DECODING DEVELOPMENTAL EPIGENETICS

Innovation and implementation of single-cell techniques to study epigenetic regulation in embryonic development

Franka Rang

DECODING DEVELOPMENTAL EPIGENETICS

Innovation and implementation of single-cell techniques to study epigenetic regulation in embryonic development

Franka Jolein Rang

Colofon

ISBN/EAN: 978-90-393-7716-1
DOI: https://doi.org/10.33540/2416

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Cover design: Lobke van Aar, lobkevanaar.nl
Layout and design: Anna Bleeker, persoonlijkproefschrift.nl

DECODING DEVELOPMENTAL EPIGENETICS

Innovation and implementation of single-cell techniques to study epigenetic regulation in embryonic development

---

Embryonale Epigenetica Ontcijferen

Innovatie en implementatie van single-cell technologieën
om epigenetische regulatie in embryonale ontwikkeling te bestuderen
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

donderdag 12 september 2024 des ochtends te 10.15 uur

door

Franka Jolein Rang

geboren op 20 december 1992
te Maastricht

Promotoren:

Prof. dr. A. van Oudenaarden
Prof. dr. J.H. Kind

Beoordelingscommissie:

Prof. dr. S. Abeln
Prof. dr. ir. J.P.W.M. Bakkers (voorzitter)
Prof. dr. T. Baubec
Prof. dr. B. van Steensel
Prof. dr. G.J.C. Veenstra

Table of contents

CHAPTER 1

Introduction

Franka J. Rang[1]

1: Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW), University Medical Center Utrecht, Oncode Institute

Epigenetics

The DNA is often referred to as the blueprint of the cell, as it encodes the information for all proteins and non-coding RNAs required for cellular structure and functioning. As such, obtaining and understanding genome sequences has been a central focus within biological research. Over the last decades, science has immensely progressed DNA sequencing technologies, which has led to a rapid increase in the availability and understanding of DNA sequences. A major milestone in this journey was the completion of the Human Genome Project, which provided the first human reference genome[1,2]. Recently, this accomplishment has been extended with the publication of the telomere-to-telomere version, which filled in highly repetitive regions and provides gapless sequences for all somatic chromosomes[3]. Alongside technological advancements, the costs and time of DNA sequencing have rapidly decreased over the years. Currently, reference genome sequences are available for numerous species and sequencing experiments are routinely performed in labs across the world. From DNA sequence, researchers have been able to extend our understanding of a wide variety of topics, including gene structure, genetic diseases, and inheritance. In addition, DNA sequence provides a record of the past and has helped to reconstruct evolutionary trees[4,5], migration patterns of human ancestors[6,7], and ancient viral infections[8].

Despite this wealth of information, DNA sequence alone does not tell the whole story. Our bodies consist of a wide variety of different cell types with vastly different functions. At any moment in time, cells use only part of their genes depending on e.g. their cell type, cell cycle phase and input from their environment. Consequently, a whole range of phenotypes emerges from a single genotype. To achieve this specific regulation of gene expression an additional layer of information is superimposed upon the genome, which we refer to as the epigenome. Amongst others, the epigenome includes chemical modifications of the DNA itself and of the histone octamers around which the DNA is wrapped to form the nucleosome. These epigenetic layers modulate how accessible the underlying genome is to transcription factors and the transcriptional machinery, thus regulating gene expression. In addition, there are features of the chromatin that are not epigenetic modifications per se, but that do greatly impact gene expression and are actively regulated in the cell. Two notable examples of such regulatory features are chromatin accessibility and the 3D organization of the genome. While not technically part of the epigenome, such layers of regulation are in practice often included when discussing epigenetic regulation.

While epigenetic mechanisms play integral roles in essentially all complex biological systems, their importance is especially clear in embryonic development. Upon fertilization of the oocyte by the sperm, the maternal and paternal genomes are joined together to form the complete genetic material of the embryo. At this point, the embryo consists of a single cell with a single copy of its genome. Over the course of development, this totipotent cell will give rise to a multitude of cell types with distinct phenotypes and functions, but with the same genotype (barring a few specific exceptions). It is through epigenetic regulation that lineage-specific

gene expression profiles are initiated and different cell types gradually emerge. In order to understand the process of embryonic development, it is thus paramount to understand the epigenetic mechanisms controlling gene expression, lineage commitment and cell fate.

As most epigenetic mechanisms directly or indirectly influence gene expression, the genome can be partitioned into an active and an inactive fraction based on its transcriptional activity and epigenetic state. Euchromatin refers to active chromatin and is characterized by an open structure, high gene density and high levels of transcription. Heterochromatin, on the other hand, is more tightly packaged, contains fewer genes, low levels of transcription and a higher density of repetitive elements. There is a large number of epigenetic mechanisms and an even larger range of proteins involved in their regulation. However, the majority of the research in this dissertation has focused on properties of heterochromatin. In the next two sections, I will therefore focus on two aspects of heterochromatin that feature most strongly throughout my research: histone post-translational modifications (PTMs) and Lamina-Associated Domains (LADs). In addition, I will expand on the potential of single-cell sequencing technologies to further our understanding of these processes.

Histone Post-Translational Modifications (PTMs)

The basic structural unit of eukaryotic chromatin is the nucleosome, which is formed by an octamer of histones with ~147 bp of DNA wrapