrently, H3K27me3 became more correlated with enhancer-associated (H3K4me1 and H3K27ac) and promoter-associated histone modifications (H3K4me3 and H3K27ac) in the primed state after capacitation (Figure 5B; panel ‘H3K27me3’), suggesting an expansion of bivalent regions in this pluripotency state. In contrast, DNA methylation became anti-correlated with the remaining epigenetic modifications in long-term cultured cells. This highlights that increased segregation between histone modification-based and methylation-based regulatory roles in conventionally cultured hPSCs could be overstated due to long-term culture conditions.

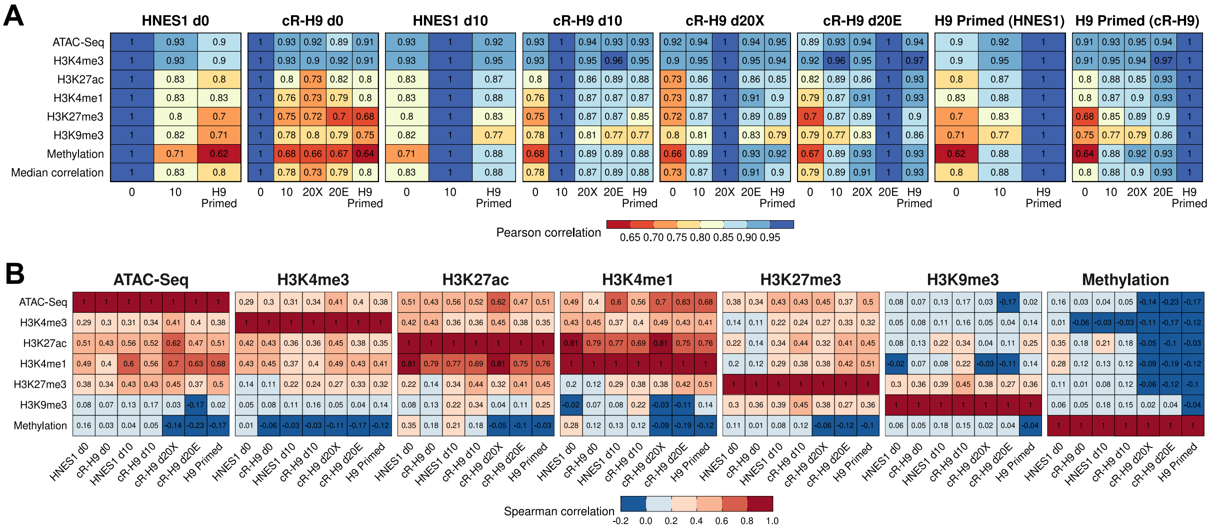

Figure 5. Overview of genome-wide epigenetic changes and interplay between epigenetic modifications across conditions.

A) Pearson correlation analysis performed between one condition and the remaining conditions for each sequencing assay. Heatmaps are shown separately for each condition of comparison. **B)** Spearman correlation analysis performed between sequencing assays for each condition. Heatmaps are shown separately for each sequencing assay of comparison.

2.3. Gene expression and promoter analysis

2.3.1. Characterization of differentially expressed genes between conditions

Aiming to explore the relationship between gene expression levels and epigenetic modifications in promoter regions, I identified five gene clusters based on differential gene expression patterns across conditions. I quantified the levels of histone modifications, chromatin accessibility, and methylation within the promoters of the genes in each cluster (Figure 6A). Gene cluster 1, with 1439 genes, showed high expression in the naïve state with a sharp expression decrease during capacitation and in long-term cultured cells. The promoters of most genes in this cluster had high H3K27ac levels and low levels of the remaining epigenetic modifications in naïve hPSCs. Gene ontology (GO) analysis of these genes identified developmental processes and cell fate commitment as the associated biological processes, suggesting that the early drivers of development may be expressed at the naïve pluripotency stage (Figure 6B). Gene cluster 2, with 543 genes, also showed high expression levels in naïve and capacitated cells but were weakly expressed in long-term capacitated cells and had lower expression in conventionally cultured H9 cells. However, these genes had a different association with promoter epigenetic modifications than cluster 1, showing low levels of active marks in their promoters in all conditions and repression being primarily associated with high H3K9me3 promoter levels. The remaining clusters represented genes that were mostly expressed in the primed state. The genes from these clusters generally had accessible promoters and higher H3K4me3 and H3K27ac levels when expressed. The genes from Cluster 4, with 891 genes, were expressed in all primed states and were associated with GO terms related to axonogenesis and neurodevelopment. Cluster 3 (471 genes) and cluster 5 (1291 genes) segregated genes expressed in capacitated and long-term cultured cells. Genes in cluster 3 were highly expressed in capacitated cells and associated with urogenital system development and extracellular structure organization, suggesting that capacitated cells had active mesoderm-associated genes repressed in long-term cultured cells. Genes in cluster 5 were highly expressed only in long-term cultured after capacitation and conventional H9 cells and were associated with ectoderm lineage terms, such as synapse assembly and organization. These differences in germ layer-associated gene expression between primed hPSCs are consistent with other studies suggesting that long-term cultured PSCs may exhibit a bias for ectoderm expression, potentially affecting the outcome when used in clinical applications[103]. Conversely, capacitated cells exhibit mesoderm and ectoderm gene expression signatures (Figures 6B; cluster 3) despite having multi-lineage differentiation potential, including endoderm[10]. However, the heterogeneity of the cell population studied may explain the mesoderm lineage bias observed in capacitated cells. Since the cells were sequenced in bulk, if a subset of cells started expressing lineage markers, those genes would show an expression increase compared to other conditions.

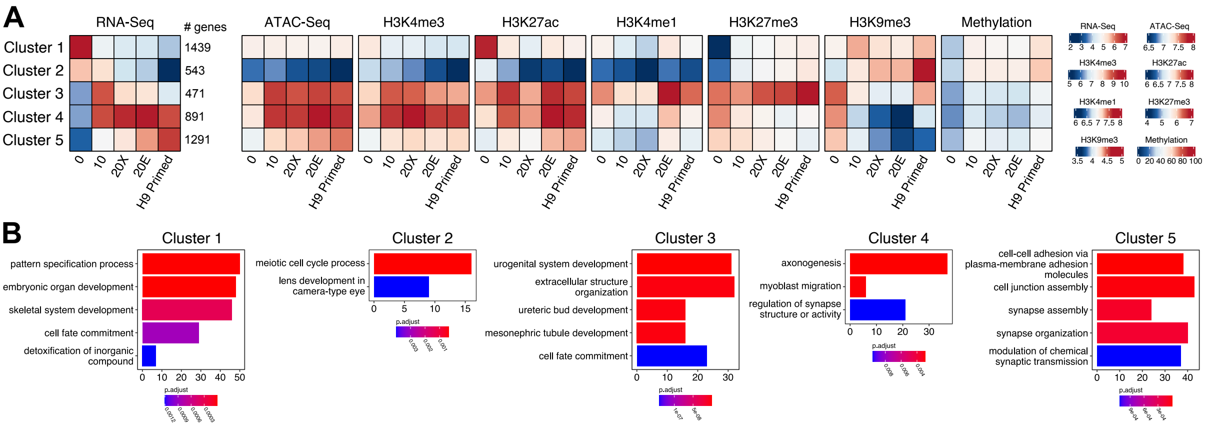

Figure 6. Differentially expressed gene clusters and associated promoter epigenetic modifications and gene ontology terms.

A) RNA-seq gene clusters defined based on their significant expression differences between any condition and their expression pattern between conditions. The heatmaps show the normalized counts of chromatin accessibility and epigenetic modifications in the promoters of the genes belonging to each cluster. **B)** Top 5 gene ontology terms associated with each RNA-seq cluster are shown. For a complete list of gene ontology terms, refer to Table 5.

2.3.2. Modelling gene expression based on promoter epigenetic modification levels

The results above suggested that there is a correlation between gene expression clusters and associated promoter epigenetic modification levels. For example, high levels of H3K27ac in promoters were associated with the expression of naïve-specific genes, whereas overall gene repression was associated with higher H3K27me3 and H3K9me3 promoter levels. Interestingly, there seemed to be only a slight correlation between promoter methylation levels and gene expression in the hPSCs examined in this study. Previous studies showed that promoter histone modification levels explain a significant percentage of gene expression variance in a multivariate linear model with an R-squared value of approximately 60-70%[104]. Therefore, to investigate how promoter epigenetic levels are associated with gene expression in this study, I employed a multivariate linear regression method to the data. I used the quantification of promoter histone modifications, DNA methylation, and chromatin accessibility to generate a model of protein-coding gene expression in each condition. The results revealed that the highest R-squared values were observed in the naïve state cells (Figure 7A). Further analysis of the model statistics indicated that the regression coefficient for H3K27ac, and thus its relationship with expression, was higher in naïve cells (Figure 7B). In contrast, the H3K4me3 regression coefficient was higher in the primed cells, which is consistent with the findings in cluster 1 of differentially expressed genes (Figure 6A). Additionally, for genes with high CpG density promoters, the H3K27ac positive regression coefficient was higher than for low CpG density promoters, as shown in Figures 7A-B and S3A-C.

By categorizing protein-coding genes based on their promoter CpG density, the overall correlation with expression improved for genes with high CpG density promoters (Figures 7A and S3D). The search for specific gene ontology terms associated with high CpG density promoters revealed terms related to embryonic organ development, chromatin organization, mRNA metabolic processes, and the Wnt signalling pathway. Genes with low CpG density promoters were associated with gene ontology terms related to sensory perception of chemical stimulus and immune response. Genes with intermediate CpG density promoters were associated with terms related to the ncRNA metabolic process, chromosome segregation, and mitochondrion organization (Figure 3E). Furthermore, the analysis also showed that lineage-specific and bivalent genes had promoters with high CpG density (Figure 7C), which supports the connection between developmental genes and high CpG density promoters[70,105]. Lineage-specific genes had consistent promoter bivalent chromatin with low expression in all conditions (Figures 7D and S3F-G). These findings highlight the putative role of epigenetic modifications in regulating gene expression. Lineage-specific genes seem to have been protected from CpG loss in their promoters, and their association with histone modifications may be an epigenetic protective or compensatory mechanism[106].

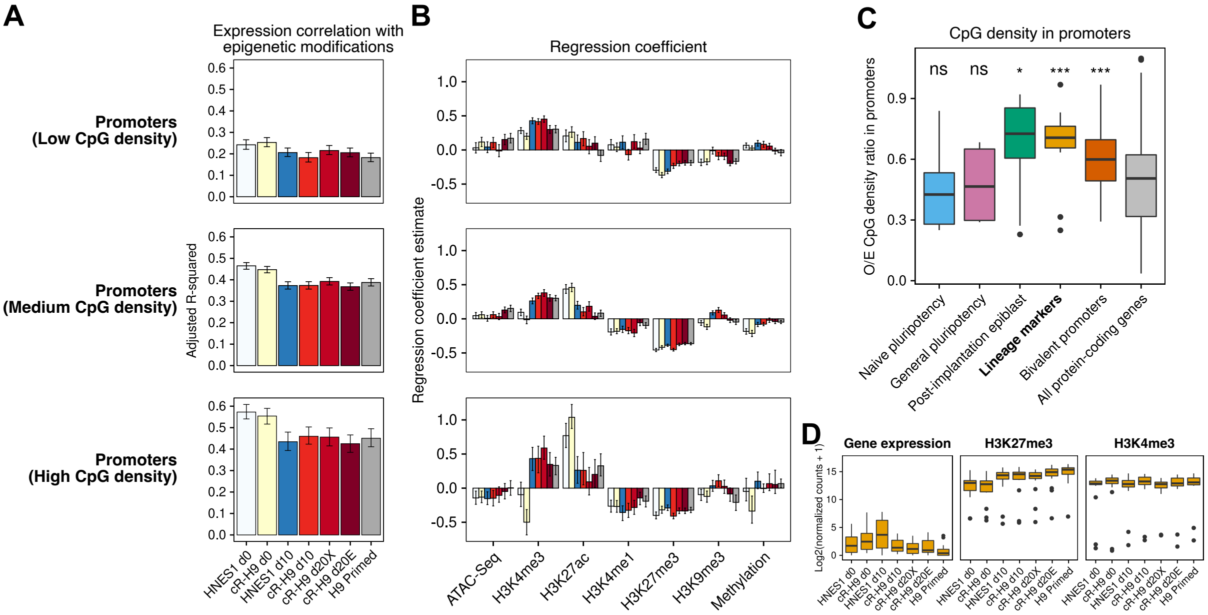

Figure 7. Association between gene expression and promoter epigenetic modifications.

A) The multivariate linear regression analysis demonstrated the relationship between gene expression and the levels of promoter epigenetic modifications in each condition. Protein-coding genes were divided into three groups based on their promoter CpG observed/expected density ratio. The adjusted R-squared values are shown for each group. **B)** The regression coefficient for each variable and condition of the multivariate linear regression model. **C)** Promoter CpG observed/expected density ratio is compared for genes associated with embryonic development and pluripotency, genes with bivalent promoters in all conditions, and all protein-coding genes. **D)** The gene expression and promoter H3K27me3 and H3K4me3 log-transformed normalized counts are shown for the lineage marker genes.

2.3.3. Promoter bivalency

Promoter bivalency is a critical mechanism for regulating gene expression in human pluripotent stem cells, involving the coexistence of two histone modifications, H3K4me3 and H3K27me3, at gene promoters. This mechanism is believed to maintain genes in a poised state, enabling rapid activation or repression during differentiation[107,108]. In this study, I observed that the number of bivalent promoters increased from the naïve state (3103 bivalent promoters in 'cR-H9 d0' cells) to the capacitated state (3847 bivalent promoters in 'cR-H9 d10' cells), with a slight increase when comparing capacitated cells to conventionally cultured H9 primed cells (3906 bivalent promoters in 'H9 Primed' cells) (Figure 8A). The promoters that became bivalent during capacitation gained H3K27me3 during the transition. Furthermore, several genes with H3K27me3 but not H3K4me3 modifications overlapping their promoters in the naïve state lost the H3K27me3 marks during capacitation. After capacitation, their promoters had higher methylation levels (data not shown). Finally, genes that maintained promoter bivalency across pluripotency states were associated with gene ontology terms related to pattern specification, neuron projection, and forebrain development (Figure 8B).

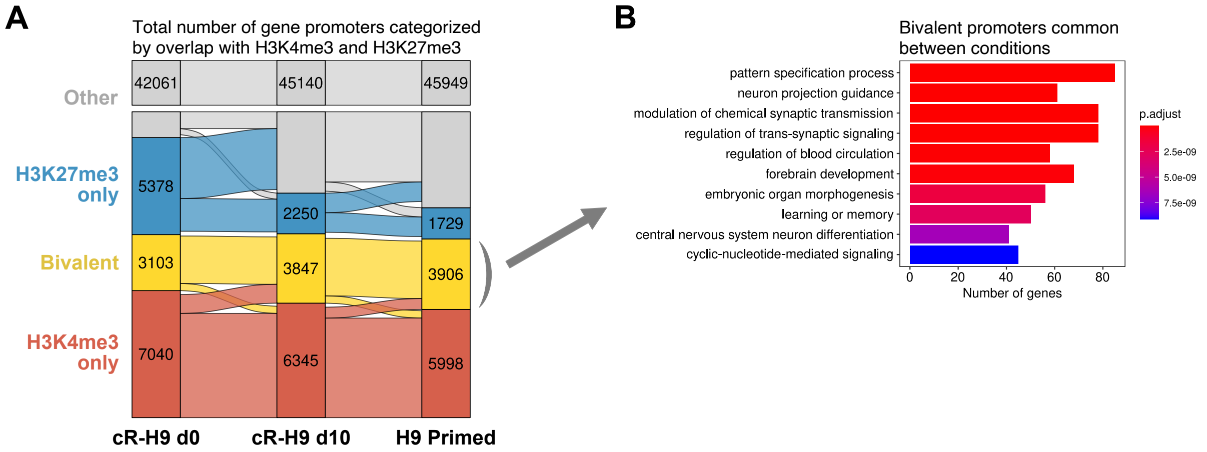

Figure 8. Comparison of promoters containing H3K4me3 and H3K27me3 marks in H9 cells and associated gene ontology terms.

A) Flow plot illustrating the number of promoters in H9 naïve, capacitated, and conventionally cultured primed cells categorized into four groups: 'Other' (gray), containing promoters without H3K27me3 and H3K4me3 ChIP-seq peaks; 'H3K27me3 only' (blue), containing promoters with H3K27me3 but not H3K4me3 peaks; 'Bivalent' (yellow), containing promoters with both H3K27me3 and H3K4me3 peaks; and 'H3K4me3 only' (red), containing promoters with H3K4me3 but not H3K27me3 peaks. **B)** Gene ontology terms associated with genes that have promoters maintaining the 'Bivalent' classification across all conditions.

2.4. Epigenetic dynamics between pluripotency states

2.4.1. Combinatory role of epigenetic modifications

To gain insights into the interaction between the epigenetic modifications across the various conditions, I employed a Hidden Markov Model (HMM) analysis. The HMM segmented the genome into 13 regions, which I classified based on the specific combination of histone modification and accessibility across all samples (see methods for details). The analysis confirmed that, in the primed state, H3K27me3 increases in CpG islands and CGI promoters when associated with promoter or enhancer marks, represented by states 1 and 2. Additionally, the model identified regions associated with CpG islands, CGI promoters, and enhancers that showed increased accessibility in the primed state (Figures 9 and S4A). These findings suggest an association between chromatin accessibility and regulatory regions in cells acquiring lineage specification competency. Moreover, CpG islands overlapping promoters acquired bivalent epigenetic patterns, as observed in Figure 8.

Notably, the HMM model identified two epigenetic states (states 11 and 12) enriched in the X chromosome, suggesting that conventional H9 cells had an eroded X chromosome and that capacitated cells, after resetting and capacitation, recovered the inactive X chromosome.

State 11 selected genomic regions with intermediate levels of H3K9me3 only in conventional H9 cells, whereas state 12 selected genomic regions with high H3K27me3 levels in all reset H9 cells and intermediate H3K9me3 levels in conventional H9 cells (Figure 9).

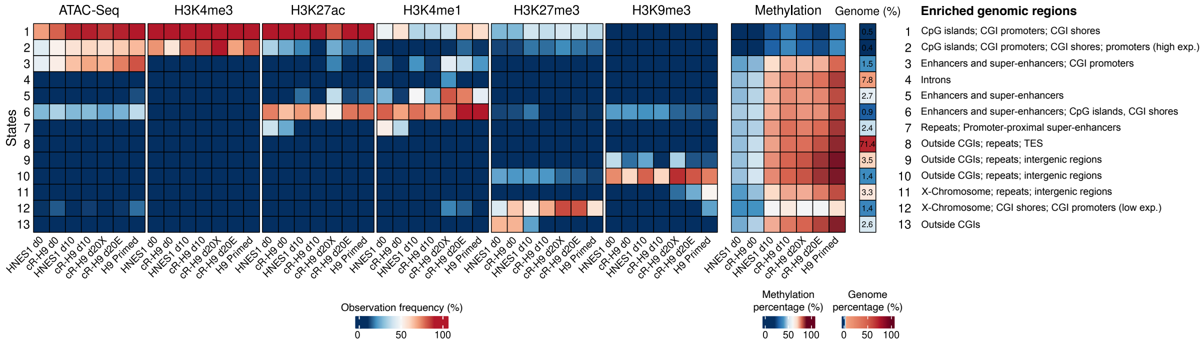

Figure 9. Chromatin state model of genomic distribution of histone modification and chromatin accessibility data across conditions.

The ChromHMM 13-state genome-wide model was built using ATAC-seq and histone modifications' ChIP-seq aligned reads. The heatmap shows the distribution of histone modification and chromatin accessibility across different conditions, including the average methylation percentage, the percentage of the genome classified with each state, and the top enriched regions. Only autosomes were considered for all analyses, except for ChromHMM, which included the X and Y chromosomes.

2.4.2. Multi-omics factor analysis (MOFA)

Following a genome-wide analysis of the epigenetic landscape in the naïve and primed pluripotency states, I aimed to identify the primary sources of variation between conditions using the multi-omics dataset. I performed a multi-omics probabilistic factor analysis using the R package MOFA2[95]. To generate the data input, I selected the most variable regions from each dataset (see methods). The model extracted one factor (factor 1) that successfully distinguished the samples based on their pluripotency state (Figures 10A-B and S4B-C). The analysis of this factor revealed that the most significant epigenetic changes during capacitation were the increase in DNA methylation and a decrease in H3K27me3 marks. Furthermore, the MOFA2 analysis highlighted that primed-specific chromatin-accessible regions overlapped with regulatory elements, particularly putative enhancers (Figures 10C-D), emphasising the results from figures 2A and 5B.

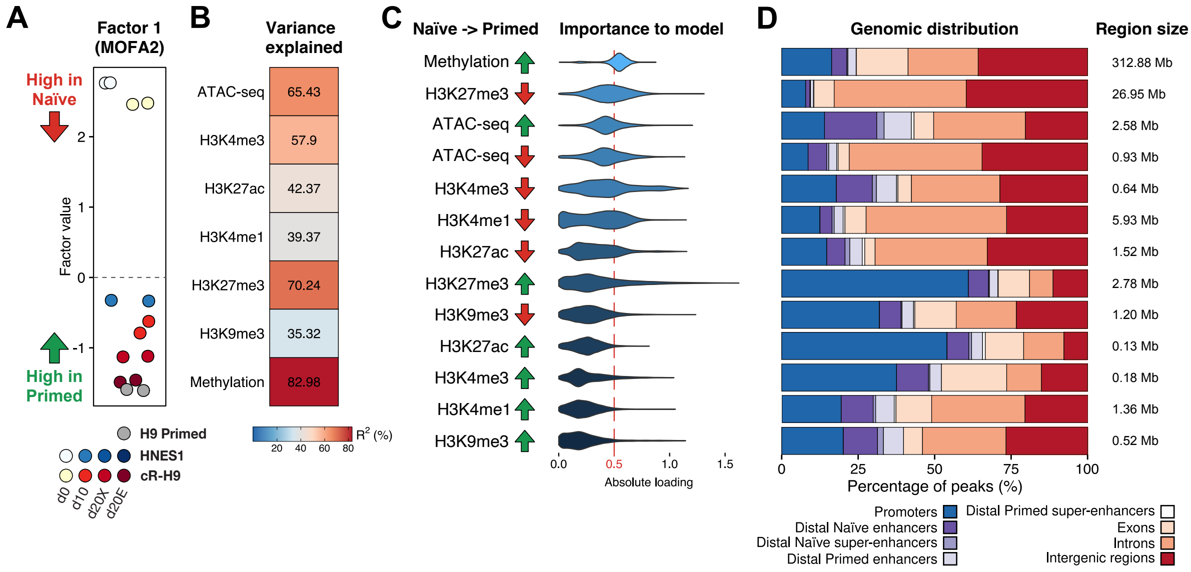

Figure 10. Analysis of MOFA2 factor 1 that distinguishes naïve and primed cells.

A) MOFA2 factor values for factor 1, extracted from differential peaks of histone modifications' ChIP-seq, ATAC-seq, and the top 200 variable CpG-containing regions based on methylation percentage between conditions. **B)** R-squared values indicating the variance explained by each sequencing assay in the sample clustering of factor 1. **C)** Top features ordered by their model loading weight (expressed as "importance to the model"). Red arrows indicate features with high levels in naïve cells and low levels in primed cells, while green arrows represent the opposite trend. **D)** Genomic distribution and size of the top features selected by having an absolute loading weight above 0.5 in factor 1. The results were limited to autosomes.

2.4.3. Enrichment analysis of MOFA factor 1 over annotated genomic regions

To identify the genomic features with the most significant epigenetic changes between pluripotency states, I collected the data points from the MOFA2 model with an absolute loading greater than 0.5 in factor 1. Then, I used the LOLA[109] package to test their overlap with public and custom genomic annotation databases. The results from the LOLA analysis showed that the primed regions with increased accessibility during capacitation were enriched for primed-specific binding sites of NANOG, OCT4 (POU5F1), and SOX2, which are canonical pluripotency transcription factors (Figure 11A). Additionally, the remaining epigenetic changes with a heavier weight in factor 1 had two primary associated features: CpG islands (CGIs) and repetitive elements.

The factor analysis confirmed that DNA methylation increased outside CGIs, and H3K27me3 marks had a relative increase almost exclusively in CGIs and CGI shores (Figure 11A). Surprisingly, regions with decreasing levels of H3K9me3 between naïve and primed had a strong enrichment in CpG islands and were associated with genes related to organ development (Figures 11A and S4D). These CpG islands were shorter and more enriched in gene bodies than H3K27me3-overlapping CpG islands, which were long and mostly near transcription start sites (TSS). Additionally, regions with increased levels of H3K4me1 between naïve and primed had a significant overlap with distal CpG islands, also known as 'orphan' CpG islands (Figure S4F), highlighting the potential role of orphan CpG islands to function as regulatory elements during the induction of developmental genes[110].

Retrotransposons were marked by different epigenetic modifications in the two pluripotency states. In the naïve state, Alu elements were overlapped by poised enhancer marks, H3K4me1 and H3K27me3, while SVA elements by active promoter marks, H3K4me3 and H3K27ac. After capacitation, almost all classes of repetitive elements increased DNA methylation levels, including retrotransposons, with Alu and SVA elements showing higher methylation levels in these cells. Finally, SVA elements, ERV1, and ERVK were more accessible in the naïve state compared to the primed state (Figures 11B and S4H), which is consistent with the higher expression of those elements in naïve hPSCs[15] (Figure 11C).

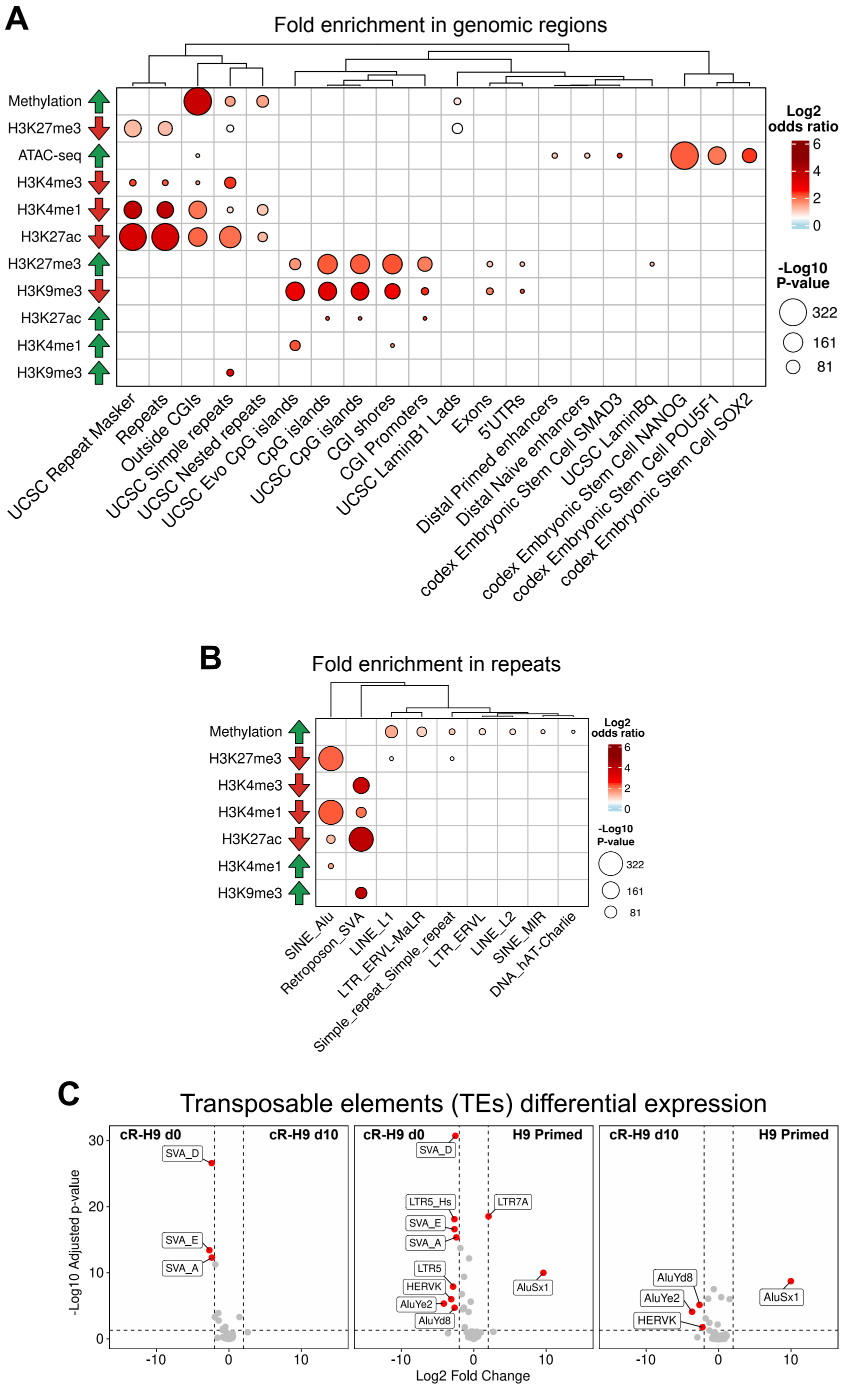

Figure 11. Analysis of epigenetic changes and transposable element expression between pluripotency states. **A)** Enrichment of the top features in factor 1 over annotated genomic regions. The features were selected based on their absolute loading weight greater than 0.5 over annotated genomic regions. The red arrows indicate features with high levels in naïve cells and low levels in primed cells, while green arrows represent the opposite trend. **B)** Enrichment of the top features in factor 1 over repeat elements. The repeat element names are divided into repeat family and, followed by an underscore, the name of the repeat class. **C)** Differential expression of transposable element classes between 'cR-H9 d0' and 'cR-H9 d10' cells, 'cR-H9 d10' and 'H9 Primed' cells, and 'cR-H9 d10' and 'H9 Primed' cells. The results in red were selected based on an absolute log2 fold change greater than 2 and an adjusted p-value less than 0.05. The results were limited to autosomes.

2.4.4. Motif enrichment in differentially accessible regions

An analysis of the read count distribution in primed-specific accessible regions with canonical pluripotency transcription factors binding sites showed that, from naïve to primed, these regions lost H3K27me3 and H3K9me3 marks and maintained low DNA methylation levels. Additionally, the H3K27ac and H3K4me1 modifications changed their profile from having the highest value at the centre of the accessibility regions in the naïve state to flanking the peak region in the primed state, a pattern commonly observed in enhancers[111] (Figure 12A).

Motif analysis revealed that the OCT4-SOX2 heterodimer motif was specific to these regions with increased accessibility in the primed state (Figure 12B). Gene ontology analysis of the genes located within 50 Kb of these regions showed an association with organ morphogenesis (Figure S4D). Conversely, regions with a significant decrease in accessibility from naïve to primed states had a considerable loss of active marks spanning the peak regions with increased DNA methylation and H3K9me3 (Figures 12A and S4E, G).

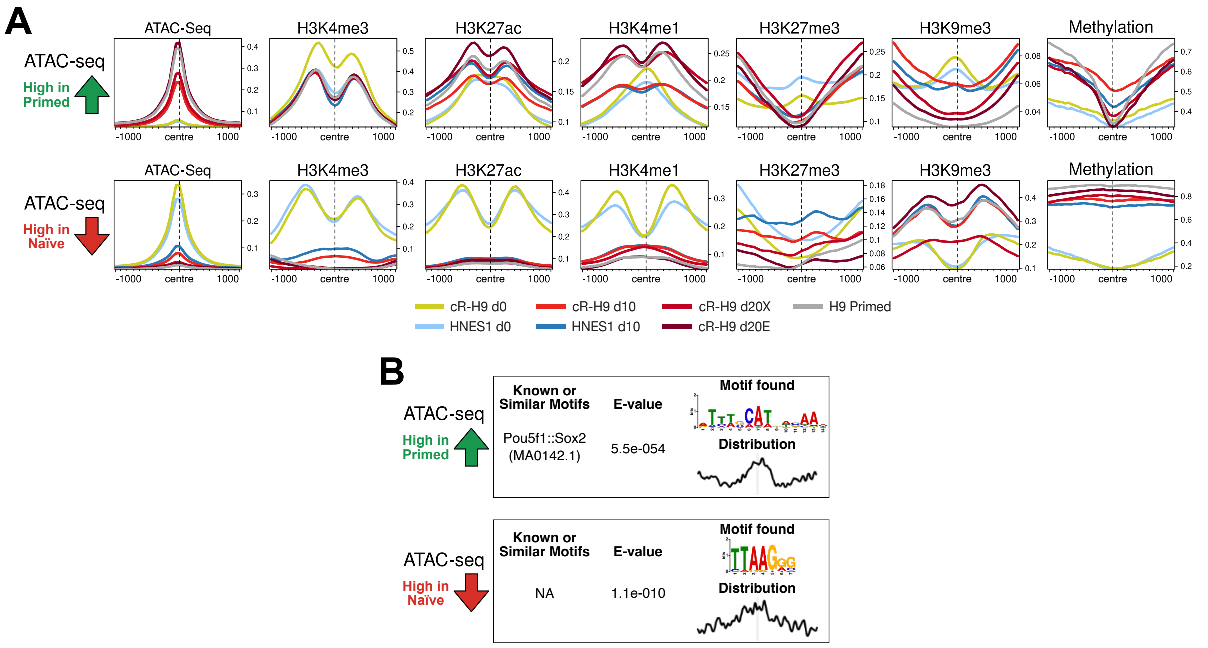

Figure 12. Epigenetic profiles and motif enrichment of accessible regions undergoing significant changes between naïve and primed states.

A) Epigenetic profiles of regions centred at the top ATAC-seq peaks, selected based on an absolute loading weight above 0.5 in factor 1. The top row shows profiles of ATAC-seq peaks with low normalized counts in the naïve state and high normalized counts in the primed state. The bottom row shows profiles of ATAC-seq peaks with high normalized counts in the naïve state and low normalized counts in the primed state. **B)** Top results of the motif analysis conducted on the same ATAC-seq peaks selected based on the highest factor 1 loading weight features. Results were limited to autosomes.

2.5. X chromosome-specific epigenetic landscape

2.5.1. Chromosome-wide epigenetic modification levels

In human pre-implantation female epiblast cells, both X chromosomes are active and their expression is regulated through transcriptional dampening[80]. However, during implantation, one of the X chromosomes is randomly silenced by the action of the long non-coding RNA XIST and remains silenced in somatic cells[112]. Interestingly, in long-term cultured conventional hPSCs, the silenced X chromosome can regain activity showing a reduction in XIST expression and a loss of H3K27me3-enriched domains[9]. This phenomenon is known as 'erosion' (see Introduction).

Figure 9 showed that the HMM model identified two epigenetic states (states 11 and 12) enriched in the X chromosome, indicating that conventionally cultured H9 cells had an eroded X chromosome. Interestingly, it also suggested that capacitated cells recovered the inactive X chromosome. Indeed, upon examining the genome-wide distribution of ChIP-seq peaks, it was possible to confirm the significantly higher percentage of H3K27me3 peaks mapped to the X chromosome in the reset naïve hPSCs compared to the conventionally cultured primed H9 cells. Furthermore, conventional H9 cells had a higher percentage of H3K9me3 peaks and more regions with high H3K9me3 counts spanning the chromosome (Figures 13A-B and S5A). Notably, XIST maintained robust expression on all days in culture in reset naïve H9 cells, while in conventional primed H9 cells, it had reduced expression, a hallmark of an eroded X chromosome (Figure 13C).

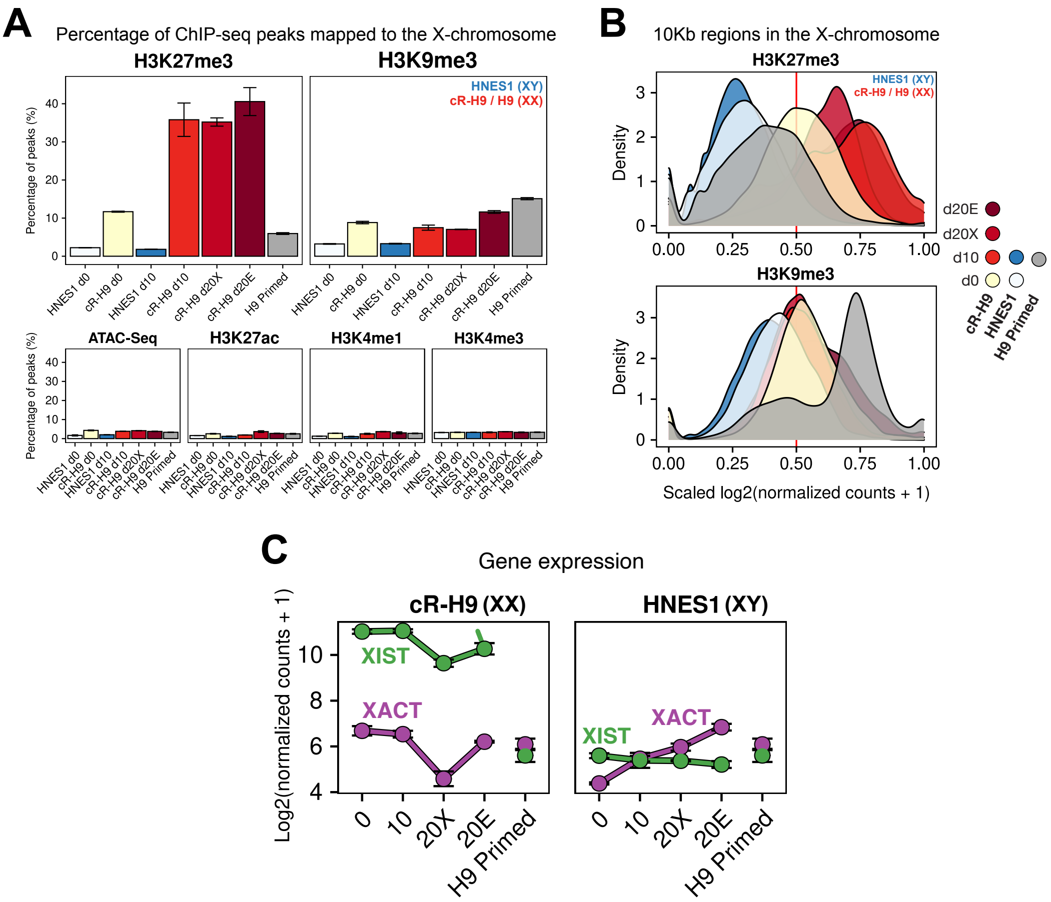

Figure 13. Epigenetic modification enrichment and XIST expression in the X chromosome.

A) The percentage of ChIP-seq peaks for H3K27me3 and H3K9me3 histone modifications mapped to the X chromosome. **B)** The distribution of log-transformed normalized counts of H3K27me3 and H3K9me3 in 10 kilobase windows across the X chromosome. The red line indicates the threshold used to categorize genomic regions as high or low H3K27me3 and H3K9me3 in Figure 14. **C)** The gene expression levels of XIST and XACT in reset H9 and HNES1 cells, in comparison to conventionally cultured H9 cells.

2.5.2. Epigenetic differences between conditions

To investigate how the X-chromosome enrichment of repressive histone modifications change between conditions, I divided the X chromosome of H9 cells into 10-kilobase regions and classified them based on H3K27me3 and H3K9me3 levels (Figures 14A and S5B-C). I then performed an enrichment analysis to identify the overlap of these regions with genomic features. The analysis showed that some regions with high levels of H3K9me3 marks and low H3K27me3 marks in the naïve state gained H3K27me3 marks during capacitation (Figure 14A and S5B-E; transition ‘1’). These regions had a significant overlap with LaminB1 LADs, nested repeats, and ERVK (Endogenous retrovirus-K). After capacitation, the distribution of the number of regions by classification remained stable between the capacitated and long-term cultured cells. However, the long-term cultured cells in the E8 medium had more regions classified for having both H3K27me3 and H3K9me3 marks than the cells cultured in the XAF medium.

Notably, most chromosome regions in conventionally cultured H9 cells had high H3K9me3 levels (Figures 14A-B; transition ‘5’). These regions were instead enriched for H3K27me3 and H3K9me3 marks on day 20E cells and overlapped with LaminB1 Lads, nested repeats, and L1 (LINE) retrotransposons (Figures 14A and S5D-E; transition ‘5’). Previous studies have shown that in mouse cells, Xist directly interacts with the Lamin B receptor and recruits the inactive X to the nuclear lamina, enabling it to spread to actively transcribed genes[113]. Therefore, the regain of H3K27me3 domains in LaminB1 Lads in the capacitated cells may reflect gene inactivation through XIST-dependent recruitment of the X chromosome to the nuclear lamina. These results suggest that capacitation, after resetting, corrected inactive X chromosome epigenetic aberrations.

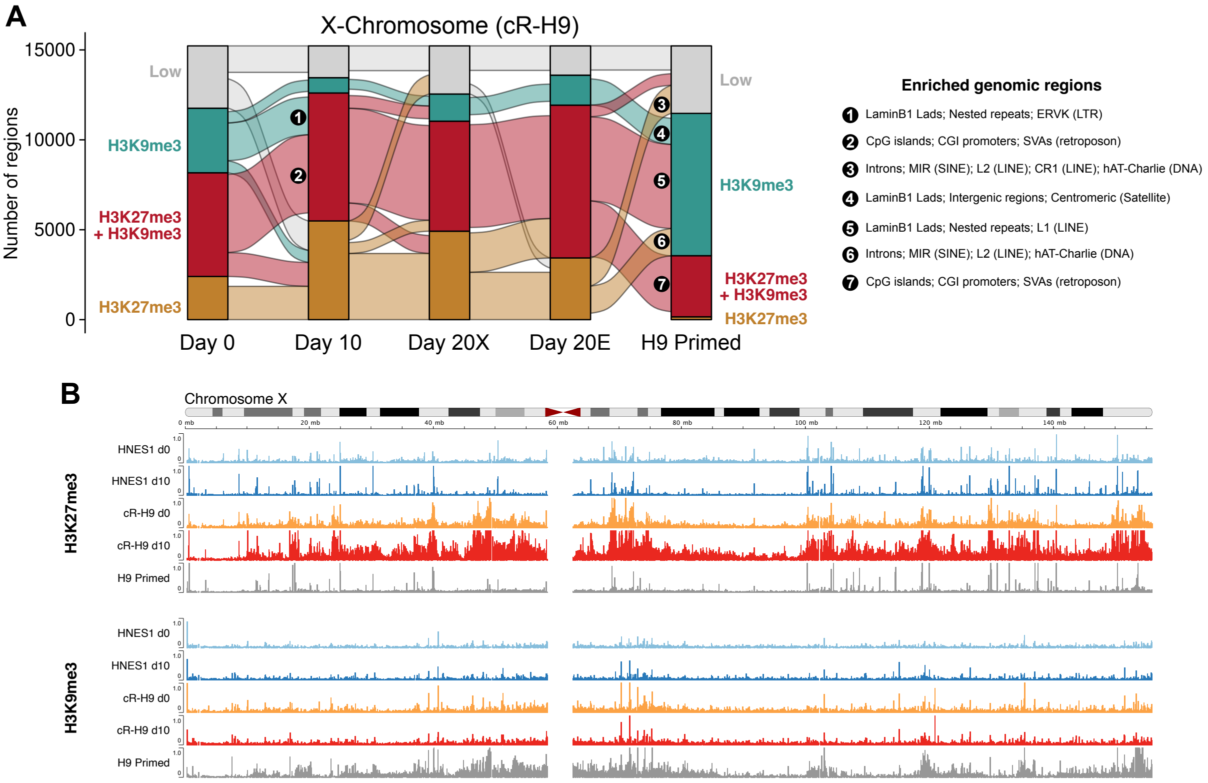

Figure 14. Differential levels of H3K27me3 and H3K9me3 in the X chromosome across experimental conditions. **A)** Flow chart illustrating the number of 10-kilobase windows in the X chromosome classified based on a scaled log2 normalized counts threshold of 0.5. The classification includes windows containing high levels of H3K27me3 marks (‘H3K27me3’), high levels of H3K9me3 marks (‘H3K9me3’), high levels of both H3K27me3 and H3K9me3 marks (‘H3K27me3 + H3K9me3’), and low levels of both marks (‘Low’). The chart shows the results for H9 cells only. The right side of the flow chart indicates the top enriched genomic regions in the X chromosome for each numbered classification transition between conditions. **B)** H3K27me3 and H3K9me3 read count profiles across the X chromosome for naïve, capacitated day 10, and conventionally cultured human pluripotent stem cells (hPSCs).

3. Discussion

Naïve human pluripotent stem cells (hPSCs) represent an early embryonic cell state resembling the blastocyst's inner cell mass (ICM)[1,2]. These cells are pluripotent, meaning they can differentiate into any cell type in the human body. They are characterized by specific gene expression patterns, including the activation of pluripotency-associated genes and naïve-specific expression[114]. In contrast, primed hPSCs are more similar to post-implantation epiblast cells and have undergone lineage priming, resulting in a more limited differentiation potential. These cells are predisposed to differentiate into specific cell lineages, such as ectoderm, mesoderm, and endoderm, which are the primary germ layers of the developing embryo[1,2]. These two types of pluripotent stem cells possess distinct transcriptional, epigenetic, and metabolic characteristics. However, our current knowledge of their disparities primarily relies on studies conducted on model organisms, which may not accurately represent the behaviour of human cells[34]. Therefore, the primary aim of this study was to comprehensively characterize the transcriptional and epigenetic profiles of naïve and primed hPSCs. To this end, we used a culture system that allowed naïve hPSCs to transition *in vitro* to the primed state with a duration comparable to human development[10]. This system enabled us to compare the differences between primed hPSCs that underwent an *in vitro* transition from the naïve to the primed state of pluripotency (referred to as capacitation) and primed hPSCs cultured conventionally in an E8 medium. This comparison holds significant importance in evaluating the potential benefits of utilizing primed cells after *in vitro* capacitation for differentiation and the potential application in induced pluripotent stem cells (iPSCs). Thus, to gain a deeper understanding of the molecular changes occurring during a stage of human development that is challenging to study due to technical and ethical limitations, we cultured HNES1 and chemically reset H9 hPSCs in a medium that facilitated their transition from a naïve to a primed state of pluripotency, as described above. As previously shown, this *in vitro* transition from naïve to primed pluripotency replicates the effects of implantation on the transcriptional machinery of human and primate epiblast cells during early development[10]. Moreover, this system provides an opportunity to characterize cell behaviour during early human development and, importantly, to identify and manipulate molecular factors involved in lineage specification in a scalable and ethical manner. Furthermore, it enables us to improve our understanding of the molecular characteristics of naïve and primed cells, which is necessary for developing new stem cell-based therapeutics.

This study primarily focused on exploring epigenetic modifications, such as DNA methylation and histone modifications, due to their significant roles in regulating gene transcription and chromatin organization[115]. These modifications are closely associated with the cell's metabolism and response to environmental cues, making them ideal targets for studying highly dynamic cellular state transitions[116]. Thus, I conducted a multi-omics analysis of five canonical histone modifications (H3K4me3, H3K27ac, H3K4me1, H3K27me3, and H3K9me3) obtained from a ChIP-Seq assay and 5-methylcytosine (DNA methylation) levels from a post-bisulfite adaptor tagging (PBAT) assay. Additionally, I examined chromatin organization data from an ATAC-Seq assay and analysed gene expression from an RNA-Seq assay. This combination of assays allowed me to characterize the epigenetic landscape in both naïve and primed cells and model the association between gene expression and the levels of promoter epigenetic modifications in the corresponding genes. Moreover, I provided evidence showing that resetting primed hPSCs to the naïve pluripotency state, followed by capacitation back to the primed pluripotency state, corrects some of the epigenetic aberrations observed in conventionally cultured primed hPSCs[32].

By performing a principal component analysis (PCA) and t-SNE (t-distributed stochastic neighbour embedding) I showed that the sequencing assays data clustered the samples into two primary groups: naïve and primed pluripotency. However, in most sequencing assays, capacitated (day 10) and long-term capacitated (day 20+) cells exhibited separate subclusters, despite clustering within the primed category. Notably, long-term capacitated cells clustered near conventionally cultured H9 cells, particularly those cultured in the E8 medium. This observation suggests that the composition of the medium impacts the transcriptional and epigenetic landscape. Intriguingly, the histone modifications H3K27ac and H3K4me1 did not display the clustering division between capacitated, long-term capacitated, and conventionally cultured primed hPSCs. According to the findings from these sequencing assays, primed samples consistently clustered together. Since H3K27ac and H3K4me1 are associated with enhancers[50], this implies that enhancer regions undergo remodelling during the transition from naïve to primed states. However, their association with these histone modifications appears to remain mostly unaffected by long-term culture conditions, potentially due to their active maintenance, as they may be relevant during the subsequent lineage differentiation processes.

An analysis of the ChIP-Seq peaks of histone modifications, which are regions with a high read enrichment compared to the genomic background, confirms previous findings indicating an increased enrichment of H3K27me3 modifications in CpG islands between naïve and primed pluripotency[117]. Simultaneously, I observed a global decrease in H3K27me3 and a global increase in DNA methylation, suggesting that in the naïve state, H3K27me3 is widely distributed throughout the genome, including CpG island regions, while DNA methylation remains predominantly low. However, during capacitation, or implantation in early human embryonic development, the H3K27me3 levels decrease with the simultaneous DNA methylation levels increase, implying that they may have a compensatory function in some genomic regions, depending on the cell's pluripotency state. Moreover, DNA methylation levels remained low in CpG islands in the naïve and primed states, while H3K27me3 exhibited an increased enrichment in those regions. This result suggests that H3K27me3 plays a compensatory and repressive role in genomic regions lacking DNA methylation. However, further research is necessary to understand the relationship between the histone modification H3K27me3 and DNA methylation in hPSCs. Additionally, I observed a noticeable reduction in H3K27me3 levels between the two pluripotency states in repetitive elements.

The analysis of the number of ChIP-Seq peaks and histone modifications at the protein level, assessed using Western blot, also revealed an overall rise in H3K4me3, H3K27ac, and H3K4me1 levels between naïve and primed pluripotency, with H3K4me3 displaying an increased enrichment in gene promoters. Since these histone modifications are associated with enhancers and active promoters, this increase could indicate a more intricate transcriptional profile of primed cells and the preparation of specific regulatory regions necessary for gene expression during lineage specification. Moreover, the ChIP-Seq peak analysis revealed that the chromatin accessibility landscape is more defined in the primed state compared to the naïve state, particularly in regions marked by epigenetic modifications associated with active promoters and enhancers. Notably, genomic regions that became accessible in the primed state displayed an enhancer-like epigenetic profile. These regions were highly enriched with the OCT4-SOX2 heterodimer motif and associated with genes involved in organ morphogenesis. These findings support the hypothesis that the OCT4-SOX2 heterodimer plays a role in the regulatory machinery during embryonic development and differentiation, extending beyond its role in inducing and maintaining pluripotency[118]. Furthermore, when examining the relationship between various epigenetic modifications and chromatin accessibility, I observed that in primed cells, DNA methylation exhibited a reduced correlation with other epigenetic modifications in long-term cultured cells. However, this diminished correlation was not as pronounced in capacitated primed cells at day 10. These findings suggest that extended culture not only leads to increased DNA methylation but also promotes the segregation of DNA methylation from regions associated with histone modifications, implying that long-term culture may have additional effects on the epigenetic landscape by influencing the relationship between DNA methylation and histone modifications.

Because I observed a global shift in the epigenetic and transcription landscape during the transition from naïve to primed state, I then aimed to characterize the levels of epigenetic modifications on gene promoters in the different conditions. This analysis would help examine the role of epigenetic modifications on gene expression. To achieve this, I generated five clusters from a differential gene expression analysis, each exhibiting distinct expression patterns across conditions. Then I measured the levels of epigenetic marks in the promoters of genes within each cluster. The findings revealed that genes exclusively expressed in the naïve state appeared to rely more on H3K27 acetylation (H3K27ac) marks at their promoters for transcription. On the other hand, genes expressed solely in the primed state showed a stronger association with H3K4 trimethylation (H3K4me3) levels, as observed in cluster 1. In the remaining clusters, higher gene expression seemed to be linked to elevated levels of "active" epigenetic modifications in the promoters, while lower gene expression was instead associated with increased levels of "repressive" epigenetic modifications. Intriguingly, a gene ontology analysis demonstrated that genes predominantly expressed in capacitated cells (day 10) with decreased expression in the long-term cultured cells (day 20+) were associated with terms related to urogenital system development and extracellular structure organization, implying that capacitated cells exhibited active mesoderm-associated genes that were repressed in long-term cultured cells. Additionally, genes in cluster 5 showed high expression exclusively in long-term cultured cells following capacitation and conventional H9 cells. Those genes were associated with ectoderm lineage terms such as synapse assembly and organization. Hence, it appears that capacitated cells display gene expression signatures related to both mesoderm and ectoderm despite possessing multi-lineage differentiation potential, including endoderm. However, the observed mesoderm lineage bias in capacitated cells could be explained by the heterogeneity of the cell population under study. As the cells were sequenced in bulk, if a subset of cells began expressing lineage markers, those specific genes would exhibit increased expression compared to the other conditions. Moreover, I observed an increase in the number of genes with both H3K27me3 and H3K4me3 peaks overlapping their promoters (bivalent genes) during capacitation, from 3101 genes in naïve cells to 3847 genes in capacitated cells. This result may be explained by the distinct roles of naïve and primed cells during early human development. Primed cells possess the ability to respond to lineage cues without committing entirely to a specific lineage. The increased promoter bivalency could reflect their behaviour. Once the cells receive lineage cues, they determine the lineage to which they will differentiate. Genes that maintain promoter bivalency across pluripotency states were associated with gene ontology terms related to pattern specification, neuron projection, and forebrain development.

Afterwards, I aimed to determine whether promoter epigenetic modification levels could be used to model gene expression and under which conditions they have the strongest association. I segmented the promoters into three categories based on their CpG density: "Low CpG density," "Medium CpG density," and "High CpG density." Through this model, I could confirm the association between H3K27ac and naïve-specific expression. However, further research is necessary to fully understand the relationship between promoter H3K27ac levels and naïve-specific expression. The model metrics also showed a significant difference in how the model performed based on the CpG density of the promoters. Genes with promoters that had higher CpG density exhibited a stronger association between gene expression and promoter epigenetic modification levels. These findings may be attributed to the different promoter sequences that potentially necessitate a closer relationship between gene expression and histone modifications, DNA methylation, and accessibility. On the other hand, promoters with high CpG density may have been evolutionarily selected to have a stronger association with gene expression and promoter epigenetic modification levels. Indeed, when analysing the promoter CpG density levels of genes associated with naïve pluripotency, general pluripotency, post-implantation epiblast expression, lineage-specific expression, bivalent promoters, and all protein-coding genes, I observed that lineage markers exhibited significantly higher promoter CpG density compared to all protein-coding genes. I obtained similar results for genes with bivalent promoters and genes associated with post-implantation epiblast expression. Furthermore, the lineage marker genes subset used in this analysis revealed that they have promoters with high levels of H3K27me3 and H3K4me3, indicating a distinct pattern of bivalent promoters and low or no expression. This suggests that histone modifications play a regulatory role in activating or repressing developmental genes, which aligns with previous observations[70]. These genes' promoters were likely evolutionarily selected to have low deamination or mutation rates, potentially involving epigenetic mechanisms in their protection[105].

After gaining an overview of the epigenetic landscape across conditions, I attempted to identify genomic features that exhibited significant changes in epigenetic modification levels and accessibility status. To achieve this, I used the Bioconductor package MOFA2 to generate a multi-factor analysis model[95]. This model produced factors that clustered the samples based on specific features in the dataset that could be extracted for further analysis. Among these factors, factor 1 effectively separated the samples into two main clusters: naïve and primed pluripotency. The naïve cluster consisted of HNES1 and cR-H9 cells from day 0, while the primed cluster included samples from day 10, day 20+, and conventional primed hPSCs. The distinction between these clusters primarily arose from changes in DNA methylation, H3K27me3, accessibility, and H3K4me3 levels throughout the genome. Notably, the most impactful changes driving the factor 1 sample clustering were: an increase in DNA methylation and a decrease in H3K27me3 between the naïve and primed states. These shifts spanned large genomic regions, with approximately 312 Mb for DNA methylation increase and 27 Mb for H3K27me3 decrease. Furthermore, regions exhibiting changes in chromatin accessibility also significantly influenced sample clustering in the model. This observation suggests a remodelling of regulatory regions and highlights the specific role of these regions in either inducing or responding to external signals that modify gene expression. In fact, upon conducting an enrichment analysis for regions with significant epigenetic changes between naïve and primed pluripotency from factor 1, I discovered that regions with increased accessibility were enriched with putative distal enhancers and genomic regions known to be bound by pluripotency-associated transcription factors (NANOG, OCT4, and SOX2) in conventional primed hPSCs. Conversely, regions that showed a significant decrease in accessibility between naïve and primed states lost their epigenetic profile associated with regulatory regions. These regions exhibited higher levels of repressive chromatin markers such as H3K4me3, H3K27ac, H3K4me1, H3K9me3, and DNA methylation. Unfortunately, no significant motif was associated with these regions. During the enrichment analysis of regions significantly associated with sample clustering in factor 1 of the multi-factor analysis, two primary enrichment categories based on genomic features emerged: CpG islands and repetitive elements. Regions exhibiting an increase in H3K27me3 and a decrease in H3K9me3 between the naïve and primed states showed significant enrichment in CpG islands. Surprisingly, regions with decreasing levels of H3K9me3 between naïve and primed demonstrated strong enrichment in CpG islands and were linked to genes involved in organ development. These CpG islands were shorter and more enriched in gene bodies compared to H3K27me3-overlapping CpG islands, which tended to be long and located mostly near transcription start sites (TSS). Furthermore, regions with increased levels of H3K4me1 between naïve and primed exhibited a significant overlap with distal CpG islands, commonly known as 'orphan' CpG islands. This finding highlights the potential role of orphan CpG islands as regulatory elements during the induction of developmental genes[110]. Repetitive regions displayed a significant overlap with regions experiencing decreasing levels of H3K27ac, H3K4me1, and H3K4me3. Specifically, the transposable elements (TEs) belonging to Short Interspersed Nuclear Elements (SINEs) and retrotransposons classes of TEs, Alu and SVA, respectively. These TEs exhibited the most significant epigenetic changes between naïve and primed pluripotency. Alu elements displayed decreased H3K27me3 and H3K4me1 levels during the naïve-to-primed transition, while SVA elements demonstrated a decrease in H3K27ac, H3K4me3, and H3K4me1, along with a simultaneous increase in H3K9me3 levels. These changes in epigenetic modifications associated with active transcription align with the differential expression of transposable elements between naïve and primed states. Specifically, SVA elements exhibited a significant expression loss from naïve to primed pluripotency. These findings are consistent with previous reports on transposable element expression in naïve and primed hPSCs[15].

During the genome-wide analysis, I observed that capacitated cells (d10 and d20+) were strongly associated with H3K27me3 ChIP-Seq peaks in their X chromosome. In contrast, conventional hPSCs displayed a higher percentage of H3K9me3 ChIP-Seq peaks in the X chromosome, compared to both naïve and capacitated cells. As shown previously, conventional hPSCs tend to undergo a loss of X chromosome inactivation over time in culture, leading to the reactivation of the silenced chromosome and the establishment of distinct epigenetic characteristics. This phenomenon, known as "erosion," involves a chromosome-wide reduction in H3K27me3 marks while preserving H3K9me3 marks[9]. Additionally, the expression of the lncRNA XIST, known to play a pivotal role in X chromosome inactivation, declines, while the lncRNA XACT expression, proposed to contribute to X chromosome activation, increases. The results presented in this dissertation revealed a substantial number of regions in capacitated cells exhibiting higher levels of H3K27me3, whereas conventional hPSCs had predominantly lower H3K27me3 levels on their X chromosomes. Additionally, conventional hPSCs displayed a greater number of regions characterized by elevated H3K9me3 levels. Moreover, there was a significant difference in XIST expression between conventional hPSCs and the other conditions. XIST showed high expression in day 0, day 10, and day 20+ cells, but it was expressed at a lower level in conventional hPSCs. On the other hand, XACT exhibited low to moderate expression across all conditions. These findings suggest that the conventional hPSCs in this study had an eroded X chromosome. Interestingly, since the cR-H9 naïve cells were derived from the conventional hPSCs H9 cells, the results indicate that a round of resetting and capacitation reversed the culture-induced X chromosome epigenetic aberrations, such as erosion of X chromosome inactivation. Thus, I then investigated the dynamics of H3K27me3 and H3K9me3 marks in specific X chromosome regions. For this purpose, I divided the X chromosome into 10-kilobase windows and categorized them based on H3K27me3 and H3K9me3 levels in each condition, tracking their states. Consistent with the number and density of these histone modifications' peaks, the conventional H9-EOS cells displayed a minor fraction of regions marked solely by H3K27me3, a reduced number of regions marked by both H3K27me3 and H3K9me3 and a significant proportion of regions marked solely by H3K9me3 compared to the naïve and capacitated cells. Moreover, despite the lower levels, H3K27me3 was still consistently maintained on the X chromosome of conventional hPSCs, primarily in combination with H3K9me3, in the regions enriched for CpG islands. While regions marked by a combination of H3K27me3/H3K9me3 in capacitated cells but by H3K9me3 only in the conventional cells were enriched in repeats and retrotransposons. Regions exclusively marked by H3K27me3 in the capacitated cells but by H3K9me3 in the conventional cells were enriched in repeats and Lamin B1 binding regions. It has been shown that *Xist* directly interacts with the Lamin B receptor in mice, recruiting the inactive X to the nuclear lamina and facilitating effective gene silencing[113]. The restoration of H3K27me3 domains in Lamin B1-binding regions in our capacitated cells might indicate gene inactivation through XIST-dependent recruitment of the X chromosome to the nuclear lamina.

However, the role of the Lamin B receptor in X chromosome silencing is still subject to debate, and the results presented here do not provide a definitive answer[119,120]. Additional experimental evidence will be necessary to determine the functional role of Lamin B1 in X chromosome silencing of capacitated and conventional hPSCs. Thus, the epigenetic state of the X chromosome seems to be more faithfully recapitulated in capacitated cells than in conventional primed female hPSCs.

Long-term *in vitro* culturing is often associated with the accumulation of epigenetic aberrations[32]. For example, naïve hPSCs frequently lose DNA methylation in imprinted regions, as we observed. Here, we explored another case of epigenetic regulation, X chromosome inactivation in female cells. X chromosome erosion commonly occurs *in vitro* cultured primed female hPSCs and represents a limitation for using female hPSCs in translational and fundamental research[9]. Strikingly, in this study resetting conventional hPSCs followed by capacitation reversed the eroded state and recovered the expected epigenetic characteristics of the inactive X chromosome. This further confirms that capacitated hPSCs represent not only transcriptional features of embryonic epiblast but also mimic some epigenetic characteristics more faithfully than the conventional primed cells. Importantly, resetting followed by capacitation represents a useful method to reduce hPSCs and iPSC epigenetic variability and study X chromosome inactivation.

To conclude, we have generated and analysed a comprehensive resource profiling global active and repressive histone modifications, DNA methylation, chromatin accessibility, and transcription in both pluripotency states. However, this study has some limitations that need to be addressed in future research. Firstly, the results presented in this dissertation are primarily based on sequencing data and have not been experimentally validated for their biological function. For instance, the observed genome-wide decrease in H3K27me3 levels may have minimal impact on the transition from naïve to primed pluripotency. Instead, this shift in epigenetic levels might be crucial for establishing appropriate conditions for differentiation and lineage specification. Conducting experiments using PCR2 inhibitors or other histone-modifying enzymes during hPSCs capacitation and differentiation would provide insight into the role of H3K27me3 during this process. Additionally, in terms of gene imprinting, we were unable to determine whether the loss of allele-specific methylation observed in H9-capacitated cells was a result of the resetting process or a consequence of naïve culture. This was due to the unavailability of single nucleotide polymorphisms (SNPs) data and the lack of epigenetic data after capacitation for the HNES1 cells. Future research should address whether capacitation affects allele-specific methylation in imprinting control regions. Regarding X chromosome inactivation, the conclusions presented in this dissertation were primarily based on epigenetic sequencing results. The assumption that both X chromosomes are active in the naïve state, that one of the X chromosomes becomes inactive in cR-H9 cells during capacitation, and that X chromosome inactivation becomes eroded in conventional H9 hPSCs were derived from the distribution of H3K27me3 and H3K9me3 reads mapped to a reference human X chromosome, as well as the expression of XIST and XACT lncRNAs. Although these are strong indications supporting the conclusions, it is still necessary to demonstrate the overall X chromosome expression in an allele-specific manner. Additionally, it would be important to provide RT-qPCR results for the lncRNAs XIST and XACT. Furthermore, conducting Fluorescence In Situ Hybridization (FISH) and immunofluorescence experiments to detect the number of XIST, XACT, and H3K27me3 loci in conventional hPSCs and hPSCs after resetting and recapitulation would be valuable.

Despite the limitations of this study, I believe that the analysis presented here will be a valuable resource for future research endeavours aimed at characterizing pluripotency and unravelling the molecular mechanisms underlying human embryogenesis. However, additional experiments are necessary to establish the causal relationship between epigenetics and pluripotency. Further research should focus on investigating enhancer-promoter interactions, perturbing epigenetic marks, exploring cell heterogeneity, examining the role of retrotransposons, and studying putative intermediate pluripotent states. These efforts will significantly enhance our understanding of the epigenetic mechanisms operating during the early stages of human development.

4. Conclusion

In conclusion, this study provides valuable insights into the epigenetic dynamics between naïve and primed pluripotency in human cells. It offers a comprehensive resource of sequencing data from multiple epigenetic modifications and chromatin accessibility that will serve as an essential guide for future research in human embryonic development and pluripotent stem cells. However, the causal relationship between epigenetics and pluripotency still requires further exploration. To deepen our understanding of epigenetic mechanisms in the early stages of human development, future research could focus on the findings highlighted in this dissertation, such as the differential role of enhancers between pluripotency states, the impact of cell heterogeneity, the function of retrotransposons, and allele-specific epigenetic differences.

Our findings reveal a clear correlation between H3K27ac and naïve-specific gene expression. To further investigate this association, it would be beneficial to explore the metabolism and signalling pathways specific to human pre-implantation epiblasts. This knowledge could assist in optimizing the conditions of naïve medium, leading to improved hPSCs for research or clinical purposes. Furthermore, our results suggest that primed hPSCs derived from transitioned naïve hPSCs may experience fewer aberrations in X chromosome inactivation due to the restoration of an eroded epigenetic landscape in capacitated cells. Therefore, capacitated cells may be a more appropriate alternative than conventionally cultured hPSCs for research purposes. Finally, the observed epigenetic differences between naïve and primed states, particularly the overall decrease in H3K27me3 and the localized enrichment in CpG islands, could have an impact on differentiation outcomes. Additional research into the function of H3K27me3 and associated proteins during capacitation could provide a better understanding of their influence on the differentiation of hPSCs.

This study characterized the epigenetic dynamics in two different pluripotent states and provided valuable insights into the molecular mechanisms that govern cell fate determination. This knowledge could improve the efficiency and safety of generating functional tissues and organs from pluripotent stem cells. Furthermore, abnormal epigenetic regulation has been linked to various diseases, such as cancer and developmental disorders. Therefore, the findings presented in this dissertation, along with the extensive sequencing data resource, can be utilized to model disease states and identify new targets for therapeutic intervention.

5. Outlook

During the writing of this dissertation, we have decided to perform additional experiments for the review process before publication. First, we will karyotype the cells in all conditions to determine if a correct and full set of chromosomes is present in the cell sample. Additionally, since we have observed a significant decrease in H3K27me3 during the transition from naïve to primed pluripotency, we plan to use a polycomb repressive complex 2 (PRC2) inhibitor in naïve and primed hPSCs to observe how the cells respond. PRC2 is a chromatin-modifying enzyme that catalyses the methylation of histone H3 at lysine 27 (H3K27me1/2/3). Furthermore, we will test the effects of PRC2 inhibition during capacitation and, in a separate experiment, use an MLL1 inhibitor, a histone H3K4 methyltransferase, in naïve and primed hPSCs. Finally, we will use fluorescence in situ hybridization (FISH) to detect the presence of the lncRNA XIST and immunostaining to detect the modified histone H3K27me3 in the X chromosome of the cR-H9 cells in all conditions. These experiments will help us to support the sequencing results and possibly guide future research.

To expand on the research findings, future investigations could conduct additional experiments using either the capacitation medium employed in this study or an improved version. For example, researchers could generate epigenetic sequencing data at multiple time points during the transitions between the naïve and primed states. In this study, we conducted bulk sequencing; however, single-cell sequencing could provide insights into cell heterogeneity during capacitation. Researchers can also evaluate the potential influence of the histone modification H3K27ac on naïve-specific expression by using transcription coactivator inhibitors such as CREB binding protein (CBP) and p300, which are two histone 3 lysine 27 acetyltransferases. Additionally, it would be interesting to compare the differential effects of capacitated and conventionally primed human pluripotent stem cells on the differentiation efficiency for each of the three primary cell layers. This comparison could be combined with highly specific small-molecule inhibitors that affect epigenetic modifications or lineage specification competency.

6. Material and methods

6.1. Cell culture

The methods from this section were conducted by collaborators from the Austin Smith laboratory (University of Exeter, Exeter, United Kingdom) and from the Wolf Reik laboratory (Babraham Institute, Cambridge, United Kingdom).

6.1.1. Cell lines

The experiments were conducted using the embryo-derived HNES1 hPSC (human pluripotent stem cell) cell line, the chemically reset cR-H9-EOS naïve hPSC line[20,98], and the H9 hPSC cell line (WA09) obtained from the WiCell Research Institute, Inc (Madison, WI, USA)[20]. The HNES1 cells were derived with informed consent under licence from the Human Embryology and Fertilisation Authority (HEFA; United Kingdom independent regulator of fertility treatment and research using human embryos).

6.1.2. Human pluripotent stem cell maintenance

Naïve hPSCs were cultured on irradiated mouse embryonic fibroblasts (MEFs) in PDLGX medium. The medium was prepared as follows: N2B27 basal medium supplemented with 1 μM PD032590, 10 ng/ml human LIF (both from Cambridge Stem Cell Institute facility), 2 μM Gö6983 (Tocris Bio-Techne, Cat. 2285), and 2 μM XAV939 (Tocris Bio-Techne, Cat. 3748), as described previously[10,121]. N2B27 basal medium was prepared as follows: Neurobasal (Cat. 21103049, ThermoFisher Scientific) and DMEM/F12 (Cat. 31331093, ThermoFisher Scientific) in the ratio 1:1, 0.5% N2 (Cat. 17202048, ThermoFisher Scientific), 1% B27 (Cat. 17504044, ThermoFisher Scientific), 2 mM L-glutamine (Cat. 25030024, ThermoFisher Scientific), 100 μM 2-mercaptoethanol (Cat. M7522, Sigma-Aldrich). Naïve hPSCs were routinely passaged using TrypLE Express (Cat. 12604021, ThermoFisher Scientific), and 0.5 μl/ml Geltrex (A1413302, ThermoFisher Scientific) was added to the culture medium during re-plating. 10 μM ROCK inhibitor (Y-27632, Cat. 688000, Millipore) was added for 24 hours after passaging. H9 hPSCs were cultured in E8 medium (prepared in-house according to Chen, G. et al., 2011[31]) on Geltrex pre-coated plates and passaged using 0.5 mM EDTA in PBS. All cells were cultured in a humidified incubator with 5% O_2 and 5% CO_2 at 37°C.

6.1.3. Capacitation

Capacitation was done as described previously[121,122]. Before capacitation, naïve hPSCs were passaged once to non-coated tissue culture plates in PDLGX medium supplemented Geltrex at 1 μl/cm^2 to reduce the number of feeder cells. For capacitation, cells were dissociated with TrypLE and plated to Geltrex-coated tissue culture plates at a seeding density of 1.6x10^4/cm^2 in PDLGX supplemented with 10 μM ROCK inhibitor. After 48 hours, cells were washed with DMEM/F12 (Cat. 31331093, ThermoFisher Scientific) supplemented with 0.1% BSA and the medium was changed to N2B27 supplemented with 2 μM XAV939 (Tocris Bio-Techne, Cat. 3748). The medium was refreshed every 1-2 days. Cells were passaged at a 1:2 ratio at confluency using TrypLE and 10 μM ROCK inhibitor.

For expansion after 10 days of capacitation, cells were cultured in either E8 or N2B27 supplemented with 2 μM XAV939, 3 ng/ml Activin A and 10 ng/ml FGF2 (XAF medium, modified from Sumi, Tomoyuki, et al. 2013[123]). During expansion, cells were cultured on Geltrex pre-coated tissue culture plates and passaged by dissociation with either 0.5 mM EDTA or TrypLE. 10 μM ROCK inhibitor was added for 24 hours after passaging.

6.2. Data generation

The methods from this section were conducted by collaborators from the Austin Smith and Wolf Reik laboratories, with the help of Maike Paramor and Vicki Murray from the Stem Cell Institute Genomics Facility. Western blots were performed by Chee-Wai Wong.

6.2.1. RNA sequencing (RNA-seq) library preparation

The RNA-seq library preparation was described in a previously published article by our collaborators[10].

6.2.2. Chromatin Immunoprecipitation (ChIP) and sequencing (ChIP-seq)

5x10^5 to 10^6 cells were used for ChIP per sample. Cross-linking was done with 1% formaldehyde in DMEM added directly to cells in culture dishes for 8 minutes at room temperature. Quenching was performed using 0.1 M glycine (final concentration). The cells were washed with ice-cold PBS and scrapped with PBS with a protease inhibitor cocktail (Cat. 11697498001, Roche). Lysis was performed in LB1 buffer (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% Igepal CA-630, 0.25% Triton X100, protease inhibitors) by rotating for 10 min at +4°C, followed by centrifugation at 2000 x g for 5 minutes. The pellet was incubated in LB2 buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, protease inhibitors) by rotating for 5 minutes at +4°C, followed by centrifugation at 2000 x g for 5 minutes. The pellet was resuspended in a buffer containing 50 mM Tris-HCl pH 8.0, 10 mM EDTA, and 1% SDS. The sonication was performed for 45 cycles, 30-second intervals on/off. Debris were removed by centrifugation, and 10% of the sonicated material was saved as input.

Before immunoprecipitation, 5 μg antibody and 100 μl Protein A beads were pre-incubated overnight at +4°C in a total volume of 250 μl, adjusted with PBS with 5 mg/ml BSA, followed by three washes using the same solution. Immunoprecipitation was done by combining the sonicated material with the bead-antibody complexes, in a total volume of 400 μl adjusted with ChIP dilution buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine, 1% Triton X100, protease inhibitors), by rotating overnight at +4°C. After the incubation, the beads were washed 6 times with RIPA buffer (50 mM HEPES-KOH pH 7.5, 500 mM LiCl, 1 mM EDTA, 1% Igepal CA-630, 0.7% sodium deoxycholate) followed by one wash in TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Reverse-crosslinking was done in elution buffer containing 1% SDS, 100 mM sodium bicarbonate, and 200 mM NaCl at 65°C for 9 to 15 hours. The input sample was also reverse-crosslinked. After this step, the beads were removed from the mixture, and the remaining DNA in the solution was treated with 8 μg RNAse A for 1 hour at 37°C, followed by 80 μg proteinase K for 2 hours at 55°C. DNA was purified using MinElute columns and then used for ChIP-Seq library preparation using the NEXTflex Rapid DNA-Seq Kit (NOVA-5144-02) according to the manufacturer's protocol with some minor modifications. Briefly, we amplified the final libraries with 4 cycles of PCR, then performed an AMPure XP-based size selection (0.5x to eliminate larger fragments followed by 1.8x to extract the smaller fragments), and then continued with 8 more PCR cycles on the size selected material. The final libraries were then cleaned up with 0.8x vol AMPure XP beads. Sequencing was carried out on HiSeq 2500 instruments (Illumina).

6.2.3. Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) library preparation

To isolate nuclei, $5x10^4$ cells were resuspended in 50 μl cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Igepal C-630), the mixture was pipetted up and down 16 times and immediately centrifuged at 500 x g for 10 minutes at +4°C. The pellet was used for tagmentation reaction, which was performed with 2.5 μl TDE1 in 50 μl total volume (FC-121-1031, Nextera DNA Library Prep Kit, Illumina). The samples were then further amplified (Nextera kit FC-121-1030) and barcoded with the corresponding indices (FC-121-1011). The final libraries were then cleaned up with 1.2x vol AMPure XP beads. Sequencing was carried out on NovaSeq 6000 instruments (Illumina).

6.2.4. Whole-Genome Bisulfite Sequencing (WGBS)

Post-bisulfite adaptor-tagging (PBAT) libraries for whole-genome DNA methylation analysis were prepared from purified genomic DNA as previously described[124–126]. Paired-end sequencing was carried out on HiSeq 2500 instruments (Illumina).

6.2.5. Western blot

Histone protein extracts (Abcam Protocol) were resolved using a gradient SDS-PAGE before being immunoblotted onto a PVDF membrane. The membrane was blocked for 1 hour in blocking solution (TBS/0.1% Tween/5% BSA) and then incubated overnight at 4°C with primary antibodies diluted in blocking solution. After washes with TBS-0.1% Tween, the membranes were incubated with secondary antibodies diluted in blocking buffer for 1 hour at room temperature. The HRP-conjugates were detected using ECL solution–which generates chemiluminescence. The following were the primary antibodies used: H3 (Abcam, #ab1791, 1:1000), H3K4me1 (Abcam, #ab8895, 1:1000), H3K4me3 (Abcam, #ab8580, 1:1000), H3K27me3 (Cell Signalling, #C36B11, 1:1000), H3K27ac (Abcam, #ab4729, 1:1000), and H3K9me3 (Abcam, #ab8898, 1:1000). The secondary antibody used was an HRP-conjugated monoclonal donkey anti-rabbit IgG (Amersham, #NA934, 1:5000).

6.3. Data processing

The methods from this section were conducted by me. Simon Andrews and Felix Krüger from the bioinformatics facility at the Babraham Institute (Cambridge, United Kingdom) helped me with data mapping and handling.

6.3.1. Genome build and annotation

Sequencing data raw reads from all assays were mapped to the GRCh38 primary assembly downloaded from Ensembl (release 98; download link: https://ftp.ensembl.org/pub/release-98/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz; link copied on the 7th of December 2022). The chromosome lengths used during the analysis were obtained from the GRCh38 primary assembly FASTA file.

Genomic features were annotated using the Homo sapiens Ensembl gene annotation release 98 (download link:

https://ftp.ensembl.org/pub/release-98/gtf/homo_sapiens/Homo_sapiens.GRCh38.98.gtf.gz; link copied on the 7th of December 2022) and the annotatr package from Bioconductor[127]. Repetitive and centromeric regions were downloaded from the UCSC (University of California Santa Cruz) Table Browser (clade: “Mammal”, genome: “Human”, assembly: “Dec. 2013

(GRCh38/hg38)”). The repetitive regions were obtained from the track “RepeatMasker” and table “rmsk” and centromeric regions from the track “Centromeres” and table “centromeres”. Human naïve and primed enhancers and super-enhancers regions were downloaded from Barakat et al., 2018[128] and converted from hg19 to hg38 using a UCSC Liftover chain (download link:

https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz, link copied on the 7th of December 2022). Human imprinted regions were previously defined in the Wolf Reik research group with assistance from the bioinformatics facility at the Babraham Institute (Table 2). The blacklisted regions used in the ChIP-seq and ATAC-seq alignment pipelines were the “hg38 ENCODE blacklist” version 2 downloaded from the Boyle lab[129].

6.3.2. Read alignment and quantification

The reads from each assay were aligned and quantified employing the community-curated nf-core pipelines[130]. The single-end RNA-seq data were processed using the ‘rnaseq’ pipeline (version 1.4.2) with the pseudo-aligner Salmon[131] selected; the ChIP-seq data (histone modifications H3K4me1, H3K4me3, H3K9me3, H3K27me3, and H3K27ac) were processed using the ‘chipseq’ pipeline (version 1.1.0) with BWA aligner[132] and MACS2 peak-caller[133] being chosen as default. Histone modifications H3K4me3 and H3K27ac peaks were called in the MACS2 ‘narrowPeak’ mode while the histone modifications H3K4me1, H3K27me3, and H3K9me3 peaks were called in the MACS2 ‘broadPeak’ mode. The input control used for peak calling in all samples was the merged purified DNA sequencing data from cR-H9-EOS cells at day 0, cR-H9-EOS cells at day 10, and the conventionally cultured (E8 medium) H9 hPSCs (each condition with two biological replicates). The ATAC-seq data was processed using the ‘atacseq’ pipeline (version 1.1.0) with the BWA aligner and MAC2 peak-caller selected as default. The ATAC-seq peaks were called in the MACS2 narrowPeak mode. The PBAT data was processed using the ‘methylseq’ pipeline (version 1.4), with Bismark[134] selected for mapping and methylation calling. The complete command line code used to run the pipelines can be found in the code availability section (Appendix C).

6.3.3. RNA-seq gene read counts

The RNA-seq read counts obtained from the Salmon transcript quantification were imported into R and converted to gene read counts using the tximport package[135]. The gene read counts were subsequentially transformed using the variance stabilizing transformation from DESeq2[136] for the correlation and clustering plots. For the remaining plots, the reads were normalized by estimating the size factors and retrieving the normalized counts with DESeq2. For the differentially expressed gene analysis, raw gene read counts were used.

6.3.4. ChIP-seq and ATAC-seq read counts in genomic regions

The ChIP-seq and ATAC-seq aligned read data were imported into R and the number of reads that overlap genomic regions (i.e., genomic windows and features) was counted with the function ‘summarizeOverlaps’ from the Bioconductor package GenomicAlignments[137] with the parameter inter.feature set as ‘FALSE’, the parameter ignore.strand set as ‘TRUE’ and the remaining parameters kept as default. The read counts were transformed using the variance stabilizing transformation from DESeq2 and used as input for the correlation plots. For the remaining plots, the DESeq2 normalized counts were used instead.

6.3.5. Methylation percentage in CpGs and genomic regions

While running the nf-core methylseq pipeline, Bismark generated a coverage file with the post-bisulfite adaptor-tagging (PBAT) sequencing read information for each cytosine. Those files were imported into R and, with the methylKit package[138], only read count data in a CpG context were selected. Also, to exclude cytosines with an extremely high number of read counts, positions belonging to the top 0.1 percentile of read counts were excluded. The normalized coverage values between samples were calculated by a scaling factor derived from differences between the median of coverage distributions using the ‘normalizeCoverage’ function from methylKit. Because several samples had low genome-wide coverage, the reads from all replicates in each CpG position were pooled, and only the CpGs with three or more pooled reads were selected. The methylation percentage in each CpGs was calculated by dividing the number of Cs (cytosines) by the CpG coverage and the result was multiplied by 100. The read counts over genomic regions were determined using the function ‘regionCounts’ function from methylKit. The function filtered out regions with less than 2 bases covered or regions with an overall coverage of fewer than 3 reads. The methylation percentage of those genomic regions was also calculated by pooling the number of Cs and dividing that number by the CpG coverage and the result was multiplied by 100.

6.3.6. Differentially expressed genes and clusters

The RNA-seq raw read counts were used to compute the differentially expressed genes using the package DESeq2[136] with the design formula having a time-point as the only variable (i.e., “~ day”). The differential genes were obtained for each time point and cell type. After computing the differentially expressed genes for each condition, only genes with an adjusted p-value below 0.05 and a log2 fold change of more than 2 were selected. The differentially expressed genes between every condition combination were clustered using the package Mfuzz[139] with 5 cluster centres and 0.8 as the minimum membership.

6.3.7. Differential ChIP-seq and ATAC-seq peaks

The peaks differentially enriched between conditions were determined using the R Bioconductor package DiffBind[140]. Peaks from ATAC-seq, H3K27ac and H3K4me3 ChIP-seq data were re-centred around the consensus summit and resized to have 200 base pairs upstream and downstream of the centre. Peaks from H3K27me3, H3K4me1 and H3K9me3 ChIP-seq data were re-centred around a consensus summit and resized to have 500 base pairs upstream and downstream of the centre. During the DiffBind analysis, no blacklist or greylist was applied to the DBA object.

6.4. Data analysis

The methods from this section were conducted by me. Simon Andrews and Felix Krüger from the bioinformatics facility at the Babraham Institute (Cambridge, United Kingdom) provided advice on how to analyse multi-omics data and Jason Ernst from UCLA (University of California, Los Angeles) helped me use the tool chromHMM[141].

6.4.1. Principal component analysis and hierarchical clustering

The principal component analysis and hierarchical clustering plots were generated using: the differentially expressed genes from the RNA-seq data; the differential peaks from the ATAC-seq and ChIP-seq analysis with an FDR (false discovery rate) of less than 0.05 and an absolute fold change of more than 2; and the top 5% most variable 200 CpG-containing genome-wide windows for the methylation data. The 200 CpG-containing genome-wide windows were generated with the software SeqMonk[142].

6.4.2. Multi-omics factor analysis (MOFA)

MOFA2[95] was used to extract factors that are responsible for sample clustering using all data assays combined. The MOFA2 model was constructed using the same input as the principal component analysis and hierarchical clustering: the differentially expressed genes from the RNA-seq data; the differential peaks from the ATAC-seq and ChIP-seq analysis with an FDR (false discovery rate) of less than 0.05 and an absolute fold change of more than 2; and the top 5% most variable 200 CpG-containing genome-wide windows for the methylation data. The model was restricted to 3 factors and the training to 2000 maximal interactions in the "slow" convergence mode. The t-distributed stochastic neighbour embedding (t-SNE) plot was generated using the model output with a perplexity equal to 4.

6.4.3. Chromatin state discovery

Two ChromHMM models were created to discover chromatin states using the histone modifications and chromatin accessibility data (ChIP-seq and ATAC-seq). One model had the ‘concatenated’ design and the other model had the ‘stacked’ design (see Introduction). In both model types, the binary files were generated using the function ‘BinarizeBam’ from the bam files obtained after read alignment. Afterwards, the function ‘LearnModel’ was used to learn the model, and the function ‘CompareModels’ to compare models with multiple states (40 states for the ‘concatenated’ design and 100 states for the ‘stacked’ design). Finally, the state segments of the model were exported using the function ‘MakeSegmentation’. The length of the segments was kept as default, i.e., 200 base pairs. Those state segments were then imported to R and used for overlap enrichment with genomic features (see Appendix C to have access to the script and design table that generated the model).

6.4.4. Enriched genomic regions

The R Bioconductor package LOLA[109] was used to compare the enrichment of selected genomic regions with public databases of annotated regions. LOLA core and extended hg38 databases were used in the analysis. Additionally, a custom database was created using the genomic annotated regions from the Homo sapiens Ensembl gene annotation release 98 and the R Bioconductor annotatr package[127]. It was also included to that custom database the following features: the regions generated from DiffBind on the ChIP-seq and ATAC-seq data, each human chromosome, repeat classes and families, CpG island types, promoter GC skew types, and RNA-seq clusters. Only significant results, with an FDR adjusted p-value of less than 0.05, were selected.

6.4.5. Multivariate regression analysis

For the regression analysis, the observed over expected ratio (O/E ratio) of the number of CpGs in all protein-coding gene promoters (defined as transcription start site, TSS, +- 2 kilobases) was first calculated using a custom script applying the function ‘oligonucleotideFrequency’ from the Bioconductor package BSgenome[143]. The promoters were then separated into three categories: promoters with high CpG density (O/E ratio > 0.7); promoters with medium CpG density (O/E ratio >= 0.35 and O/E ratio <= 0.7); and promoters with low CpG density (O/E ratio < 0.35). Then the R function ‘boot’[144] was used with the linear formula:

$$`RNA-seq' \sim 'ATAC-seq' + H3K27ac + H3K27me3 + H3K4me1 + H3K4me3 + H3K9me3 + Methylation$$

and with 1000 bootstrap replicates. For the RNA-seq data, the gene expression logarithm (log) normalized counts were used and for the ATAC-seq and ChIP-seq data, it was used the log normalized counts that overlapped promoters (with at least 1 base pair overlap). For the methylation data, the methylation ratio inside the promoter region was calculated as previously described for genomic windows (see methods subsection "Methylation percentage in CpGs and genomic regions"). All datasets were re-scale to have values between 0 and 1 before the regression analysis. The relative importance of regressors was calculated using the function 'boot.relimp' from the R package relaimpo[145]. For this analysis, only protein-coding genes were used.

6.4.6. Gene function profiling

The gene ontology (GO) analysis was generated using the function 'enrichGO' from the Bioconductor package clusterProfiler[146]. In all GO analyses: the database org.Hs.eg.db[147] was used; only 'biological process' was selected as the sub ontologies; the adjusted p-value procedures used was the Benjamini-Hochberg; and the q-value had a cutoff of 0.01. The results were simplified with the function 'simplify' from the clusterProfiler package to avoid term redundancy.

The Kyoto encyclopedia of genes and genomes (KEGG) pathway analysis was generated using the 'enrichKEGG' function, also from the Bioconductor package clusterProfiler. In all the KEGG analyses: the organism selected was "hsa"; the adjusted p-value procedures used was the Benjamini-Hochberg; and the q-value had a cutoff of 0.01.

6.4.7. Bivalency and developmental genes

Bivalent promoters were defined by selecting only the promoters (transcription start site, TSS, +- 2 kilobases) containing both H3K27me3 and H3K4me3 MACS2 peaks (from the ChIP-seq dataset) using the function 'subsetByOverlaps' from the Bioconductor package IRanges with parameters kept as default (with at least 1 base pair overlap). The list of developmental genes analysed and categorized as 'naïve pluripotency', 'general pluripotency', 'post-implantation epiblast', and 'lineage markers'-associated genes were extracted from Rostovskaya et al., 2019[10] (see gene list in Table 3).

6.4.8. Imprinting control regions allelic analysis

The human imprinted control regions (ICRs) used in this analysis were previously defined in the Wolf Reik research group with assistance from the bioinformatics facility at the Babraham Institute (see the ICRs list in Table 2). To perform the allelic differential analysis, the single nucleotide polymorphisms (SNPs) from the H9 cell line (provided by the Babraham Institute bioinformatics facility) were used to assign the sequencing reads to each allele using the software SNPsplit[148]. The reads assigned to each allele from the PBAT (with DNA methylation data), ChIP-seq (with H3K4me3, H3K4me1, H3K27ac, H3K27me3, and H3K9me3 data), and ATAC-seq assays were then imported into the software SeqMonk[142]. The reads overlapping the imprinted control regions were counted in each allele and the Bioconductor package EdgeR was used to perform a statistical significance test between allele counts in each ICR. A p-value cut-off of 0.05 was used. The results were then converted from hg19 to hg38 using a UCSC Liftover chain (download link:

https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz, link copied on the 7th of December 2022).

6.4.9. Motif enrichment

Motif enrichment analysis was performed using the software MEME-ChIP[149] from the MEME Suite[150] in the differential enrichment mode. The sequence alphabet was defined as "DNA, RNA or Protein", the primary sequences were the significant MACS2 peaks, the control sequences were the non-significant MACS2 peaks, and the database used was the JASPAR2022[151] CORE vertebrates non-redundant. The remaining options were kept as default.

6.4.10. Transposable elements expression

The expression and differential expression of transposable element (TE) subfamilies were performed using the package SalmonTE[152] using 'hs' (homo sapiens) as reference and, for the differential analysis, the analysis type selected was 'DE'. The remaining parameters were kept as default. The RNA-seq fastq output files were used as input.

7. Supplemental figures

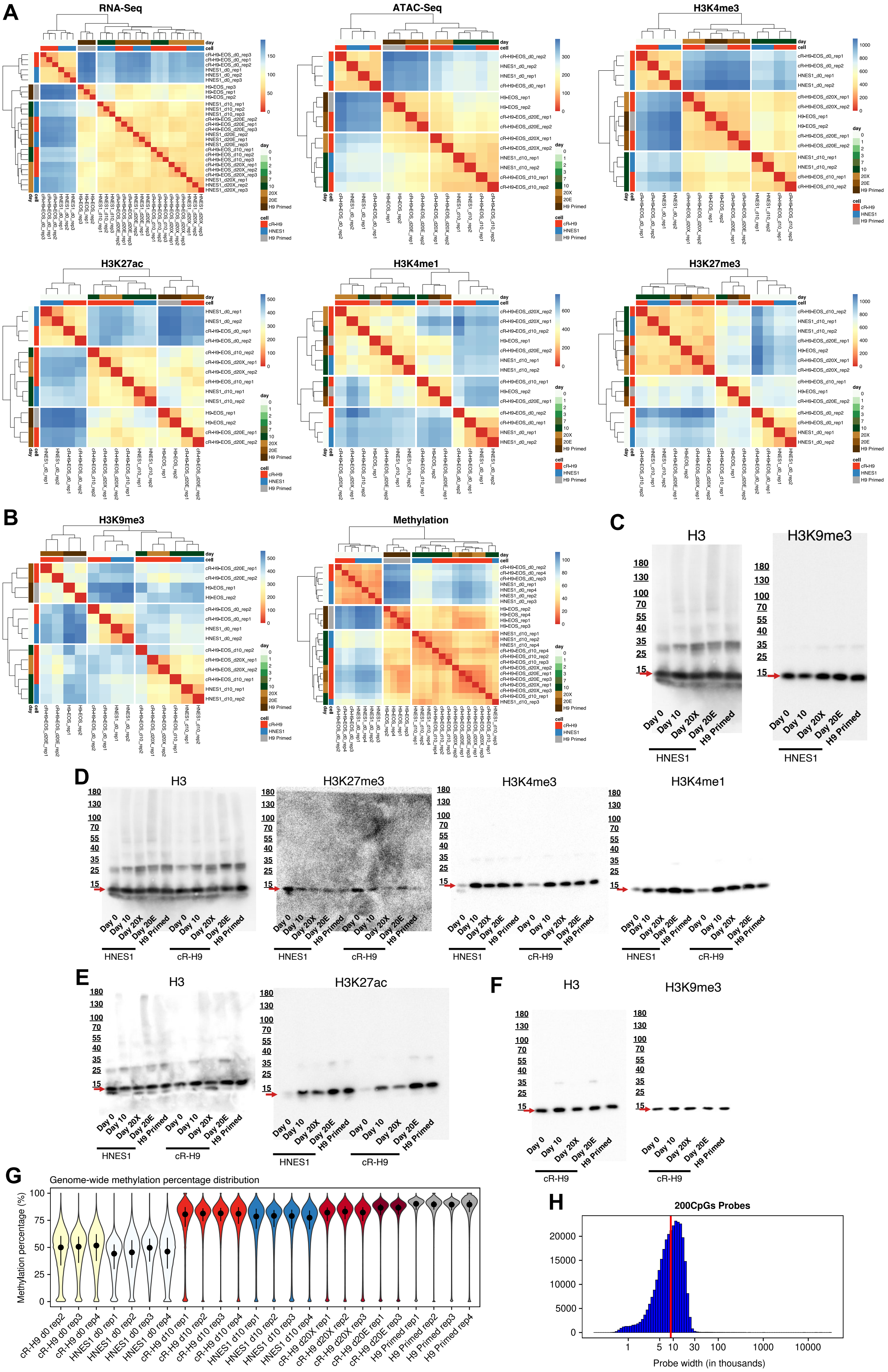
A
RNA-Seq
ATAC-Seq
H3K4me3
H3K27ac
H3K4me1
H3K27me3
day
cell
0
1
2
3
7
10
20X
20E
H9 Primed
cR-H9
HNES1
H9 Primed
B
H3K9me3
Methylation
C
H3
H3K9me3
180
130
100
70
55
40
35
25
15
Day 0
Day 10
Day 20X
Day 20E
H9 Primed
HNES1
D
H3
H3K27me3
H3K4me3
H3K4me1
HNES1
cR-H9
E
H3
H3K27ac
HNES1
cR-H9
F
H3
H3K9me3
cR-H9
G
Genome-wide methylation percentage distribution
Methylation percentage (%)
100
75
50
25
0
cR-H9 d0 rep2
cR-H9 d0 rep3
cR-H9 d0 rep4
HNES1 d0 rep1
HNES1 d0 rep2
HNES1 d0 rep3
HNES1 d0 rep4
cR-H9 d10 rep1
cR-H9 d10 rep2
cR-H9 d10 rep3
cR-H9 d10 rep4
HNES1 d10 rep1
HNES1 d10 rep2
HNES1 d10 rep3
HNES1 d10 rep4
cR-H9 d20X rep1
cR-H9 d20X rep2
cR-H9 d20X rep3
cR-H9 d20E rep1
cR-H9 d20E rep3
H9 Primed rep1
H9 Primed rep2
H9 Primed rep3
H9 Primed rep4
H
200CpGs Probes
20000
15000
10000
5000
0
1
5
10
30
100
1000
10000
Probe width (in thousands)

Figure S1.

A) and **B)** Hierarchical clustering based on Euclidean distance between sample replicates for each sequencing assay. The figures were generated using the top 500 most variable differentially expressed genes, histone modifications ChIP-seq differential peaks, ATAC-seq differential peaks, and DNA methylation average in 200 CpG-containing genomic windows between all conditions. Western blot results with antibodies recognising: **C)** H3 and H3K9me3 in HNES1 and conventional H9 cells. **D)** H3, H3K27me3, H3K4me3, and H3K4me1; **E)** H3 and H3K27ac; **F)** H3 and H3K9me3 in cR-H9 and conventional H9 cells. **G)** Probe width distribution of the 200 CpG probes generated with SeqMonk using the PBAT reads from all samples. **H)** Genome-wide methylation percentage distribution in the 200 CpG probes from SeqMonk.

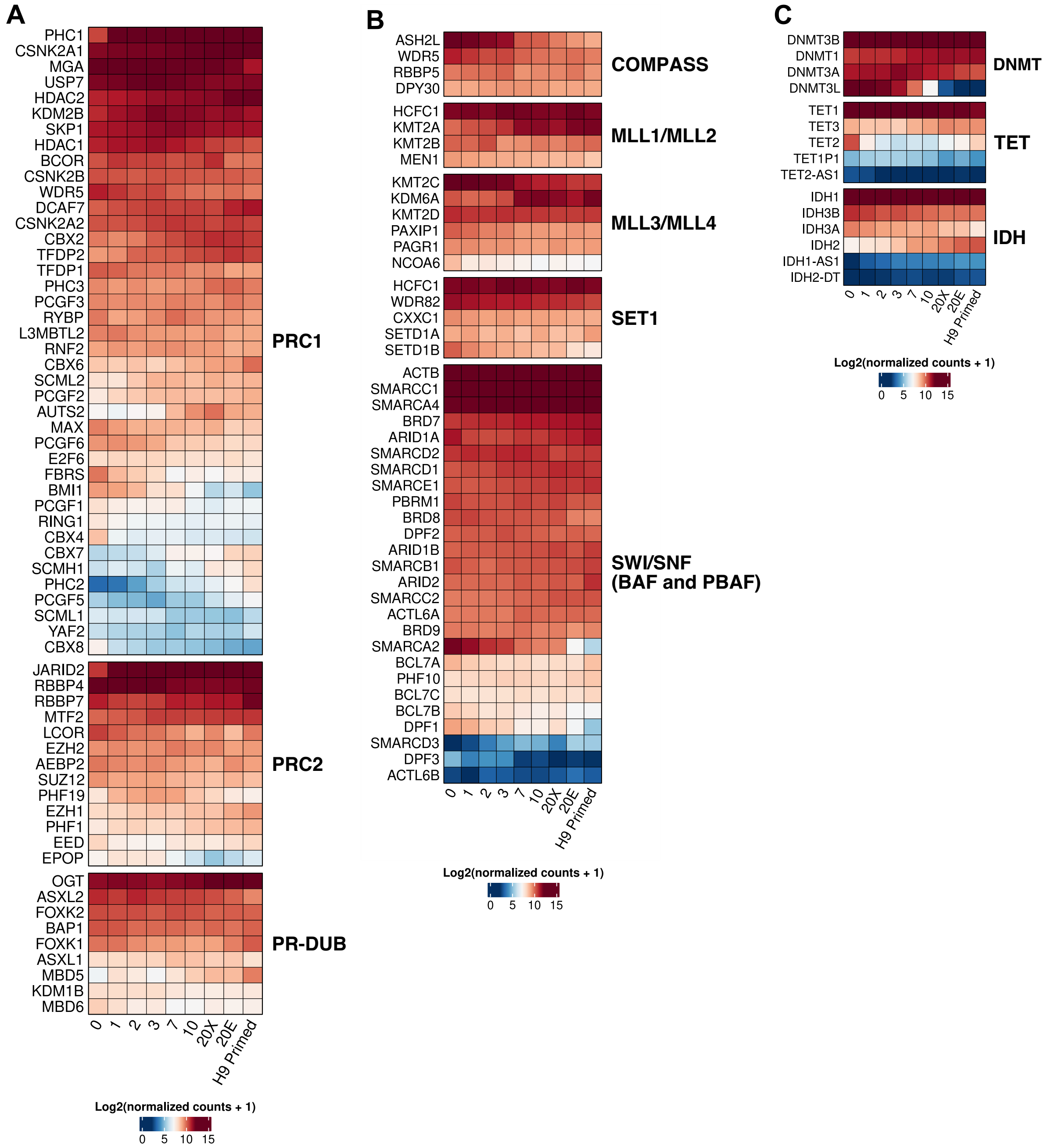

Figure S2.

Heatmaps with the log-normalized expression counts of genes to which their protein products belong to the: **A)** Polycomb-group proteins (PcG): PRC1, PRC1, and PR-DUB complex; **B)** Trithorax-group proteins (TrxG): COMPASS, MLL1/MLL2, MLL3/MLL4, SET1, and SWI/SNF complex; **C)** and the DNMT, TET, and IDH protein families.

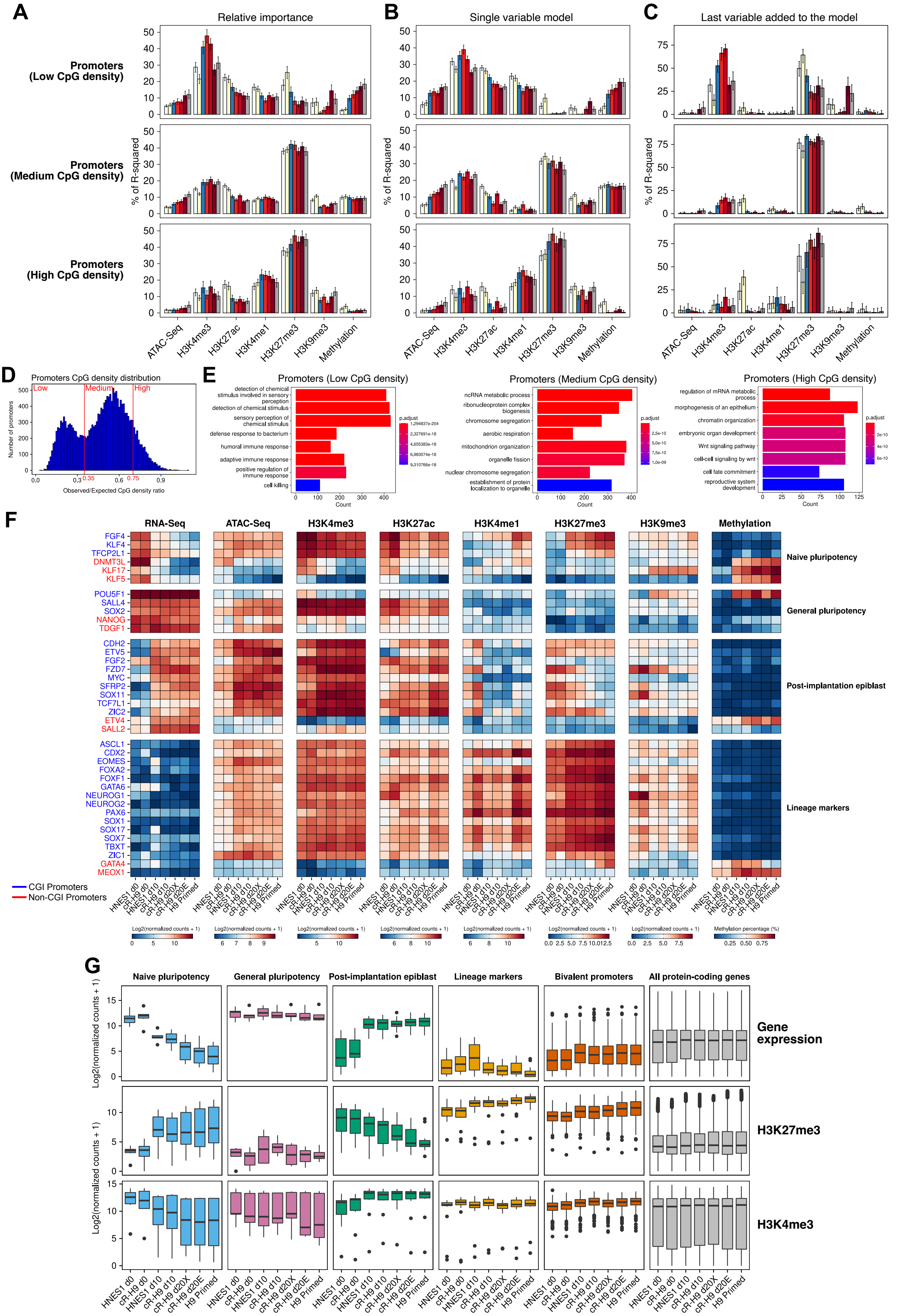
A
Relative importance
B
Single variable model
C
Last variable added to the model
Promoters (Low CpG density)
Promoters (Medium CpG density)
Promoters (High CpG density)
% of R-squared
ATAC-Seq
H3K4me3
H3K27ac
H3K4me1
H3K27me3
H3K9me3
Methylation
D
Promoters CpG density distribution
Low
Medium
High
Number of promoters
Observed/Expected CpG density ratio
0.35
0.75
E
Promoters (Low CpG density)
detection of chemical stimulus involved in sensory perception
detection of chemical stimulus
sensory perception of chemical stimulus
defense response to bacterium
humoral immune response
adaptive immune response
positive regulation of immune response
cell killing
Promoters (Medium CpG density)
ncRNA metabolic process
ribonucleoprotein complex biogenesis
chromosome segregation
aerobic respiration
mitochondrion organization
organelle fission
nuclear chromosome segregation
establishment of protein localization to organelle
Promoters (High CpG density)
regulation of mRNA metabolic process
morphogenesis of an epithelium
chromatin organization
embryonic organ development
Wnt signaling pathway
cell-cell signaling by wnt
cell fate commitment
reproductive system development
p.adjust
Count
F
RNA-Seq
ATAC-Seq
H3K4me3
H3K27ac
H3K4me1
H3K27me3
H3K9me3
Methylation
FGF4
KLF4
TFCP2L1
DNMT3L
KLF17
KLF5
Naive pluripotency
POU5F1
SALL4
SOX2
NANOG
TDGF1
General pluripotency
CDH2
ETV5
FGF2
FZD7
MYC
SFRP2
SOX11
TCF7L1
ZIC2
ETV4
SALL2
Post-implantation epiblast
ASCL1
CDX2
EOMES
FOXA2
FOXF1
GATA6
NEUROG1
NEUROG2
PAX6
SOX1
SOX17
SOX7
TBXT
ZIC1
GATA4
MEOX1
Lineage markers
CGI Promoters
Non-CGI Promoters
HNES1 d0
cR-H9 d0
HNES1 d10
cR-H9 d10
cR-H9 d20X
cR-H9 d20E
H9 Primed
Log2(normalized counts + 1)
Methylation percentage (%)
G
Naive pluripotency
General pluripotency
Post-implantation epiblast
Lineage markers
Bivalent promoters
All protein-coding genes
Log2(normalized counts + 1)
Gene expression
H3K27me3
H3K4me3

Figure S3.

A) The relative importance of each variable in the multivariate linear regression model that correlates gene expression with the promoter epigenetic modification levels for each condition, as measured by the 'lmg' metric from the R package relaimpo. B) The percentage of the multivariate linear regression R-squared explained by each variable when the model only uses a single variable. C) The percentage of R-squared explained by each variable when it is the last variable added to the multivariate linear regression model. D) The distribution of promoter numbers is shown based on their observed/expected CpG density ratio. E) Gene ontology terms for genes with promoters with low, medium, and high CpG density. F) Gene expression and promoter ATAC-seq and ChIP-seq log-transformed normalized counts, as well as the methylation percentage for genes associated with embryonic development and pluripotency in each condition. G) Gene expression and promoter H3K27me3 and H3K4me3 log-transformed normalized counts for genes associated with embryonic development and pluripotency, genes with bivalent promoters in all conditions, and all protein-coding genes.

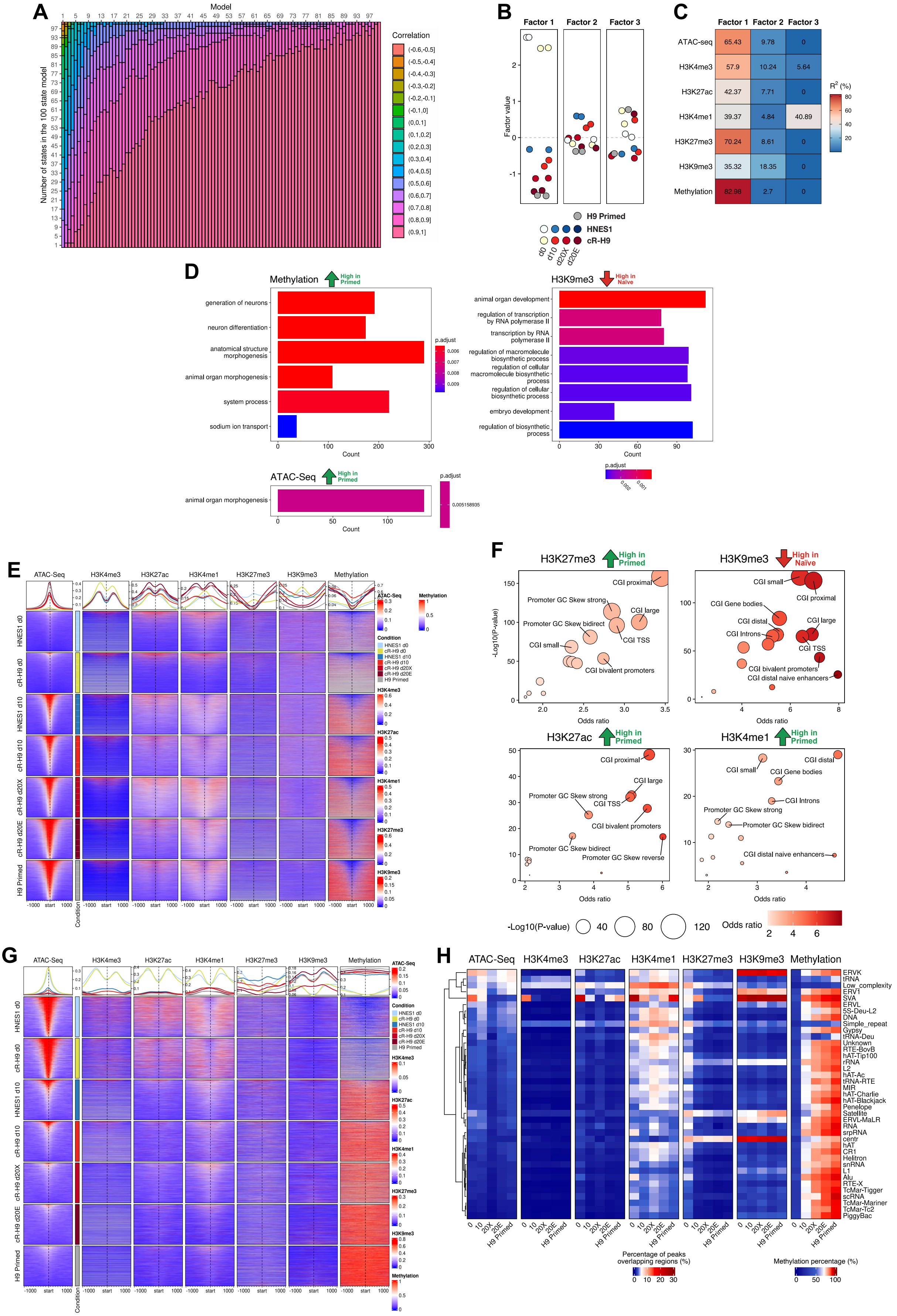
A
Model
Number of states in the 100 state model
Correlation
(-0.6,-0.5]
(-0.5,-0.4]
(-0.4,-0.3]
(-0.3,-0.2]
(-0.2,-0.1]
(-0.1,0]
(0,0.1]
(0.1,0.2]
(0.2,0.3]
(0.3,0.4]
(0.4,0.5]
(0.5,0.6]
(0.6,0.7]
(0.7,0.8]
(0.8,0.9]
(0.9,1]
B
Factor 1
Factor 2
Factor 3
Factor value
H9 Primed
HNES1
cR-H9
d0
d10
d20X
d20E
C
Factor 1 Factor 2 Factor 3
ATAC-seq 65.43 9.78 0
H3K4me3 57.9 10.24 5.64
H3K27ac 42.37 7.71 0
H3K4me1 39.37 4.84 40.89
H3K27me3 70.24 8.61 0
H3K9me3 35.32 18.35 0
Methylation 82.98 2.7 0
R² (%)
D
Methylation
High in Primed
generation of neurons
neuron differentiation
anatomical structure morphogenesis
animal organ morphogenesis
system process
sodium ion transport
Count
p.adjust
H3K9me3
High in Naïve
animal organ development
regulation of transcription by RNA polymerase II
transcription by RNA polymerase II
regulation of macromolecule biosynthetic process
regulation of cellular macromolecule biosynthetic process
regulation of cellular biosynthetic process
embryo development
regulation of biosynthetic process
ATAC-Seq
High in Primed
animal organ morphogenesis
E
ATAC-Seq
H3K4me3
H3K27ac
H3K4me1
H3K27me3
H3K9me3
Methylation
HNES1 d0
cR-H9 d0
HNES1 d10
cR-H9 d10
cR-H9 d20X
cR-H9 d20E
H9 Primed
Condition
-1000 start 1000
F
H3K27me3
High in Primed
H3K9me3
High in Naïve
H3K27ac
High in Primed
H3K4me1
High in Primed
CGI proximal
Promoter GC Skew strong
CGI large
Promoter GC Skew bidirect
CGI TSS
CGI small
CGI bivalent promoters
CGI Gene bodies
CGI distal
CGI Introns
CGI distal naive enhancers
Promoter GC Skew reverse
-Log10(P-value)
Odds ratio
-Log10(P-value) 40 80 120
Odds ratio 2 4 6
G
ATAC-Seq
H3K4me3
H3K27ac
H3K4me1
H3K27me3
H3K9me3
Methylation
Condition
H
ATAC-Seq H3K4me3 H3K27ac H3K4me1 H3K27me3 H3K9me3 Methylation
ERVK
tRNA
Low_complexity
ERV1
SVA
ERVL
5S-Deu-L2
DNA
Simple_repeat
Gypsy
tRNA-Deu
Unknown
RTE-BovB
hAT-Tip100
rRNA
L2
hAT-Ac
tRNA-RTE
MIR
hAT-Charlie
hAT-Blackjack
Penelope
Satellite
ERVL-MaLR
RNA
srpRNA
centr
hAT
CR1
Helitron
snRNA
L1
Alu
RTE-X
TcMar-Tigger
scRNA
TcMar-Mariner
TcMar-Tc2
PiggyBac
0 10 20X 20E H9 Primed
Percentage of peaks overlapping regions (%)
0 10 20 30
Methylation percentage (%)
0 50 100

Figure S4.

A) MOFA2 factor values for the three factors extracted by analysing histone modifications' ChIP-seq differential peaks, ATAC-seq differential peaks, and DNA methylation most variable regions between conditions. B) The variance explained (R-squared values) by each sequencing assay for the clustering samples in each MOFA2 factor. C) Gene ontology of the closest genes (less than 50 kb) to MOFA2 factor 1 top loading weight features. Only the factor 1 features that returned gene ontology terms are represented. D) Enrichment of factor 1 top loading weight features with CpG island and GC skew-associated genomic regions. Only the factor 1 features with significant overlap with CpG islands were selected. E) Epigenetic profile and heatmap of regions centred at the ATAC-seq peaks with high normalized counts in the primed state, selected by loading weight in factor 1. G) Epigenetic profile and heatmap of regions centred at the ATAC-seq peaks with high normalized counts in the naïve state, selected by loading weight in factor 1. H) Methylation percentage and percentage of ATAC-seq and ChIP-seq peaks overlapping repeat elements. All results were limited to autosomes.

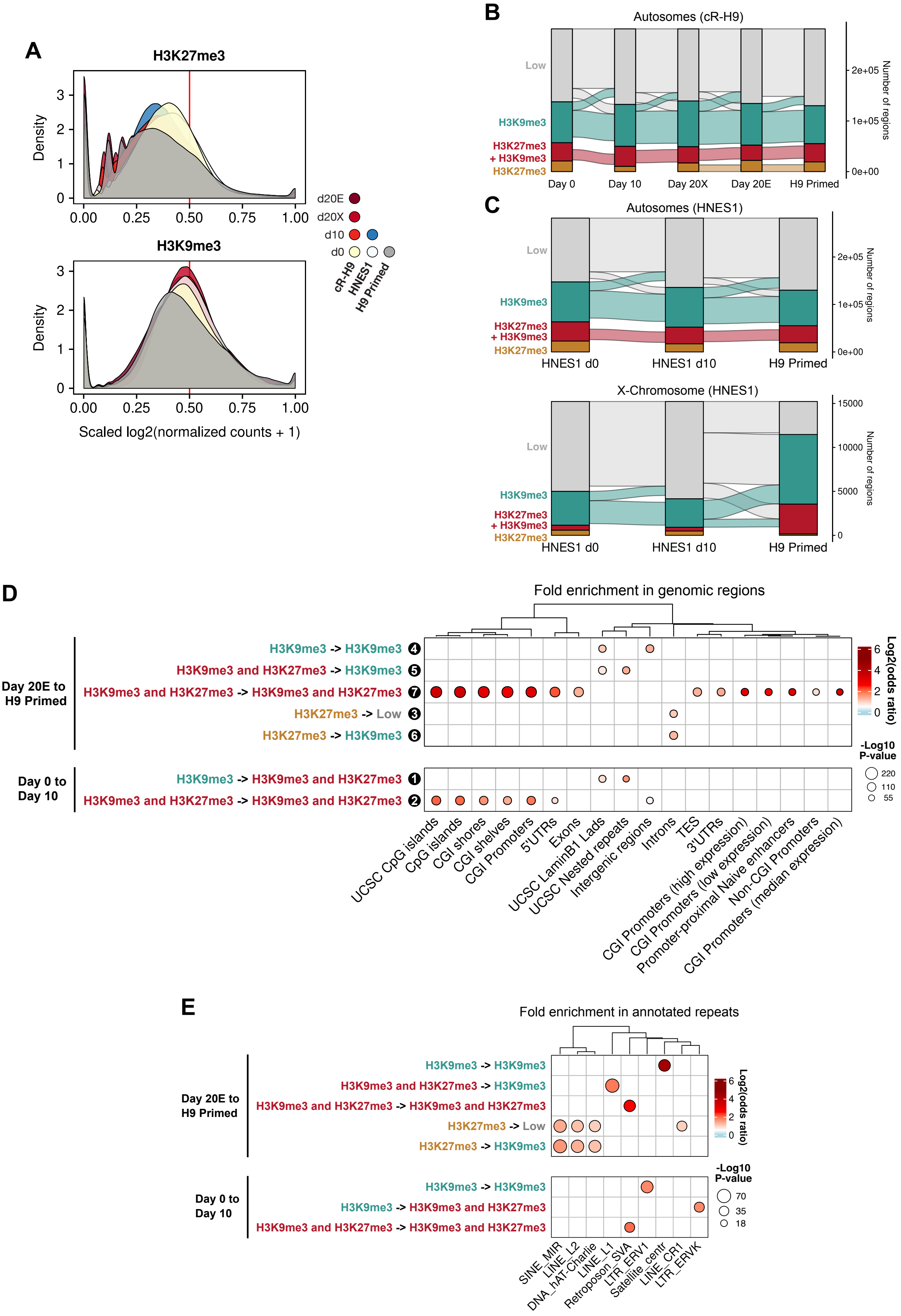
A
H3K27me3
Density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
H3K9me3
Density
Scaled log2(normalized counts + 1)
d20E
d20X
d10
d0
cR-H9
HNES1
H9 Primed
B
Autosomes (cR-H9)
Low
H3K9me3
H3K27me3
+ H3K9me3
H3K27me3
Day 0
Day 10
Day 20X
Day 20E
H9 Primed
2e+05
1e+05
0e+00
Number of regions
C
Autosomes (HNES1)
HNES1 d0
HNES1 d10
H9 Primed
X-Chromosome (HNES1)
15000
10000
5000
0
D
Fold enrichment in genomic regions
Day 20E to H9 Primed
H3K9me3 -> H3K9me3 ❹
H3K9me3 and H3K27me3 -> H3K9me3 ❺
H3K9me3 and H3K27me3 -> H3K9me3 and H3K27me3 ❼
H3K27me3 -> Low ❸
H3K27me3 -> H3K9me3 ❻
Day 0 to Day 10
H3K9me3 -> H3K9me3 and H3K27me3 ❶
H3K9me3 and H3K27me3 -> H3K9me3 and H3K27me3 ❷
UCSC CpG islands
CpG islands
CGI shores
CGI shelves
CGI Promoters
5'UTRs
Exons
UCSC LaminB1 Lads
UCSC Nested repeats
Intergenic regions
Introns
TES
3'UTRs
CGI Promoters (high expression)
CGI Promoters (low expression)
Promoter-proximal Naive enhancers
Non-CGI Promoters
CGI Promoters (median expression)
Log2(odds ratio)
6
4
2
0
-Log10 P-value
220
110
55
E
Fold enrichment in annotated repeats
Day 20E to H9 Primed
H3K9me3 -> H3K9me3
H3K9me3 and H3K27me3 -> H3K9me3
H3K9me3 and H3K27me3 -> H3K9me3 and H3K27me3
H3K27me3 -> Low
H3K27me3 -> H3K9me3
Day 0 to Day 10
H3K9me3 -> H3K9me3
H3K9me3 -> H3K9me3 and H3K27me3
H3K9me3 and H3K27me3 -> H3K9me3 and H3K27me3
SINE_MIR
LINE_L2
DNA_hAT-Charlie
LINE_L1
Retroposon_SVA
LTR_ERV1
Satellite_centr
LINE_CR1
LTR_ERVK
Log2(odds ratio)
-Log10 P-value
70
35
18

Figure S5.

A) Distribution of the H3K27me3 and H3K9me3 log-transformed normalized values in 10 Kb windows over all autosomes. B) Flow chart with the number of 10 Kb windows in autosomes and X chromosome classified as having high levels of H3K27me3 marks, ‘H3K27me3’, high levels of H3K9me3 marks, ‘H3K9me3’, high levels of both H3K27me3 and H3K9me3 marks, ‘H3K27me3 + H3K9me3’, and low levels of H3K27me3 and H3K9me3 marks, ‘Low’. The chart only shows the HNES1 and conventional H9 results. C) Enrichment of the X chromosome regions with classification transition from day 0 to day 10 and from day 20E to conventional hPSCs over annotated genomic regions. D) Enrichment of the X chromosome regions with classification transition from day 0 to day 10 and from day 20E to conventional hPSCs over annotated repetitive elements. F) Flow chart with the number of 10 Kb windows in autosomes classified into the same categories as in figure D.

8. Tables

Condition	RNA-seq	ATAC-seq	ChIP-seq (H3K4me3)	ChIP-seq (H3K4me1)	ChIP-seq (H3K9me3)	ChIP-seq (H3K27me3)	ChIP-seq (H3K27ac)	PBAT
0	3	2	2	2	2	2	2	3 (cR-H9), 4 (HNES1)
1	3	-	-	-	-	-	-	-
2	3	-	-	-	-	-	-	-
3	3	-	-	-	-	-	-	-
7	3	-	-	-	-	-	-	-
10	3	2	2	2	2	2	2	4
20X	3	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	3 (cR-H9 only)
20E	3	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)	2 (cR-H9 only)
H9 Primed	3	2	2	2	2	2	2	4

Table 1. Number of biological replicates for each condition.

The added notes ‘cR-H9 only’, ‘cR-H9’, and ‘HNES1’ indicate that only that cell type has the number of biological replicates indicated. The condition categories include ‘0’ (0 days), ‘1’ (1 day), ‘2’ (2 days), ‘3’ (3 days), ‘7’ (7 days), ‘10’ (10 days), ‘20X’ (more than 10 days in XAF medium after the 10 days in the XAV medium), ‘20E’ (more than 10 days in E8 medium after the 10 days in the XAV medium), and ‘H9 Primed’ (H9 cells cultured in E8 medium).

Status	Gene	Chr	Start	End	Inf	CG content	#CpG	Origin
Novel DMRs	PPIEL	1	39558954	39559868	4	0.54	39	M
Novel DMRs near known imprinted loci	DIRAS3 Ex2	1	68046822	68047803	8	0.52	39	M
Known imprinted DMRs	DIRAS3	1	68049750	68051862	17	0.5	88	M
Placental-specific DMRs	GPR1-AS	2	206202243	206204721	3	0.49	86	M
Known imprinted DMRs	ZDBF2	2	206249859	206271820	8	0.45	439	P
Placental-specific DMRs	MCCC1	3	183097937	183099839	13	0.54	94	M
Known imprinted DMRs	NAP1L5	4	88697033	88698086	15	0.57	57	M
Placental-specific DMRs	PDE4D	5	59037947	59040727	7	0.54	145	M
Known imprinted DMRs	FAM50B	6	3848848	3850125	25	0.65	90	M
Placental-specific DMRs	LIN28B	6	104952756	104954684	8	0.45	62	M
Placental-specific DMRs	AIM1	6	106510070	106514099	19	0.54	203	M
Known imprinted DMRs	PLAGL1	6	144006941	144008751	16	0.58	143	M
Known imprinted DMRs	IGF2R	6	160005526	160006529	2	0.7	74	M
Novel DMRs	WDR27	6	169654408	169655522	2	0.56	58	M
Known imprinted DMRs	GRB10	7	50781029	50783615	9	0.6	171	M
Known imprinted DMRs	PEG10	7	94656225	94658648	53	0.6	119	M
Known imprinted DMRs	MEST	7	130490281	130494547	55	0.54	226	M
Placental-specific DMRs	AGBL3	7	134986273	134987260	12	0.59	74	M
Novel DMRs	HTR5A	7	155071009	155071672	6	0.62	55	M
Novel DMRs	CXORF56 pseudogene/ ERLIN2	8	37747474	37748570	7	0.45	37	M
Placental-specific DMRs	ZFAT	8	134694984	134697871	3	0.6	111	M
Known imprinted DMRs	TRAPPC9	8	140098048	140100982	8	0.62	193	M
Placental-specific DMRs	GLIS3	9	4297279	4300182	9	0.63	235	M
Placental-specific DMRs	DCAF10	9	37800143	37802940	5	0.56	157	M
Known imprinted DMRs	INPP5F	10	119818534	119819215	4	0.59	52	M
Placental-specific DMRs	FAM196A/DOCK1	10	127195141	127196978	10	0.72	198	M

Known imprinted DMRs	H19	11	1997582	2003510	48	0.6	250	P
Known imprinted DMRs	IGF2 DMR2	11	2132761	2133882	9	0.65	63	P
Known imprinted DMRs	IGF2 DMR0	11	2147103	2148538	1	0.62	33	P
Known imprinted DMRs	KvDMR1	11	2698718	2701029	30	0.67	192	M
Placental-specific DMRs	ZC3H12C	11	110092001	110094058	9	0.66	198	M
Placental-specific DMRs	N4BP2L1	13	32426557	32428311	13	0.66	136	M
Known imprinted DMRs	RB1	13	48318205	48321627	12	0.59	195	M
Known imprinted DMRs	IG-DMR	14	100809090	100811721	0	0.52	64	P
Known imprinted DMRs	MEG3	14	100824187	100827641	33	0.6	188	P
Novel DMRs near known imprinted loci	MEG8	14	100904404	100905082	1	0.66	43	M
Known imprinted DMRs	MKRN3/MIR4508	15	23561939	23567348	12	0.44	109	M
Known imprinted DMRs	MAGEL2	15	23647278	23648882	6	0.55	51	M
Known imprinted DMRs	NDN	15	23686304	23687612	8	0.65	108	M
Novel DMRs near known imprinted loci	SNRPN intragenic CpG32	15	24101589	24101995	1	0.59	30	M
Novel DMRs near known imprinted loci	SNRPN intragenic CpG29	15	24426725	24427532	4	0.59	39	M
Novel DMRs near known imprinted loci	SNRPN intragenic CpG30	15	24477606	24477924	1	0.66	29	M
Novel DMRs near known imprinted loci	SNRPN intragenic CpG40	15	24772777	24773739	4	0.51	67	M
Known imprinted DMRs	SNRPN	15	24823417	24824334	8	0.42	19	M
Known imprinted DMRs	SNRPN	15	24847861	24848682	4	0.49	44	M
Known imprinted DMRs	SNRPN	15	24877880	24878758	5	0.47	45	M
Known imprinted DMRs	SNURF	15	24954857	24956829	7	0.6	113	M
Placental-specific DMRs	RGMA	15	93071769	93073630	8	0.61	134	M
Known imprinted DMRs	IGF1R	15	98865267	98866421	7	0.51	55	M
Novel DMRs near known imprinted loci	ZNF597	16	3431801	3432388	2	0.54	29	M

Known imprinted DMRs	ZNF597/NAA60	16	3442828	3444463	11	0.54	76	P
Placental-specific DMRs	FAM20A	17	68600014	68601502	4	0.72	162	M
Placental-specific DMRs	ZNF396	18	35376546	35377616	9	0.64	86	M
Placental-specific DMRs	DNMT1	19	10192830	10195739	10	0.55	129	M
Known imprinted DMRs	ZNF331	19	53537256	53538958	11	0.64	125	M
Novel DMRs near known imprinted loci	ZNF331	19	53553832	53555171	4	0.66	102	M
Placental-specific DMRs	MIR512-1 cluster	19	53647261	53652354	6	0.53	216	M
Known imprinted DMRs	PEG3	19	56837125	56841903	36	0.59	221	M
Known imprinted DMRs	MCTS2P/HM13	20	31546860	31548130	9	0.48	47	M
Known imprinted DMRs	BLCAP/NNAT	20	37520202	37522126	35	0.55	135	M
Known imprinted DMRs	L3MBTL	20	43513725	43515400	25	0.65	84	M
Known imprinted DMRs	GNAS	20	58838984	58843557	23	0.57	257	P
Known imprinted DMRs	NESP-AS/GNAS-AS1	20	58850594	58852978	62	0.61	128	M
Known imprinted DMRs	GNAS XL	20	58853850	58856408	6	0.65	200	M
Known imprinted DMRs	GNAS Ex1A	20	58888210	58890146	38	0.67	198	M
Novel DMRs	WRB	21	39385584	39386350	4	0.61	43	M
Novel DMRs	NHP2L1	22	41681770	41682869	8	0.54	63	M

Table 2. List of human allele-specific differentially methylated regions (DMRs), indicating known, novel, and placental-specific imprinted regions.

Each column provides the following information: ‘Status’ describes the detected DMRs; ‘Gene’ provides the DMR name attributed by matching the name of the nearest gene; ‘Chr’ indicates the chromosome number; ‘Start’ and ‘End’ give the chromosomal start and end positions, respectively; ‘Inf’ shows the number of informative alleles; ‘CG content’ provides the ratio of the number of cytosines (C) and guanines (G) over the total number of nucleotides; ‘#CpG’ indicates the number of CpGs present within the DMR; and ‘Origin’ specifies the parental allele where the imprinting is observed, with ‘M’ indicating maternal and ‘P’ indicating paternal.

Gene name	Gene category
NANOG	General pluripotency
POU5F1	General pluripotency
SALL4	General pluripotency
SOX2	General pluripotency
TDGF1	General pluripotency
ASCL1	Lineage markers
CDX2	Lineage markers
EOMES	Lineage markers
FOXA2	Lineage markers
FOXF1	Lineage markers
GATA4	Lineage markers
GATA6	Lineage markers
MEOX1	Lineage markers
NEUROG1	Lineage markers
NEUROG2	Lineage markers
PAX6	Lineage markers
SOX1	Lineage markers
SOX17	Lineage markers
SOX7	Lineage markers
TBXT	Lineage markers
ZIC1	Lineage markers
DNMT3L	Naïve pluripotency
FGF4	Naïve pluripotency
KLF17	Naïve pluripotency
KLF4	Naïve pluripotency
KLF5	Naïve pluripotency
TFCP2L1	Naïve pluripotency
CDH2	Post-implantation epiblast
ETV4	Post-implantation epiblast
ETV5	Post-implantation epiblast
FGF2	Post-implantation epiblast
FZD7	Post-implantation epiblast
MYC	Post-implantation epiblast
SALL2	Post-implantation epiblast
SFRP2	Post-implantation epiblast
SOX11	Post-implantation epiblast
TCF7L1	Post-implantation epiblast
ZIC2	Post-implantation epiblast

Table 3. List of genes classified by their association with naïve pluripotency, primed pluripotency, post-implantation epiblast, and lineage-specific expression.

Gene symbol	Ensembl gene id	Protein complex group	Protein complex subgroup	Protein complex component
RING1	ENSG00000204227	PcG	PRC1	core
RNF2	ENSG00000121481	PcG	PRC1	core
PCGF1	ENSG00000115289	PcG	PRC1	core
PCGF2	ENSG00000277258	PcG	PRC1	core
PCGF3	ENSG00000185619	PcG	PRC1	core
BMI1	ENSG00000168283	PcG	PRC1	core
PCGF5	ENSG00000180628	PcG	PRC1	core
PCGF6	ENSG00000156374	PcG	PRC1	core
CBX2	ENSG00000173894	PcG	PRC1	canonical
CBX4	ENSG00000141582	PcG	PRC1	canonical
CBX6	ENSG00000183741	PcG	PRC1	canonical
CBX7	ENSG00000100307	PcG	PRC1	canonical
CBX8	ENSG00000141570	PcG	PRC1	canonical
PHC1	ENSG00000111752	PcG	PRC1	canonical
PHC2	ENSG00000134686	PcG	PRC1	canonical
PHC3	ENSG00000173889	PcG	PRC1	canonical
SCMH1	ENSG00000010803	PcG	PRC1	canonical
RYBP	ENSG00000163602	PcG	PRC1	Variant PRC1 component
YAF2	ENSG00000015153	PcG	PRC1	Variant PRC1 component
KDM2B	ENSG00000089094	PcG	PRC1	Variant PRC1 component
DCAF7	ENSG00000136485	PcG	PRC1	Variant PRC1 component
WDR5	ENSG00000196363	PcG	PRC1	Variant PRC1 component
SCML1	ENSG00000047634	PcG	PRC1	canonical
SCML2	ENSG00000102098	PcG	PRC1	canonical
USP7	ENSG00000187555	PcG	PRC1	Variant PRC1 component
BCOR	ENSG00000183337	PcG	PRC1	Variant PRC1 component
SKP1	ENSG00000113558	PcG	PRC1	Variant PRC1 component
MAX	ENSG00000125952	PcG	PRC1	Variant PRC1 component
MGA	ENSG00000174197	PcG	PRC1	Variant PRC1 component
L3MBTL2	ENSG00000100395	PcG	PRC1	Variant PRC1 component
E2F6	ENSG00000169016	PcG	PRC1	Variant PRC1 component
TFDP1	ENSG00000198176	PcG	PRC1	Variant PRC1 component

TFDP2	ENSG00000114126	PcG	PRC1	Variant PRC1 component
HDAC1	ENSG00000116478	PcG	PRC1	Variant PRC1 component
HDAC2	ENSG00000196591	PcG	PRC1	Variant PRC1 component
FBRS	ENSG00000156860	PcG	PRC1	Variant PRC1 component
AUTS2	ENSG00000158321	PcG	PRC1	Variant PRC1 component
CSNK2A1	ENSG00000101266	PcG	PRC1	Variant PRC1 component
CSNK2A2	ENSG00000070770	PcG	PRC1	Variant PRC1 component
CSNK2B	ENSG00000204435	PcG	PRC1	Variant PRC1 component
EZH1	ENSG00000108799	PcG	PRC2	core
EZH2	ENSG00000106462	PcG	PRC2	core
SUZ12	ENSG00000178691	PcG	PRC2	core
EED	ENSG00000074266	PcG	PRC2	core
RBBP4	ENSG00000162521	PcG	PRC2	core
RBBP7	ENSG00000102054	PcG	PRC2	core
PHF1	ENSG00000112511	PcG	PRC2	PRC2.1
PHF19	ENSG00000119403	PcG	PRC2	PRC2.1
MTF2	ENSG00000143033	PcG	PRC2	PRC2.1
JARID2	ENSG00000008083	PcG	PRC2	PRC2.2
AEBP2	ENSG00000139154	PcG	PRC2	PRC2.2
EPOP	ENSG00000273604	PcG	PRC2	co-purify
LCOR	ENSG00000196233	PcG	PRC2	co-purify
BAP1	ENSG00000163930	PcG	PR-DUB	core
ASXL1	ENSG00000171456	PcG	PR-DUB	PR-DUB1
ASXL2	ENSG00000143970	PcG	PR-DUB	PR-DUB2
FOXK1	ENSG00000164916	PcG	PR-DUB	accessory
FOXK2	ENSG00000141568	PcG	PR-DUB	accessory
OGT	ENSG00000147162	PcG	PR-DUB	accessory
KDM1B	ENSG00000165097	PcG	PR-DUB	accessory
MBD5	ENSG00000204406	PcG	PR-DUB	accessory
MBD6	ENSG00000166987	PcG	PR-DUB	accessory
SMARCA2	ENSG00000080503	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCA4	ENSG00000127616	TrxG	SWI/SNF (BAF and PBAF)	NA

ARID1A	ENSG00000117713	TrxG	SWI/SNF (BAF and PBAF)	NA
ARID1B	ENSG00000049618	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCC1	ENSG00000173473	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCC2	ENSG00000139613	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCB1	ENSG00000099956	TrxG	SWI/SNF (BAF and PBAF)	NA
PHF10	ENSG00000130024	TrxG	SWI/SNF (BAF and PBAF)	NA
DPF1	ENSG00000011332	TrxG	SWI/SNF (BAF and PBAF)	NA
DPF3	ENSG00000205683	TrxG	SWI/SNF (BAF and PBAF)	NA
DPF2	ENSG00000133884	TrxG	SWI/SNF (BAF and PBAF)	NA
ACTL6A	ENSG00000136518	TrxG	SWI/SNF (BAF and PBAF)	NA
ACTL6B	ENSG00000077080	TrxG	SWI/SNF (BAF and PBAF)	NA
PBRM1	ENSG00000163939	TrxG	SWI/SNF (BAF and PBAF)	NA
ARID2	ENSG00000189079	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCD1	ENSG00000066117	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCD2	ENSG00000108604	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCD3	ENSG00000082014	TrxG	SWI/SNF (BAF and PBAF)	NA
SMARCE1	ENSG00000073584	TrxG	SWI/SNF (BAF and PBAF)	NA
ACTB	ENSG00000075624	TrxG	SWI/SNF (BAF and PBAF)	NA
BCL7A	ENSG00000110987	TrxG	SWI/SNF (BAF and PBAF)	NA
BCL7B	ENSG00000106635	TrxG	SWI/SNF (BAF and PBAF)	NA
BCL7C	ENSG00000099385	TrxG	SWI/SNF (BAF and PBAF)	NA
BRD7	ENSG00000166164	TrxG	SWI/SNF (BAF and PBAF)	NA

BRD8	ENSG00000112983	TrxG	SWI/SNF (BAF and PBAF)	NA
BRD9	ENSG00000028310	TrxG	SWI/SNF (BAF and PBAF)	NA
WDR5	ENSG00000196363	TrxG	COMPASS	core
ASH2L	ENSG00000129691	TrxG	COMPASS	core
RBBP5	ENSG00000117222	TrxG	COMPASS	core
DPY30	ENSG00000162961	TrxG	COMPASS	core
SETD1A	ENSG00000099381	TrxG	SET1/COMPASS	accessory
SETD1B	ENSG00000139718	TrxG	SET1/COMPASS	accessory
HCFC1	ENSG00000172534	TrxG	SET1/COMPASS	accessory
WDR82	ENSG00000164091	TrxG	SET1/COMPASS	accessory
CXXC1	ENSG00000154832	TrxG	SET1/COMPASS	accessory
KMT2A	ENSG00000118058	TrxG	MLL1/MLL2 COMPASS	accessory
KMT2B	ENSG00000272333	TrxG	MLL1/MLL2 COMPASS	accessory
HCFC1	ENSG00000172534	TrxG	MLL1/MLL2 COMPASS	accessory
MEN1	ENSG00000133895	TrxG	MLL1/MLL2 COMPASS	accessory
KMT2C	ENSG00000055609	TrxG	MLL3/MLL4 COMPASS	accessory
KMT2D	ENSG00000167548	TrxG	MLL3/MLL4 COMPASS	accessory
NCOA6	ENSG00000198646	TrxG	MLL3/MLL4 COMPASS	accessory
PAGR1	ENSG00000280789	TrxG	MLL3/MLL4 COMPASS	accessory
KDM6A	ENSG00000147050	TrxG	MLL3/MLL4 COMPASS	accessory
PAXIP1	ENSG00000157212	TrxG	MLL3/MLL4 COMPASS	accessory
DNMT1	ENSG00000130816	NA	NA	NA
DNMT3A	ENSG00000119772	NA	NA	NA
DNMT3B	ENSG00000088305	NA	NA	NA
DNMT3L	ENSG00000142182	NA	NA	NA
IDH1	ENSG00000138413	NA	NA	NA
IDH1-AS1	ENSG00000231908	NA	NA	NA
IDH2	ENSG00000182054	NA	NA	NA

IDH2-DT	ENSG00000259685	NA	NA	NA
IDH3A	ENSG00000166411	NA	NA	NA
IDH3B	ENSG00000101365	NA	NA	NA
TET1	ENSG00000138336	NA	NA	NA
TET1P1	ENSG00000232204	NA	NA	NA
TET2	ENSG00000168769	NA	NA	NA
TET2-AS1	ENSG00000251586	NA	NA	NA
TET3	ENSG00000187605	NA	NA	NA

Table 4. List of genes and their corresponding protein products that belong to the Polycomb-group proteins (PcG), Trithorax-group proteins (TrxG), and the DNMT, TET, and IDH protein families.

The table contains the following information: ‘Gene name’ gives the official gene symbol; ‘Protein name’ gives the name of the protein product; ‘Protein family’ indicates the protein family to which the protein product belongs to, either PcG, TrxG, DNMT, TET, or IDH.

ID	Description	GeneRatio	BgRatio	qvalue	Cluster
GO:0048568	embryonic organ development	49/787	427/18480	1.25E-06	1
GO:0007389	pattern specification process	48/787	434/18480	3.30E-06	1
GO:0045165	cell fate commitment	30/787	258/18480	4.86E-04	1
GO:0048705	skeletal system morphogenesis	27/787	218/18480	4.86E-04	1
GO:0061687	detoxification of inorganic compound	7/787	17/18480	0.00200809	1
GO:0048706	embryonic skeletal system development	18/787	121/18480	0.00200809	1
GO:0003205	cardiac chamber development	21/787	162/18480	0.0027124	1
GO:0010273	detoxification of copper ion	6/787	14/18480	0.00485252	1
GO:1990169	stress response to copper ion	6/787	14/18480	0.00485252	1
GO:1903046	meiotic cell cycle process	16/306	202/18480	8.19E-04	2
GO:0002088	lens development in camera-type eye	10/306	84/18480	0.00189675	2
GO:0001655	urogenital system development	31/344	338/18480	7.32E-10	3
GO:0072001	renal system development	29/344	302/18480	7.32E-10	3
GO:0001657	ureteric bud development	16/344	91/18480	7.56E-09	3
GO:0072164	mesonephric tubule development	16/344	92/18480	7.56E-09	3
GO:0045165	cell fate commitment	23/344	258/18480	2.23E-07	3
GO:0061564	axon development	31/344	466/18480	3.12E-07	3
GO:0060485	mesenchyme development	24/344	291/18480	3.49E-07	3
GO:0007369	gastrulation	19/344	184/18480	4.17E-07	3
GO:0042476	odontogenesis	16/344	130/18480	5.91E-07	3
GO:0048754	branching morphogenesis of an epithelial tube	17/344	150/18480	6.28E-07	3
GO:0001667	ameboidal-type cell migration	30/344	475/18480	1.10E-06	3
GO:0001763	morphogenesis of a branching structure	18/344	195/18480	3.96E-06	3
GO:0060537	muscle tissue development	26/344	403/18480	5.64E-06	3
GO:0007389	pattern specification process	27/344	434/18480	6.12E-06	3
GO:0043062	extracellular structure organization	22/344	301/18480	6.12E-06	3
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	24/344	354/18480	6.12E-06	3
GO:0060560	developmental growth involved in morphogenesis	19/344	233/18480	8.18E-06	3
GO:0042303	molting cycle	13/344	106/18480	8.18E-06	3
GO:0042633	hair cycle	13/344	106/18480	8.18E-06	3
GO:0110110	positive regulation of animal organ morphogenesis	8/344	33/18480	1.10E-05	3
GO:0050920	regulation of chemotaxis	18/344	221/18480	1.49E-05	3
GO:0006940	regulation of smooth muscle contraction	10/344	64/18480	1.86E-05	3
GO:0051960	regulation of nervous system development	26/344	444/18480	1.89E-05	3
GO:0002040	sprouting angiogenesis	16/344	185/18480	2.37E-05	3
GO:0050767	regulation of neurogenesis	23/344	365/18480	2.37E-05	3
GO:0042060	wound healing	25/344	423/18480	2.39E-05	3
GO:0048545	response to steroid hormone	22/344	339/18480	2.44E-05	3

GO:1903034	regulation of response to wounding	15/344	167/18480	3.05E-05	3
GO:0050804	modulation of chemical synaptic transmission	25/344	434/18480	3.50E-05	3
GO:0008544	epidermis development	21/344	326/18480	4.52E-05	3
GO:0060326	cell chemotaxis	20/344	306/18480	6.38E-05	3
GO:0090130	tissue migration	22/344	364/18480	6.51E-05	3
GO:0030850	prostate gland development	8/344	45/18480	7.11E-05	3
GO:0045445	myoblast differentiation	10/344	84/18480	1.44E-04	3
GO:0048638	regulation of developmental growth	20/344	329/18480	1.54E-04	3
GO:0016049	cell growth	25/344	480/18480	1.54E-04	3
GO:0033002	muscle cell proliferation	17/344	248/18480	1.61E-04	3
GO:0090287	regulation of cellular response to growth factor stimulus	19/344	303/18480	1.61E-04	3
GO:0009914	hormone transport	19/344	306/18480	1.80E-04	3
GO:0009410	response to xenobiotic stimulus	24/344	459/18480	2.03E-04	3
GO:0006939	smooth muscle contraction	11/344	110/18480	2.14E-04	3
GO:0042634	regulation of hair cycle	6/344	26/18480	2.14E-04	3
GO:0008361	regulation of cell size	14/344	182/18480	2.51E-04	3
GO:0048332	mesoderm morphogenesis	9/344	74/18480	2.68E-04	3
GO:0002690	positive regulation of leukocyte chemotaxis	10/344	94/18480	2.86E-04	3
GO:0001558	regulation of cell growth	22/344	412/18480	2.97E-04	3
GO:0021953	central nervous system neuron differentiation	13/344	162/18480	3.09E-04	3
GO:0046879	hormone secretion	18/344	295/18480	3.13E-04	3
GO:0007517	muscle organ development	19/344	326/18480	3.40E-04	3
GO:0007492	endoderm development	9/344	77/18480	3.46E-04	3
GO:0010720	positive regulation of cell development	18/344	299/18480	3.66E-04	3
GO:0010453	regulation of cell fate commitment	6/344	30/18480	4.22E-04	3
GO:0031589	cell-substrate adhesion	20/344	363/18480	4.33E-04	3
GO:0050900	leukocyte migration	20/344	366/18480	4.79E-04	3
GO:0001942	hair follicle development	9/344	81/18480	4.88E-04	3
GO:0030858	positive regulation of epithelial cell differentiation	8/344	64/18480	5.58E-04	3
GO:0007218	neuropeptide signaling pathway	10/344	104/18480	5.63E-04	3
GO:0007498	mesoderm development	11/344	129/18480	6.59E-04	3
GO:0050878	regulation of body fluid levels	20/344	379/18480	6.81E-04	3
GO:0048588	developmental cell growth	15/344	233/18480	7.21E-04	3
GO:0048660	regulation of smooth muscle cell proliferation	13/344	180/18480	7.32E-04	3
GO:0031128	developmental induction	6/344	34/18480	7.40E-04	3
GO:1905952	regulation of lipid localization	13/344	182/18480	8.04E-04	3
GO:0048738	cardiac muscle tissue development	15/344	236/18480	8.06E-04	3
GO:0010595	positive regulation of endothelial cell migration	11/344	133/18480	8.14E-04	3
GO:0045661	regulation of myoblast differentiation	7/344	51/18480	8.50E-04	3
GO:0050890	cognition	17/344	298/18480	9.22E-04	3

GO:0021936	regulation of cerebellar granule cell precursor proliferation	4/344	12/18480	9.80E-04	3
GO:0048568	embryonic organ development	21/344	427/18480	0.00101895	3
GO:0060353	regulation of cell adhesion molecule production	5/344	23/18480	0.00101895	3
GO:0009612	response to mechanical stimulus	14/344	216/18480	0.00102422	3
GO:1902893	regulation of pri-miRNA transcription by RNA polymerase II	7/344	54/18480	0.00110472	3
GO:0045666	positive regulation of neuron differentiation	9/344	94/18480	0.00115534	3
GO:0051402	neuron apoptotic process	15/344	247/18480	0.00115696	3
GO:0001654	eye development	19/344	369/18480	0.00120853	3
GO:0061614	pri-miRNA transcription by RNA polymerase II	7/344	55/18480	0.00120853	3
GO:0045926	negative regulation of growth	15/344	249/18480	0.00123714	3
GO:0002064	epithelial cell development	14/344	221/18480	0.0012491	3
GO:0110148	biomineralization	12/344	170/18480	0.00141054	3
GO:0035902	response to immobilization stress	5/344	25/18480	0.00141054	3
GO:0031667	response to nutrient levels	22/344	474/18480	0.00142102	3
GO:0021510	spinal cord development	9/344	99/18480	0.00153603	3
GO:0043524	negative regulation of neuron apoptotic process	11/344	147/18480	0.00155776	3
GO:0060352	cell adhesion molecule production	5/344	26/18480	0.00159383	3
GO:0001659	temperature homeostasis	12/344	174/18480	0.00161197	3
GO:0009266	response to temperature stimulus	12/344	175/18480	0.00169344	3
GO:0051480	regulation of cytosolic calcium ion concentration	18/344	353/18480	0.00176599	3
GO:0032412	regulation of ion transmembrane transporter activity	15/344	268/18480	0.00235198	3
GO:0070997	neuron death	18/344	362/18480	0.00236772	3
GO:0035051	cardiocyte differentiation	11/344	156/18480	0.00242057	3
GO:0032370	positive regulation of lipid transport	8/344	84/18480	0.00242974	3
GO:0050886	endocrine process	8/344	84/18480	0.00242974	3
GO:0042063	gliogenesis	16/344	301/18480	0.0024607	3
GO:1990845	adaptive thermogenesis	11/344	157/18480	0.0024607	3
GO:1904862	inhibitory synapse assembly	4/344	16/18480	0.0024607	3
GO:0007162	negative regulation of cell adhesion	16/344	304/18480	0.00266154	3
GO:0050918	positive chemotaxis	7/344	65/18480	0.00268478	3
GO:0001666	response to hypoxia	16/344	306/18480	0.00282748	3
GO:0050673	epithelial cell proliferation	20/344	437/18480	0.00285294	3
GO:2000177	regulation of neural precursor cell proliferation	8/344	87/18480	0.00285294	3
GO:0010038	response to metal ion	18/344	371/18480	0.00286486	3
GO:0030193	regulation of blood coagulation	7/344	66/18480	0.00286486	3
GO:0007568	aging	17/344	339/18480	0.00287302	3
GO:0007596	blood coagulation	13/344	217/18480	0.00288569	3
GO:0001503	ossification	19/344	405/18480	0.00290083	3
GO:0032409	regulation of transporter activity	16/344	311/18480	0.00313725	3
GO:0061448	connective tissue development	14/344	250/18480	0.00319195	3

GO:0071560	cellular response to transforming growth factor beta stimulus	14/344	250/18480	0.00319195	3
GO:0050817	coagulation	13/344	222/18480	0.00337733	3
GO:0035296	regulation of tube diameter	10/344	141/18480	0.00361725	3
GO:0097746	blood vessel diameter maintenance	10/344	141/18480	0.00361725	3
GO:0060688	regulation of morphogenesis of a branching structure	6/344	50/18480	0.00364565	3
GO:0071559	response to transforming growth factor beta	14/344	256/18480	0.00376765	3
GO:0050818	regulation of coagulation	7/344	71/18480	0.00390049	3
GO:1901214	regulation of neuron death	16/344	321/18480	0.00397699	3
GO:0050869	negative regulation of B cell activation	5/344	34/18480	0.00420601	3
GO:0090050	positive regulation of cell migration involved in sprouting angiogenesis	5/344	35/18480	0.00467755	3
GO:0003151	outflow tract morphogenesis	7/344	74/18480	0.00470883	3
GO:0045834	positive regulation of lipid metabolic process	10/344	148/18480	0.00477845	3
GO:0043627	response to estrogen	7/344	75/18480	0.00502471	3
GO:0060840	artery development	8/344	100/18480	0.00557506	3
GO:0021532	neural tube patterning	5/344	37/18480	0.00572455	3
GO:0050866	negative regulation of cell activation	12/344	211/18480	0.00607026	3
GO:0002790	peptide secretion	13/344	242/18480	0.00620368	3
GO:0051412	response to corticosterone	4/344	22/18480	0.00633047	3
GO:0055093	response to hyperoxia	4/344	22/18480	0.00633047	3
GO:0009750	response to fructose	3/344	10/18480	0.00654222	3
GO:0044857	plasma membrane raft organization	3/344	10/18480	0.00654222	3
GO:0043583	ear development	12/344	214/18480	0.00655815	3
GO:0016055	Wnt signaling pathway	19/344	444/18480	0.00655815	3
GO:0007422	peripheral nervous system development	7/344	80/18480	0.00672497	3
GO:0198738	cell-cell signaling by wnt	19/344	446/18480	0.00685208	3
GO:0036303	lymph vessel morphogenesis	4/344	23/18480	0.00719752	3
GO:0032496	response to lipopolysaccharide	16/344	345/18480	0.00719752	3
GO:0043434	response to peptide hormone	18/344	415/18480	0.00748453	3
GO:0010876	lipid localization	19/344	451/18480	0.0076318	3
GO:0097191	extrinsic apoptotic signaling pathway	12/344	219/18480	0.0077208	3
GO:0061053	somite development	7/344	83/18480	0.00805124	3
GO:0000302	response to reactive oxygen species	12/344	221/18480	0.00820291	3
GO:0033138	positive regulation of peptidyl-serine phosphorylation	8/344	108/18480	0.00820291	3
GO:0009755	hormone-mediated signaling pathway	11/344	191/18480	0.00820291	3
GO:0001765	membrane raft assembly	3/344	11/18480	0.00824	3
GO:0099550	trans-synaptic signaling, modulating synaptic transmission	3/344	11/18480	0.00824	3
GO:0034329	cell junction assembly	18/344	420/18480	0.00824814	3
GO:0034612	response to tumor necrosis factor	13/344	254/18480	0.00861221	3
GO:0060393	regulation of pathway-restricted SMAD protein phosphorylation	6/344	62/18480	0.00872725	3

GO:0050680	negative regulation of epithelial cell proliferation	10/344	164/18480	0.00872725	3
GO:0040013	negative regulation of locomotion	17/344	389/18480	0.00889866	3
GO:1901654	response to ketone	11/344	194/18480	0.00889866	3
GO:0001756	somitogenesis	6/344	63/18480	0.00918262	3
GO:0030239	myofibril assembly	6/344	63/18480	0.00918262	3
GO:2000242	negative regulation of reproductive process	6/344	63/18480	0.00918262	3
GO:0045927	positive regulation of growth	13/344	258/18480	0.00945342	3
GO:0045598	regulation of fat cell differentiation	9/344	139/18480	0.00974856	3
GO:0031099	regeneration	11/344	197/18480	0.00978359	3
GO:0045995	regulation of embryonic development	6/344	64/18480	0.00978359	3
GO:1903800	positive regulation of production of miRNAs involved in gene silencing by miRNA	3/344	12/18480	0.00994994	3
GO:0031623	receptor internalization	8/344	113/18480	0.00995495	3
GO:0061564	axon development	42/616	466/18480	2.18E-05	4
GO:0051451	myoblast migration	6/616	13/18480	0.00151594	4
GO:0007188	adenylate cyclase-modulating G protein-coupled receptor signaling pathway	22/616	234/18480	0.00848547	4
GO:0060042	retina morphogenesis in camera-type eye	10/616	57/18480	0.00853434	4
GO:1990138	neuron projection extension	18/616	171/18480	0.00853434	4
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	38/736	274/18480	8.20E-08	5
GO:0034329	cell junction assembly	44/736	420/18480	7.17E-06	5
GO:0007416	synapse assembly	25/736	180/18480	5.64E-05	5
GO:0050808	synapse organization	41/736	427/18480	1.62E-04	5
GO:0090075	relaxation of muscle	9/736	34/18480	0.00311216	5
GO:0050804	modulation of chemical synaptic transmission	37/736	434/18480	0.00596568	5
GO:0099177	regulation of trans-synaptic signaling	37/736	435/18480	0.00596568	5
GO:0021988	olfactory lobe development	8/736	31/18480	0.00902076	5
GO:1901888	regulation of cell junction assembly	22/736	205/18480	0.00937615	5

Table 5. Gene ontology terms that were significantly enriched from genes belonging to the RNA sequencing clusters. Each row represents a different gene ontology term. Only gene ontology terms with a significant enrichment (adjusted p-value < 0.05) are included in the table.

Software package name	Version	Documentation	Reference
AnnotationDbi	v1.56.2	https://bioconductor.org/packages/AnnotationDbi	153
annotatr	1.20.0	https://bioconductor.org/packages/annotatr	127
BEclear	v2.10.0	https://bioconductor.org/packages/BEclear	154
Biobase	v2.54.0	https://bioconductor.org/packages/Biobase	155
boot	v1.3-28	https://cran.r-project.org/package=boot	144
BSgenome	v1.62.0	https://bioconductor.org/packages/BSgenome	143
BSgenome.Hsapiens.UCSC.hg38	v1.4.4	https://bioconductor.org/packages/BSgenome.Hsapiens.UCSC.hg38	156
bsseq	v1.30.0	https://bioconductor.org/packages/bsseq	157
ChIPseeker	v1.30.3	https://bioconductor.org/packages/ChIPseeker	158
ChromHMM	v1.22	http://compbio.mit.edu/ChromHMM/	141
circlize	v0.4.15	https://cran.r-project.org/package=circlize	159
clusterProfiler	v4.2.2	https://bioconductor.org/packages/clusterProfiler	146
colorspace	v2.0-3	https://cran.r-project.org/package=colorspace	160
ComplexHeatmap	v2.10.0	https://bioconductor.org/packages/ComplexHeatmap	161
conflicted	v1.1.0	https://cran.r-project.org/package=conflicted	162
data.table	v1.14.2	https://cran.r-project.org/package=data.table	163
DEGreport	v1.30.3	https://bioconductor.org/packages/DEGreport	164
DESeq2	v1.34.0	https://bioconductor.org/packages/DESeq2	136
DEXSeq	v1.40.0	https://bioconductor.org/packages/DEXSeq	165
DiffBind	v3.4.11	https://bioconductor.org/packages/DiffBind	140
dmrseq	v1.14.0	https://bioconductor.org/packages/dmrseq	166
DRIMSeq	v1.22.0	https://bioconductor.org/packages/DRIMSeq	167
edgeR	v3.36.0	https://bioconductor.org/packages/edgeR	102
EnrichedHeatmap	v1.24.0	https://bioconductor.org/packages/EnrichedHeatmap	168
ensembldb	v2.18.4	https://bioconductor.org/packages/ensembldb	169
furrr	v0.3.0	https://cran.r-project.org/package=furrr	170
future	v1.25.0	https://cran.r-project.org/package=future	171
GenomicAlignments	v1.30.0	https://bioconductor.org/packages/GenomicAlignments	137
GenomicFeatures	v1.46.5	https://bioconductor.org/packages/GenomicFeatures	137
GenomicRanges	v1.46.1	https://bioconductor.org/packages/GenomicRanges	137
ggalluvial	v0.12.3	https://cran.r-project.org/package=ggalluvial	172
ggfortify	v0.4.14	https://cran.r-project.org/package=ggfortify	173
ggplot2	v3.3.6	https://cran.r-project.org/package=ggplot2	174
ggthemes	v4.2.4	https://cran.r-project.org/package=ggthemes	175

gridExtra	v2.3	https://cran.r-project.org/package=gridExtra	176
IGV	v2.13.0	https://software.broadinstitute.org/software/igv/	177
janitor	v2.1.0	https://cran.r-project.org/package=janitor	178
LOLA	v1.24.0	https://bioconductor.org/packages/LOLA	109
magrittr	v2.0.3	https://cran.r-project.org/package=magrittr	179
MEME Suite	v5.4.1	https://meme-suite.org/	150
MEME-ChIP	v5.4.1	https://meme-suite.org/meme/doc/meme-chip.html	149
methylKit	v1.20.0	https://bioconductor.org/packages/methylKit	138
Mfuzz	v2.54.0	https://bioconductor.org/packages/Mfuzz	139
MOFA2	v1.4.0	https://bioconductor.org/packages/MOFA2	95
Nextflow	v19.10.0	https://www.nextflow.io/	180
nf-core/atacseq	v1.1.0	https://nf-co.re/atacseq	181
nf-core/chipseq	v1.1.0	https://nf-co.re/chipseq	182
nf-core/methylseq	v1.4	https://nf-co.re/methylseq	183
nf-core/rnaseq	v1.4.2	https://nf-co.re/rnaseq	184
org.Hs.eg.db	v3.14.0	https://bioconductor.org/packages/org.Hs.eg.db	147
parallelly	v1.31.1	https://cran.r-project.org/package=parallelly	185
patchwork	v1.1.1	https://cran.r-project.org/package=patchwork	186
pheatmap	v1.0.12	https://cran.r-project.org/package=pheatmap	187
profileplyr	v1.10.2	https://bioconductor.org/packages/profileplyr	188
purrr	v0.3.4	https://cran.r-project.org/package=purrr	189
R	v4.1.3	https://www.r-project.org/	190
RColorBrewer	v1.1-3	https://cran.r-project.org/package=RColorBrewer	191
relaimpo	v2.2-6	https://cran.r-project.org/package=relaimpo	145
reshape2	v1.4.4	https://cran.r-project.org/package=reshape2	192
reticulate	v1.25	https://cran.r-project.org/package=reticulate	193
rlist	v0.4.6.2	https://cran.r-project.org/package=rlist	194
Rstudio server	v2022.02.1 +461	https://posit.co/products/open-source/rstudio-server/	195
rtracklayer	v1.54.0	https://bioconductor.org/packages/rtracklayer	196
SalmonTE	v0.4	http://liuzlab.org/salmonte/	152
scales	v1.2.0	https://cran.r-project.org/package=scales	197
SeqMonk	v1.48.1	https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/	142
SNPsplit	v0.4.0	https://www.bioinformatics.babraham.ac.uk/projects/SNPsplit/	148
stringr	v1.4.0	https://cran.r-project.org/package=stringr	198
tidyverse	v1.3.1	https://cran.r-project.org/package=tidyverse	199

TxDb.Hsapiens.UCSC.hg38.knownGene	v3.14.0	https://bioconductor.org/packages/TxDb.Hsapiens.UCSC.hg38.knownGene	200
tximport	v1.22.0	https://bioconductor.org/packages/tximport	135

Table 6. List of software packages used in data processing and analysis.

Bibliography

1. Nichols, J. & Smith, A. Pluripotency in the Embryo and in Culture. *Cold Spring Harb Perspect Biol* **4**, a008128 (2012).
2. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat Rev Mol Cell Biol* **17**, 155–169 (2016).
3. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).
4. Molè, M. A., Weberling, A. & Zernicka-Goetz, M. Chapter Four - Comparative analysis of human and mouse development: From zygote to pre-gastrulation. in *Current Topics in Developmental Biology* (ed. Solnica-Krezel, L.) vol. 136 113–138 (Academic Press, 2020).
5. Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental Biology* **375**, 54–64 (2013).
6. Rossant, J. & Tam, P. P. L. New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. *Cell Stem Cell* **20**, 18–28 (2017).
7. Shahbazi, M. N. & Zernicka-Goetz, M. Deconstructing and reconstructing the mouse and human early embryo. *Nature Cell Biology* **20**, 878–887 (2018).
8. Chovanec, P. *et al.* Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. *Nat Commun* **12**, 2098 (2021).
9. Vallot, C. *et al.* Erosion of X Chromosome Inactivation in Human Pluripotent Cells Initiates with XACT Coating and Depends on a Specific Heterochromatin Landscape. *Cell Stem Cell* **16**, 533–546 (2015).
10. Rostovskaya, M., Stirparo, G. G. & Smith, A. Capacitation of human naïve pluripotent stem cells for multi-lineage differentiation. *Development* **146**, dev172916 (2019).
11. The Developing Human - 11th Edition. https://www.elsevier.com/books/the-developing-human/moore/978-0-323-61154-1.

12. Human Embryology and Developmental Biology. https://shop.elsevier.com/books/human-embryology-and-developmental-biology/carlson/978-0-323-52375-2.
13. Sadler, T. W. & Langman, J. *Langman's medical embryology*. (Lippincott Williams & Wilkins, 2004).
14. Developmental Biology - Michael J.F. Barresi; Scott F. Gilbert - Oxford University Press. //global.oup.com/ushe/product/developmental-biology-9781605358246.
15. Theunissen, T. W. *et al.* Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**, 502–515 (2016).
16. Zhu, Z. & Huangfu, D. Human pluripotent stem cells: an emerging model in developmental biology. *Development* **140**, 705–717 (2013).
17. Takahashi, S., Kobayashi, S. & Hiratani, I. Epigenetic differences between naïve and primed pluripotent stem cells. *Cellular and Molecular Life Sciences* **75**, 1191–1203 (2018).
18. Battle, S. L. *et al.* Enhancer Chromatin and 3D Genome Architecture Changes from Naive to Primed Human Embryonic Stem Cell States. *Stem Cell Reports* **12**, 1129–1144 (2019).
19. Dong, C. *et al.* Derivation of trophoblast stem cells from naïve human pluripotent stem cells. *eLife* **9**, e52504 (2020).
20. Guo, G. *et al.* Epigenetic resetting of human pluripotency. *Development* **144**, 2748–2763 (2017).
21. Takashima, Y. *et al.* Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. *Cell* **158**, 1254–1269 (2014).
22. Rossant, J. Genetic Control of Early Cell Lineages in the Mammalian Embryo. *Annual Review of Genetics* **52**, 185–201 (2018).
23. Guo, G. *et al.* Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063–1069 (2009).
24. Wang, X. *et al.* The transcription factor TFCP2L1 induces expression of distinct target genes and promotes self-renewal of mouse and human embryonic stem cells. *Journal of Biological Chemistry* **294**, 6007–6016 (2019).

25. Chen, H. *et al.* Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. *Nat Commun* **6**, 7095 (2015).
26. Nichols, J. & Smith, A. Naive and Primed Pluripotent States. *Cell Stem Cell* **4**, 487–492 (2009).
27. Theunissen, T. W. *et al.* Nanog overcomes reprogramming barriers and induces pluripotency in minimal conditions. *Curr Biol* **21**, 65–71 (2011).
28. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
29. Dodsworth, B. T., Flynn, R. & Cowley, S. A. The Current State of Naïve Human Pluripotency. *Stem Cells* **33**, 3181–3186 (2015).
30. Thomson, J. A. *et al.* Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science* **282**, 1145–1147 (1998).
31. Chen, G. *et al.* Chemically defined conditions for human iPSC derivation and culture. *Nat Methods* **8**, 424–429 (2011).
32. Bar, S. & Benvenisty, N. Epigenetic aberrations in human pluripotent stem cells. *The EMBO Journal* **38**, e101033 (2019).
33. Lee, J.-H. *et al.* Lineage-Specific Differentiation Is Influenced by State of Human Pluripotency. *Cell Reports* **19**, 20–35 (2017).
34. Liu, W., Deng, C., Godoy-Parejo, C., Zhang, Y. & Chen, G. Developments in cell culture systems for human pluripotent stem cells. *World Journal of Stem Cells* **11**, 968–981 (2019).
35. 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 (2011).
36. Kalkan, T. *et al.* Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144**, 1221–1234 (2017).
37. Mulas, C., Kalkan, T. & Smith, A. NODAL Secures Pluripotency upon Embryonic Stem Cell Progression from the Ground State. *Stem Cell Reports* **9**, 77–91 (2017).

38. Xu, Z. *et al.* Wnt/β-catenin signaling promotes self-renewal and inhibits the primed state transition in naïve human embryonic stem cells. *Proceedings of the National Academy of Sciences* **113**, E6382–E6390 (2016).
39. Davidson, K. C. *et al.* Wnt/β-catenin signaling promotes differentiation, not self-renewal, of human embryonic stem cells and is repressed by Oct4. *Proceedings of the National Academy of Sciences* **109**, 4485–4490 (2012).
40. Hikasa, H. & Sokol, S. Y. Wnt Signaling in Vertebrate Axis Specification. *Cold Spring Harb Perspect Biol* **5**, a007955 (2013).
41. Logan, C. Y. & Nusse, R. The Wnt signaling pathway in development and disease. *Annu Rev Cell Dev Biol* **20**, 781–810 (2004).
42. Clevers, H. Wnt/beta-catenin signaling in development and disease. *Cell* **127**, 469–480 (2006).
43. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33**, 245–254 (2003).
44. Bird, A. Perceptions of epigenetics. *Nature* **447**, 396–398 (2007).
45. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat Biotechnol* **28**, 1057–1068 (2010).
46. 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 (2019).
47. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**, 23–38 (2013).
48. Shi, D.-Q. Q., Ali, I., Tang, J. & Yang, W.-C. C. New Insights into 5hmC DNA Modification: Generation, Distribution and Function. *Frontiers in Genetics* **8**, (2017).
49. Song, C.-X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends in Biochemical Sciences* **38**, 480–484 (2013).
50. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
51. Yao, Q., Chen, Y. & Zhou, X. The roles of microRNAs in epigenetic regulation. *Current Opinion in Chemical Biology* **51**, 11–17 (2019).

52. Wang, C. *et al.* LncRNA Structural Characteristics in Epigenetic Regulation. *International Journal of Molecular Sciences* **18**, 2659 (2017).
53. Francis, N. J. & Sihou, D. Inheritance of Histone (H3/H4): A Binary Choice? *Trends in Biochemical Sciences* **46**, 5–14 (2021).
54. Fitz-James, M. H. & Cavalli, G. Molecular mechanisms of transgenerational epigenetic inheritance. *Nat Rev Genet* **23**, 325–341 (2022).
55. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21**, 381–395 (2011).
56. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165**, 1012–1026 (2016).
57. Xia, W. *et al.* Resetting histone modifications during human parental-to-zygotic transition. *Science* **365**, 353–360 (2019).
58. Xu, R., Li, C., Liu, X. & Gao, S. Insights into epigenetic patterns in mammalian early embryos. *Protein Cell* **12**, 7–28 (2021).
59. Inoue, A., Chen, Z., Yin, Q. & Zhang, Y. Maternal Eed knockout causes loss of H3K27me3 imprinting and random X inactivation in the extraembryonic cells. *Genes Dev* **32**, 1525–1536 (2018).
60. Saha, B. *et al.* EED and KDM6B Coordinate the First Mammalian Cell Lineage Commitment To Ensure Embryo Implantation. *Molecular and Cellular Biology* **33**, 2691–2705 (2013).
61. Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620–631 (2018).
62. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
63. Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).
64. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**, 204–220 (2013).

65. Pan, G. *et al.* Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell* **1**, 299–312 (2007).
66. Voigt, P., Tee, W.-W. W. W. & Reinberg, D. A double take on bivalent promoters. *Genes and Development* **27**, 1318–1338 (2013).
67. Dunican, D. S. *et al.* Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells. *BMC Biol* **18**, 25 (2020).
68. Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome Res.* **31**, 2170–2184 (2021).
69. Bernhart, S. H. *et al.* Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Sci Rep* **6**, 37393 (2016).
70. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).
71. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
72. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**, 21–32 (2001).
73. Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nat Rev Genet* **15**, 517–530 (2014).
74. Õunap, K. Silver-Russell Syndrome and Beckwith-Wiedemann Syndrome: Opposite Phenotypes with Heterogeneous Molecular Etiology. *MSY* **7**, 110–121 (2016).
75. Van de Pette, M. *et al.* Epigenetic changes induced by in utero dietary challenge result in phenotypic variability in successive generations of mice. *Nat Commun* **13**, 2464 (2022).
76. Nicholls, R. D. & Knepper, J. L. Genome Organization, Function, and Imprinting in Prader-Willi and Angelman Syndromes. *Annual Review of Genomics and Human Genetics* **2**, 153–175 (2001).
77. Galupa, R. & Heard, E. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annual Review of Genetics* **52**, 535–566 (2018).

78. Fang, H., Disteche, C. M. & Berletch, J. B. X Inactivation and Escape: Epigenetic and Structural Features. *Frontiers in Cell and Developmental Biology* **7**, (2019).
79. van den Berg, I. M. *et al.* X Chromosome Inactivation Is Initiated in Human Preimplantation Embryos. *The American Journal of Human Genetics* **84**, 771–779 (2009).
80. Patrat, C., Ouimette, J.-F. & Rougeulle, C. X chromosome inactivation in human development. *Development* **147**, dev183095 (2020).
81. Migeon, B. R. X-linked diseases: susceptible females. *Genet Med* **22**, 1156–1174 (2020).
82. Vallot, C. *et al.* XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. *Nat Genet* **45**, 239–241 (2013).
83. Casanova, M. *et al.* A primate-specific retroviral enhancer wires the XACT lncRNA into the core pluripotency network in humans. *Nat Commun* **10**, 5652 (2019).
84. Bousard, A. *et al.* The role of Xist-mediated Polycomb recruitment in the initiation of X-chromosome inactivation. *EMBO reports* **20**, e48019 (2019).
85. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* **21**, 1592–1600 (2011).
86. Pintacuda, G. & Cerase, A. X Inactivation Lessons from Differentiating Mouse Embryonic Stem Cells. *Stem Cell Rev and Rep* **11**, 699–705 (2015).
87. Sahakyan, A. *et al.* Human Naive Pluripotent Stem Cells Model X Chromosome Dampening and X Inactivation. *Cell Stem Cell* **20**, 87–101 (2017).
88. Sarel-Gallily, R. & Benvenisty, N. Large-Scale Analysis of X Inactivation Variations between Primed and Naïve Human Embryonic Stem Cells. *Cells* **11**, 1729 (2022).
89. Mekhoubad, S. *et al.* Erosion of dosage compensation impacts human iPSC disease modeling. *Cell Stem Cell* **10**, 595–609 (2012).
90. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628 (2008).
91. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).

92. 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. *Nature Methods* **10**, 1213–1218 (2013).

93. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **6**, 468–481 (2011).

94. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14**, e8124 (2018).

95. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 111 (2020).

96. Velten, B. *et al.* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat Methods* **19**, 179–186 (2022).

97. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216 (2012).

98. Guo, G. *et al.* Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. *Stem Cell Reports* **6**, 437–446 (2016).

99. Theunissen, T. W. *et al.* Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. *Cell Stem Cell* **15**, 471–487 (2014).

100. Pastor, W. A. *et al.* Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. *Cell Stem Cell* **18**, 323–329 (2016).

101. Ishihara, T., Griffith, O. W., Suzuki, S. & Renfree, M. B. Presence of H3K4me3 on Paternally Expressed Genes of the Paternal Genome From Sperm to Implantation. *Frontiers in Cell and Developmental Biology* **10**, (2022).

102. 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 (2010).

103. Mawaribuchi, S., Aiki, Y., Ikeda, N. & Ito, Y. mRNA and miRNA expression profiles in an ectoderm-biased substate of human pluripotent stem cells. *Sci Rep* **9**, 11910 (2019).

104. McLeay, R. C., Lesluyes, T., Cuellar Partida, G. & Bailey, T. L. Genome-wide in silico prediction of gene expression. *Bioinformatics* **28**, 2789–2796 (2012).

105. Angeloni, A. & Bogdanovic, O. Sequence determinants, function, and evolution of CpG islands. *Biochemical Society Transactions* **49**, 1109–1119 (2021).

106. Eckersley-Maslin, M. A. *et al.* Epigenetic priming by Dppa2 and 4 in pluripotency facilitates multi-lineage commitment. *Nat Struct Mol Biol* **27**, 696–705 (2020).

107. Regulation, functions and transmission of bivalent chromatin during mammalian development. *Nature Reviews Molecular Cell Biology* doi:10.1038/s41580-022-00518-2.

108. Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics* **36**, 118–131 (2020).

109. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).

110. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat Commun* **12**, 4344 (2021).

111. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* (2019) doi:10.1038/s41576-019-0173-8.

112. Sahakyan, A., Yang, Y. & Plath, K. The Role of Xist in X-Chromosome Dosage Compensation. *Trends in Cell Biology* **28**, 999–1013 (2018).

113. Chen, C.-K. *et al.* Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science* **354**, 468–472 (2016).

114. Ghosh, A. & Som, A. Decoding molecular markers and transcriptional circuitry of naive and primed states of human pluripotency. *Stem Cell Research* **53**, 102334 (2021).

115. Henikoff, S. & Greally, J. M. Epigenetics, cellular memory and gene regulation. *Current Biology* **26**, R644–R648 (2016).

116. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489–499 (2019).

117. Kumar, B. *et al.* Polycomb repressive complex 2 shields naïve human pluripotent cells from trophectoderm differentiation. *Nat Cell Biol* **24**, 845–857 (2022).

118. Tapia, N. *et al.* Dissecting the role of distinct OCT4-SOX2 heterodimer configurations in pluripotency. *Sci Rep* **5**, 13533 (2015).

119. Wang, C.-Y., Froberg, J. E., Blum, R., Jeon, Y. & Lee, J. T. Comment on "Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing". *Science* **356**, eaal4976 (2017).

120. Nesterova, T. B. *et al.* Systematic allelic analysis defines the interplay of key pathways in X chromosome inactivation. *Nat Commun* **10**, 3129 (2019).

121. Rostovskaya, M. Maintenance of Human Naïve Pluripotent Stem Cells. in *Human Naïve Pluripotent Stem Cells* (ed. Rugg-Gunn, P.) 73–90 (Springer US, 2022). doi:10.1007/978-1-0716-1908-7_6.

122. Rostovskaya, M. Capacitation of Human Naïve Pluripotent Stem Cells. in *Human Naïve Pluripotent Stem Cells* (ed. Rugg-Gunn, P.) 117–131 (Springer US, 2022). doi:10.1007/978-1-0716-1908-7_9.

123. Sumi, T., Oki, S., Kitajima, K. & Meno, C. Epiblast Ground State Is Controlled by Canonical Wnt/β-Catenin Signaling in the Postimplantation Mouse Embryo and Epiblast Stem Cells. *PLoS One* **8**, e63378 (2013).

124. von Meyenn, F. *et al.* Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification. *Developmental Cell* **39**, 104–115 (2016).

125. Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research* **40**, e136 (2012).

126. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817–820 (2014).

127. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

128. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**, 276-288.e8 (2018).

129. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* **9**, 9354 (2019).

130. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**, 276–278 (2020).

131. 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 (2017).

132. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

133. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

134. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

135. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).

136. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

137. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**, e1003118 (2013).

138. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13**, R87 (2012).

139. Kumar, L. & E. Futschik, M. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* **2**, 5–7 (2007).

140. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).

141. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12**, 2478–2492 (2017).

142. Babraham Bioinformatics - SeqMonk Mapped Sequence Analysis Tool. https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/.

143. Pagès, H. & Maduka), P. C. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. (2023) doi:10.18129/B9.bioc.BSgenome.

144. Canty, A. & support), B. R. (author of parallel. boot: Bootstrap Functions (Originally by Angelo Canty for S). (2022).

145. Grömping, U. Relative Importance for Linear Regression in *R* : The Package **relaimpo**. *J. Stat. Soft.* **17**, (2006).

146. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

147. org.Hs.eg.db. *Bioconductor* http://bioconductor.org/packages/org.Hs.eg.db/.

148. Krueger, F. & Andrews, S. R. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res* **5**, 1479 (2016).

149. Ma, W., Noble, W. S. & Bailey, T. L. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat Protoc* **9**, 1428–1450 (2014).

150. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Research* **43**, W39–W49 (2015).

151. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165–D173 (2022).

152. Jeong, H.-H., Yalamanchili, H. K., Guo, C., Shulman, J. M. & Liu, Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. in *Biocomputing 2018* 168–179 (WORLD SCIENTIFIC, 2017). doi:10.1142/9789813235533_0016.

153. Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. (2023) doi:10.18129/B9.bioc.AnnotationDbi.

154. Akulenko, R., Merl, M. & Helms, V. BEclear: Batch Effect Detection and Adjustment in DNA Methylation Data. *PLOS ONE* **11**, e0159921 (2016).

155. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115–121 (2015).

156. BSgenome.Hsapiens.UCSC.hg38. *Bioconductor* http://bioconductor.org/packages/BSgenome.Hsapiens.UCSC.hg38/.

157. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13**, R83 (2012).

158. Wang, Q. *et al.* Exploring Epigenomic Datasets by ChIPseeker. *Current Protocols* **2**, e585 (2022).

159. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

160. Zeileis, A. *et al.* colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software* **96**, 1–49 (2020).

161. Gu, Z. Complex heatmap visualization. *iMeta* **1**, e43 (2022).

162. Wickham, H. & RStudio. conflicted: An Alternative Conflict Resolution Strategy. (2021).

163. Dowle, M. *et al.* data.table: Extension of ‘data.frame’. (2022).

164. Report of DEG analysis. https://lpantano.github.io/DEGreport/.

165. Reyes, A. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences* **110**, 15377–15382 (2013).

166. Korthauer, K., Chakraborty, S., Benjamini, Y. & Irizarry, R. A. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics* **20**, 367–383 (2019).

167. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. Preprint at https://doi.org/10.12688/f1000research.8900.2 (2016).

168. Gu, Z., Eils, R., Schlesner, M. & Ishaque, N. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics* **19**, 234 (2018).

169. Rainer, J., Gatto, L. & Weichenberger, C. X. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* **35**, 3151–3153 (2019).

170. Apply Mapping Functions in Parallel using Futures. https://furrr.futureverse.org/index.html.

171. Bengtsson, H. A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal* **13**, 273–291 (2021).

172. Brunson, J. C. ggalluvial: Layered Grammar for Alluvial Plots. *Journal of Open Source Software* **5**, 2017 (2020).

173. Tang, Y., Horikoshi, M. & Li, W. ggfortify: Unified Interface to Visualize Statistical Results of Popular R Packages. *The R Journal* **8**, 474–485 (2016).

174. Create Elegant Data Visualisations Using the Grammar of Graphics. https://ggplot2.tidyverse.org/.

175. Arnold (<https://orcid.org/0000-0001-9953-3904>), J. B. *et al.* ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. (2021).

176. Auguie, B. & Antonov, A. gridExtra: Miscellaneous Functions for 'Grid' Graphics. (2017).

177. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).

178. Firke, S. *et al.* janitor: Simple Tools for Examining and Cleaning Dirty Data. (2021).

179. magrittr), S. M. B. (Original author and creator of, Wickham, H., Henry, L. & RStudio. magrittr: A Forward-Pipe Operator for R. (2022).

180. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316–319 (2017).

181. Patel, H., Ewels, P. & Peltzer, A. nf-core/atacseq: nf-core/atacseq v1.1.0 - Iron Shark. (2019) doi:10.5281/zenodo.3529420.

182. Wang, C. *et al.* nf-core/chipseq: nf-core/chipseq v1.1.0 - Platinum Pig. (2019) doi:10.5281/zenodo.3529400.

183. Ewels, P. *et al.* nf-core/methylseq: nf-core/methylseq version 1.4 [Mercury Rattlesnake]. (2019) doi:10.5281/zenodo.3548445.

184. Ewels, P. *et al.* nf-core/rnaseq: nf-core/rnaseq version 1.4.2. (2019) doi:10.5281/zenodo.3503887.

185. Bengtsson, H. parallelly: Enhancing the ‘parallel’ Package. (2022).

186. Pedersen, T. L. patchwork: The Composer of Plots. (2022).

187. Kolde, R. pheatmap: Pretty Heatmaps. (2019).

188. Carroll, T. & Barrows, D. profileplyr: Visualization and annotation of read signal over genomic ranges with profileplyr. (2023) doi:10.18129/B9.bioc.profileplyr.

189. Wickham, H., Henry, L. & RStudio. purrr: Functional Programming Tools. (2023).

190. R: The R Project for Statistical Computing. https://www.r-project.org/.

191. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2022).

192. Wickham, H. Reshaping Data with the reshape Package. *Journal of Statistical Software* **21**, 1–20 (2007).

193. Kalinowski, T. *et al.* reticulate: Interface to ‘Python’. (2023).

194. Ren, K. rlist: A Toolbox for Non-Tabular Data Manipulation. (2021).

195. RStudio Team. *RStudio: Integrated Development Environment for R.* (RStudio, PBC., 2020).

196. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

197. Wickham, H., Seidel, D. & RStudio. scales: Scale Functions for Visualization. (2022).

198. Wickham, H. & RStudio. stringr: Simple, Consistent Wrappers for Common String Operations. (2022).

199. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

200. TxDb.Hsapiens.UCSC.hg38.knownGene. *Bioconductor* http://bioconductor.org/packages/TxDb.Hsapiens.UCSC.hg38.knownGene/.

201. Sloutskin, A., Shir-Shapira, H., Freiman, R. N. & Juven-Gershon, T. The Core Promoter Is a Regulatory Hub for Developmental Gene Expression. *Frontiers in Cell and Developmental Biology* **9**, (2021).

202. Fritsch, E. F., Lawn, R. M. & Maniatis, T. Molecular cloning and characterization of the human β-like globin gene cluster. *Cell* **19**, 959–972 (1980).

203. deBoer, E., Antoniou, M., Mignotte, V., Wall, L. & Grosveld, F. The human beta-globin promoter; nuclear protein factors and erythroid specific induction of transcription. *The EMBO Journal* **7**, 4203–4212 (1988).

204. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biology* **22**, 108 (2021).

205. Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077–3086 (2002).

206. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725–1735 (2003).

207. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *Journal of Molecular Biology* **196**, 261–282 (1987).

208. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & Development* **25**, 1010–1022 (2011).

209. Jones, P. A. & Laird, P. W. Cancer-epigenetics comes of age. *Nat Genet* **21**, 163–167 (1999).

210. Hartono, S. R., Korf, I. F. & Chédin, F. GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. *Nucleic Acids Research* gkv811 (2015) doi:10.1093/nar/gkv811.

211. Li, W.-C. *et al.* Sequence analysis of origins of replication in the Saccharomyces cerevisiae genomes. *Frontiers in Microbiology* **5**, (2014).

212. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol* **21**, 459–474 (2020).

213. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development* **9**, 657–663 (1999).

214. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).

215. Wang, H. *et al.* SVA Elements: A Hominid-specific Retroposon Family. *Journal of Molecular Biology* **354**, 994–1007 (2005).

216. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics* **54**, 539–561 (2020).

217. Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. *Nat Rev Genet* **20**, 760–772 (2019).

218. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397–405 (2008).

219. Bonder, M. J. *et al.* Single cell DNA methylation ageing in mouse blood. 2023.01.30.526343 Preprint at https://doi.org/10.1101/2023.01.30.526343 (2023).

220. Galle, E. *et al.* H3K18 lactylation marks tissue-specific active enhancers. *Genome Biology* **23**, 207 (2022).

Appendix A – Description of genomic features

Gene promoters

A gene promoter is a region of DNA located upstream of a gene that has a significant role in initiating transcription. Promoters contain specific nucleotide sequences that serve as binding sites for the RNA polymerase and other transcription factors, which help initiate transcription. One important feature of gene promoters is their variability in sequence and structure, allowing them to regulate gene expression in response to various environmental cues and developmental signals[201]. For example, some promoters contain enhancer elements that can increase or decrease gene expression depending on their interaction with regulatory proteins. deBoer et al. (1988) showed that the beta-globin promoter contains several distinct functional elements, including a TATA box and a CAAT box, that are conserved across species and play a critical role in transcription initiation[202,203].

Enhancers

Enhancers are regulatory elements in DNA that control gene expression by increasing or decreasing the transcription of target genes. They are typically located far from the genes they regulate and can be located upstream, downstream or even within an intron of a gene. Enhancers work by binding transcription factors, which are proteins that can either activate or repress gene expression. These transcription factors can recruit other proteins to help modify the chromatin structure, allowing the RNA polymerase to access the promoter region and initiate transcription[204]. For example, the β-globin gene is regulated by a complex enhancer located 6-22 kb upstream of the transcription start site. This enhancer contains binding sites for several transcription factors, including GATA1, NF-E2, and TAL1, and is essential for proper expression of the β-globin gene[205]. Another example is the Sonic hedgehog (Shh) gene, which is regulated by a conserved enhancer called ZRS (ZPA regulatory sequence). ZRS is located 800 kb upstream of the Shh gene and contains binding sites for several transcription factors, including GLI3 and HAND2. Mutations in the ZRS enhancer have been associated with limb malformations in humans[206]. In summary, enhancers are essential for appropriate development and homeostasis, and their misregulation can lead to developmental defects and disease.

CpG islands

CpG islands are regions of DNA that contain a high frequency of cytosine (C) and guanine (G) nucleotides linked by a phosphate group (p). They are typically located in the promoter regions of genes and are often associated with gene expression regulation[207]. CpG islands are usually at least 200 base pairs long, with a high GC content (greater than 50%). Although most CpG sites in the human genome are methylated, CpG islands are usually unmethylated in normal cells, allowing for the proper regulation of gene expression[208]. Methylation of CpG islands has been linked to several diseases, including cancer and neurological disorders. For example, in some types of cancer, CpG islands in tumour suppressor genes can become hypermethylated, leading to the silencing of these genes and promoting cancer development[209].

GC skew

GC skew is a measure of the asymmetry in the distribution of the nucleotides guanine (G) and cytosine (C) on the leading and lagging strands of DNA. It is calculated using the formula [(G - C) / (G + C)], where G and C represent the number of occurrences of those nucleotides at a particular position in the DNA sequence. A positive GC skew indicates that there are more Gs than Cs in the leading strand, while a negative GC skew indicates that there are more Cs than Gs in the leading strand[210].

One of its primary roles is in identifying the direction and location of replication origins during DNA replication. This is due to the difference in synthesis rates between the leading and lagging strands, which results in a distinct GC content in each strand[211]. Another important role of GC skew is in predicting regions of the genome that are prone to mutations. Regions with high GC content, particularly CpG dinucleotides, are more susceptible to mutations that can cause genetic diseases like cancer. Moreover, GC skew is associated with the formation of G-quadruplexes. G-quadruplexes are secondary DNA structures formed in guanine-rich sequences. These structures are associated with gene regulation, telomere maintenance, and DNA replication. Regions of the genome with high GC skew are more likely to contain G-quadruplexes, making GC skew a tool for predicting the locations of G-quadruplexes in the genome. However, G-quadruplexes can also contribute to genomic instability, particularly when they occur in regions of DNA where they can impede replication or transcription[212]. This instability can lead to DNA damage, genetic mutations, and the development of diseases like cancer.

Repetitive DNA sequences

The human genome contains different types of repetitive DNA sequences. These DNA segments are present in multiple copies throughout the genome and make up a significant portion of the genome. They play various roles in genetic processes such as gene expression, DNA replication, and genome stability. The main types of repeats found in the human genome are: Satellite DNA, consisting of short DNA sequences (usually less than 100 base pairs) repeated in tandem arrays with varying lengths from several kilobases to megabases[213]. This type of repeat is found in the centromeres and telomeres of chromosomes, where they are involved in chromosome structure and segregation during cell division; Simple sequence repeats (SSRs), also known as microsatellites, consisting of short tandem repeats (STRs) of 1-6 nucleotide units scattered throughout the genome. These sequences can vary in length between individuals, thus, being commonly used in genetic studies as markers for disease diagnosis, forensic analysis, and population genetics; Interspersed repeats, are DNA sequences that are interspersed throughout the genome and can be further divided into long interspersed nuclear elements (LINEs), which are the most abundant interspersed repeats in the human genome, making up about 20% of the genome[214]. LINEs are autonomous retrotransposons that can copy and paste themselves into new locations in the genome, potentially causing mutations and genetic diversity; and short interspersed nuclear elements (SINEs), which are non-autonomous retrotransposons that require the help of LINEs to transpose. SINEs are about 300 base pairs in length and comprise about 13% of the human genome. The most common SINE in humans is the Alu element; DNA transposons: DNA transposons are a type of mobile genetic element that can move within the genome through a "cut-and-paste" mechanism. These sequences are relatively rare in the human genome, making up less than 3% of the genome.

A SINE with a relevant role in naïve pluripotency is the SVA (SINE-VNTR-Alu). It is a composite interspersed repeat containing three elements: a short interspersed nuclear element (SINE), a variable number tandem repeat (VNTR), and an Alu element. SVAs are relatively rare, comprising less than 0.2% of the human genome, but are thought to be involved in gene regulation and evolution. The SINE component of SVA is derived from an Alu element, a primate-specific retrotransposon[215].

Transposable elements

Transposable elements (TEs) are genetic sequences that can move, or transpose, within a genome, causing mutations and increasing genetic diversity. TEs are found in the genomes of all living organisms, from bacteria to humans, and constitute a significant proportion of many genomes, including the human genome. There are two main types of TEs: DNA transposons and retrotransposons[216]. DNA transposons move by a "cut-and-paste" mechanism, where the transposon is excised from its original location and inserted into a new location in the genome. Conversely, retrotransposons move by a "copy-and-paste" mechanism where the transposon is first transcribed into RNA, then reverse-transcribed into DNA, and then inserted into a new location in the genome.

TEs can have significant impacts on the genome, both positive and negative[217]. TEs can contribute to genetic diversity as they are transposed into new genomic locations and provide novel genetic functions. On the other hand, TEs can also cause deleterious mutations or disrupt gene expression and regulatory elements, leading to diseases or disorders.

Some notable examples of TEs in the human genome include Alu, LINE-1 retrotransposons, and the human endogenous retroviruses (HERVs), which are remnants of ancient retroviral infections. It has been suggested that this may play a role in the evolution of the genome, as it can provide the raw material for the creation of new genes and regulatory elements. However, the exact mechanisms by which TEs contribute to genome evolution are still not fully understood[218].

Appendix B – Abbreviations

5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
ATAC	Assay for transposase accessible chromatin
ATAC-seq	Assay for transposase accessible chromatin using sequencing
BIC	Bayesian Information Criterion
BWA	Burrows-Wheeler aligner
C	Cytosine
cDNA	Complementary DNA
CGI	CpG island
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation sequencing
CpG	5'–C–phosphate–G–3' DNA sites
cR-H9-EOS	Chemically reset H9-EOS cell line
CSPh	Cellules souches pluripotentes humaines
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNase-seq	DNase I hypersensitive sites sequencing
DNMT	DNA methyltransferase
EOS	Early Transposon promoter and OCT4 (POU5F1) and SOX2 enhancers lentiviral vector
FDR	False discovery rate
FGF	Fibroblast growth factor
FISH	Fluorescence in situ hybridization
G	Guanine
GEO	Gene expression omnibus
GO	Gene ontology
GRCh38	Genome reference consortium human build 38
H3K27ac	Histone H3 lysine 27 acetylation
H3K27me3	Histone H3 lysine 27 trimethylation
H3K4me1	Histone H3 lysine 4 monomethylation
H3K4me3	Histone H3 lysine 4 trimethylation
H3K9me3	Histone H3 lysine 9 trimethylation
HEFA	Human embryology and fertilisation authority
HERVs	human endogenous retroviruses

hESC	Human embryonic stem cell
HMM	Hidden Markov model
hPSC	Human pluripotent stem cell
ICM	Inner cell mass
ICR	Imprinting control region
Kb	Kilobase
KEGG	Kyoto encyclopedia of genes and genomes
Lads	Lamina-associated domains
LIF	Leukemia inhibitory factor
LINE	Long interspersed nuclear element
lncRNA	Long non-coding RNA
LOLA	Locus overlap analysis
LTR	Long terminal repeat
MACS	Model-based analysis of ChIP-seq
Mb	Megabase
MEFs	Mouse embryonic fibroblasts
mESCs	Mouse embryonic stem cells
MOFA	Multi-omic factor analysis
O/E	Observed to expected
PBAT	Post-bisulfite adaptor-tagging
PCA	Principal component analysis
PcG	Polycomb-group proteins
PKC	Protein kinase C
PMF	Probabilistic matrix factorization
PRC2	Polycomb repressive complex 2
PWS	Prader-Willi syndrome
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
Shh	Sonic hedgehog
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
STR	Short tandem repeat
SVA	SINE-VNTR-Alu
t-SNE	t-distributed stochastic neighbour embedding
TE	Transposable element

TES	Transcription end site
TET	Ten-eleven translocation
TrxG	Trithorax-group proteins
TSS	Transcription start site
UCSC	University of California Santa Cruz
UTR	Untranslated region
VNTR	Variable number tandem repeat
WGBS	Whole-genome bisulfite sequencing
XACT	X-active coating transcript
XCI	X chromosome inactivation
XX	Chromosomes X and X
XY	Chromosomes X and Y
ZGA	Zygotic gene activation
ZRS	ZPA regulatory sequence

Appendix C – Code and data availability

Gene expression (RNA sequencing) datasets have been published[10] and are available in the Gene Expression Omnibus (GEO) database under the accession number GSE123055 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE123055). The ChIP-seq, ATAC-seq, and PBAT raw and processed data used in this dissertation have been deposited and are available in GEO under the accession number GSE218512 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE218512; secure token for reviewers: yzovuskurzaxfcd). All scripts required to reproduce the figures are available at https://github.com/vonMeyennLab/hPSC_epigenetic_dynamics.

Appendix D – Scientific Contributions and Collaborations

Parts of this dissertation were performed in collaboration with colleagues, research groups, and supported by core facilities. Those parts and the researchers involved are described below:

- Cell culture, which included hPSC (human pluripotent stem cell) maintenance and capacitation (see methods), was conducted by collaborators from the Austin Smith laboratory (University of Exeter, Exeter, United Kingdom) and the Wolf Reik laboratory (Babraham Institute, Cambridge, United Kingdom).

- Data preparation–which included chromatin immunoprecipitation (ChIP), assay for transposase accessible chromatin (ATAC), and whole-genome bisulfite (WGBS) library preparation and sequencing–was also conducted by collaborators from the Austin Smith and Wolf Reik laboratories, with the help of Maike Paramor and Vicki Murray from the Stem Cell Institute Genomics Facility.

- RNA sequencing and library preparation were performed by members of the Austin Smith laboratory. The methodology can be consulted in the Rostovskaya et al. 2019 article[10].

- The western blots were conducted by Chee-Wai Wong from the Laboratory of Nutrition and Metabolic Epigenetics at ETH Zürich, Switzerland.

- Data processing, such as sequencing quality control, read alignment, read counts, and methylation percentage in CpGs, was performed by me. I was also responsible for the data analysis and for all figures in this dissertation (except the figures with the western blot bands and microscopy). Furthermore, Simon Andrews and Felix Krüger from the bioinformatics facility at the Babraham Institute (Cambridge, United Kingdom) helped me with data mapping and handling and provided advice on how to analyse multi-omics data. Jason Ernst from UCLA (University of California, Los Angeles) provided helpful tips regarding the tool chromHMM[141].

Appendix E – Funding

The work presented in this dissertation was supported by MRC UK Regenerative Medicine Platform (UKRMP, MR/L012537/1); MRC Pluripotent stem cells and engineered cells Hub (PSEC, MR/R015724/1); King’s Prize Fellowship (King's Health Partners, Wellcome Trust and London Law Trust to Ferdinand von Meyenn); MRC (MR/S005536/1 to Ferdinand von Meyenn); ETH Zurich (core funding to Ferdinand von Meyenn); MRC (MR/V02969X/1 to Maria Rostovskaya); BBSRC (BBS/E/B/000C0421, BBS/E/B/000C0422, Core Capability Grant to Wolf Reik), Wellcome Investigator Award (210754/Z/18/Z to Wolf Reik).

Appendix F – Publications

- João Agostinho de Sousa, Maria Rostovskaya, Chee-Wai Wong, Wolf Reik, Austin Smith, Ferdinand von Meyenn. **Epigenetic dynamics during capacitation of naïve human pluripotent stem cells**. (In revision at the journal "Science Advances").

  Abstract:

  Human pluripotent stem cells (hPSCs) are of fundamental relevance in regenerative medicine and the primary source for many novel cellular therapies. The development of naïve culture conditions has led to the expectation that these naïve hPSCs could overcome some of the limitations found in conventional (primed) hPSCs culture conditions, including recurrent epigenetic anomalies. Recent work has shown that transition to the primed state (or capacitation) is necessary for naïve hPSCs to acquire multi-lineage differentiation competence. This pluripotent state transition may recapitulate essential features of human peri-implantation development. Here we studied epigenetic changes during the transition between naïve and primed pluripotency, examining global genomic redistribution of histone modifications, chromatin accessibility, and DNA methylation, and correlating these with gene expression. We identify CpG islands, enhancers, and retrotransposons as hotspots of epigenetic dynamics between pluripotency states. Our results further reveal that hPSC resetting and subsequent capacitation rescue X chromosome-linked epigenetic erosion and reduce the ectoderm-biased gene expression of conventional primed hPSCs.

- Marc Jan Bonder, Stephen Clark, Felix Krüger, Siyuan Luo, João Agostinho de Sousa, Aida M. Hashtroud, Thomas M. Stubbs, Anne-Katrien Stark, Steffen Rulands, Oliver Stegle, Wolf Reik, Ferdinand von Meyenn. **Single cell DNA methylation ageing in mouse blood**. (Preprint published in bioRxiv and manuscript submitted to the journal "Nature Ageing")[219].

  Abstract:

  Ageing is the accumulation of changes and overall decline of the function of cells, organs and organisms over time. At the molecular and cellular level, the concept of biological age has been established and biomarkers of biological age have been identified, notably epigenetic DNA-methylation based clocks. With the emergence of single-cell DNA methylation profiling methods, the possibility to study biological age of individual cells has been proposed, and a first proof-of-concept study, based on limited single cell datasets mostly from early developmental origin, indicated the feasibility and relevance of this approach to better understand organismal changes and cellular ageing heterogeneity.

Here we generated a large single-cell DNA methylation and matched transcriptome dataset from mouse peripheral blood samples, spanning a broad range of ages (10-101 weeks of age). We observed that the number of genes expressed increased at older ages, but gene specific changes were small. We next developed a robust single cell DNA methylation age predictor (scEpiAge), which can accurately predict age in a broad range of publicly available datasets, including very sparse data and it also predicts age in single cells. Interestingly, the DNA methylation age distribution is wider than technically expected in 19% of single cells, suggesting that epigenetic age heterogeneity is present *in vivo* and may relate to functional differences between cells. In addition, we observe differences in epigenetic ageing between the major blood cell types. Our work provides a foundation for better single-cell and sparse data epigenetic age predictors and highlights the significance of cellular heterogeneity during ageing.

- Anna B Rueegg, Vera B van der Weijden, Joao A de Sousa, Ferdinand von Meyenn, Hubert Pausch, Susanne Ulbrich. **Slowed, but not stopped: developmental progression continues during embryonic diapause in the roe deer**. (Submitted to the journal “Nature Communications Biology”).

  Abstract:

  Embryonic diapause in mammals is a temporary developmental delay occurring at the blastocyst stage. In contrast to other diapausing species that display a full halt in development, the blastocyst of the European roe deer (Capreolus capreolus) continuously proliferates over a period of 4-5 months. Meanwhile, the inner cell mass undergoes notable morphological changes. We hypothesized that not only embryonic cell proliferation, but also developmental progression continues. In embryos across diapause and elongation, we evaluated the mRNA transcript abundance of genes important for early developmental processes, such as axis formation and gastrulation. We performed immunohistochemical stainings on individual embryos to support our findings. Morphological rearrangements of the epiblast during diapause overlapped with gene expression patterns and changes in cell polarity, as indicated by the localization of f-actin. Our results indicate that primitive endoderm formation occurred during diapause in embryos composed of around 3’000 cells. Gastrulation likely coincided with elongation and thus proceeded after embryo reactivation. We thus evidence developmental progression at slow pace during roe deer diapause. Studying the latter may unravelling key requirements for embryo survival, the dispensability of maternal signals, as well as the link between proliferation and differentiation.

- Galle E, Wong CW, Ghosh A, Desgeorges T, Melrose K, Hinte LC, Castellano-Castillo D, Engl M, de Sousa JA, Ruiz-Ojeda FJ, De Bock K. **H3K18 lactylation marks tissue-specific active enhancers**. Genome biology. 2022 Dec;23(1):1-28[220].

  Abstract:

  Histone lactylation has been recently described as a novel histone posttranslational modification linking cellular metabolism to epigenetic regulation. Given the expected relevance of this modification and current limited knowledge of its function, we generate genome-wide datasets of H3K18la distribution in various *in vitro* and *in vivo* samples, including mouse embryonic stem cells, macrophages, adipocytes, and mouse and human skeletal muscle. We compare them to profiles of well-established histone modifications and gene expression patterns. Supervised and unsupervised bioinformatics analysis shows that global H3K18la distribution resembles H3K27ac, although we also find notable differences. H3K18la marks active CpG island containing promoters of highly expressed genes across most tissues assessed, including many housekeeping genes, and positively correlates with H3K27ac and H3K4me3 as well as with gene expression. In addition, H3K18la is enriched at active enhancers that lie in proximity to genes that are functionally important for the respective tissue. Overall, our data suggests that H3K18la is not only a marker for active promoters, but also a mark of tissue specific active enhancers.

Appendix G – Curriculum Vitae

João Agostinho de Sousa

Uetlibergstrasse 111b, 8045 Zürich, Switzerland
+41775279473 jpadesousa@gmail.com
www.linkedin.com/in/joao-agostinhodesousa
https://github.com/jpadesousa

PRACTICAL EXPERIENCE

Bioinformatician and stem cell researcher *ETH Zürich, Switzerland* Aug 2019 – Present

I collaborated on different research projects with a focus on bioinformatics, epigenetics, ageing, and stem cells at the Laboratory of Nutrition and Metabolic Epigenetics.

- Multi-omics data modelling and prediction
- Project management and collaborative research
- Development new next-generation sequencing HPC workflows
- Result interpretation and scientific writing

Bioinformatician and cancer researcher *Instituto de Medicina Molecular, Portugal* Sep 2017 – Oct 2018

I analysed an extensive cancer dataset to test if there is an association between point mutations in cancer primary tissues and enrichment in hydroxymethylcytosine (5hmC, an oxidative derivative of DNA methylation).

- Developed R, Python, and Bash scripts for multi-omics data
- Researched the relationship between epigenetics and cancer
- Analysed genomic data from primary tissue of cancer patients' samples

Concierge *InterContinental Hotels & Resorts *****, Portugal* May 2016 – Oct 2016

Provided a personalised service, travel scheduling, and consulting to the hotel guests.

Cycling and camping traveller *Japan* Nov 2015 – Jan 2016

Travelled around Japan on a bicycle to learn the more about the Japanese culture and language.

Hotel Receptionist *The Elevator Hostel, Portugal* Jun 2014 – Nov 2015

Managed a 60-bed hotel together with a receptionist team.

Photojournalist *Jornal i, Portugal* Apr 2014 – Jun 2014

In this internship, I did editorial photography, interviews, and news coverage. My work was featured on 3 newspaper covers.

Housekeeper *Granada Inn Backpackers, Spain* Aug 2013 – Nov 2013

Cleaning and organisation, tour guide, and Spanish practice.

Freelance photographer *Portugal* Aug 2012 – Jun 2014

Photographed events, music concerts, travel, and editorial for small companies and artists.

COMPUTER SKILLS

R, Python, C, Nextflow, Git	**Programming**
Machine learning (currently learning), RStudio, Visual Studio Code, Jupyter	**Data science**
Slurm, LSF, Linux, Singularity, Docker, Lmod, Conda environments	**Server management**
Adobe Photoshop, Adobe Lightroom	**Image editing**
Autodesk Autocad, Adobe Illustrator	**Technical drawing**

EDUCATION

Bioinformatics and Epigenetics Doctoral candidate — Aug 2019 - Present

ETH Zürich, Switzerland

Doctoral project: “Characterization and modelling of the epigenetic dynamics during the transition from naïve to primed pluripotency”

Biochemistry MSc — Sep 2016 – Oct 2018

Faculdade de Ciências, Universidade de Lisboa, Portugal

Master thesis: “Association of DNA modifications with somatic mutations in cancer”

Civil Engineering BSc — Sep 2007 – Jun 2012

Instituto Superior Técnico, Universidade de Lisboa, Portugal

LANGUAGES

Full professional proficiency	**English**
Native proficiency	**Portuguese**
Professional working proficiency	**Spanish**
Limited working proficiency	**French**
Elementary proficiency (currently learning)	**German**

SELECTED PUBLICATIONS

João Agostinho de Sousa, Maria Rostovskaya, Chee-Wai Wong, Wolf Reik, Austin Smith, Ferdinand von Meyenn. **Epigenetic dynamics during capacitation of naïve human pluripotent stem cells**. (Submitted to the journal “Science Advances”).	Feb 2023
Marc Jan Bonder, Stephen Clark, Felix Krüger, Siyuan Luo, João Agostinho de Sousa, Aida M. Hashtroud, Thomas M. Stubbs, Anne-Katrien Stark, Steffen Rulands, Oliver Stegle, Wolf Reik, Ferdinand von Meyenn. **Single cell DNA methylation ageing in mouse blood**. (Preprint published in bioRxiv and manuscript submitted to the journal “Nature Ageing”).	Feb 2023
Anna B Rueegg, Vera B van der Weijden, Joao A de Sousa, Ferdinand von Meyenn, Hubert Pausch, Susanne Ulbrich. **Slowed, but not stopped: developmental progression continues during embryonic diapause in the roe deer**. (Submitted to the journal “Nature Communications Biology”).	Dec 2022

Galle E, Wong CW, Ghosh A, Desgeorges T, Melrose K, Hinte LC, Castellano-Castillo D, Engl M, de Sousa JA, Ruiz-Ojeda FJ, De Bock K. **H3K18 lactylation marks tissue-specific active enhancers**. Genome biology. 2022 Dec;23(1):1-28.	Oct 2022

INTERESTS / HOBBIES

Bouldering, hiking, running, boxing	**Sports**
Guitar	**Music**
Physics, statistics, engineering, history, biographies	**Reading**

COMMUNICATION

Scientific Presentation *BioMed Retreat 2021, ETH Zürich* Characterization and modelling of the epigenetic dynamics during the transition from naïve to primed pluripotency	Feb 2021
Scientific poster *BioMed Retreat 2019, ETH Zürich* Epigenetic changes during naïve to primed human embryonic stem cell transition	Nov 2019
Seminar presentation *Computational Biology and Bioinformatics Seminars, Instituto de Medicina Molecular* Association of the epigenetic mark 5hmC local density with somatic mutations in cancer	Sep 2018

FURTHER EDUCATION

Course Applied statistical regression *ETH Zürich*	Sep 2022 – Feb 2023
Lecture Series Space Research and Exploration *ETH Zürich*	Sep 2022 – Feb 2023
Course Masterclass in Scientific Writing and Publishing in High-Impact Journals *ETH Zürich*	Dec 2022
Course Molecular Biology Methods *ETH Zürich*	Sep 2022
Symposium Clinics meets Data Science *Comprehensive Cancer Center Zürich, University Hospital Zürich*	Jun 2022
Course Biological Methods for Engineers (Lab techniques) *ETH Zürich*	Nov – Dec 2021
Course Bioinformatics and Next Generation Sequencing *ETH Zürich*	Nov – Dec 2020
Colloquium in Translational Science *ETH Zürich*	May & Dec 2020
Symposium Nuclear and mitochondria instability in health and cancer *Montargil, Portugal*	Jul 2018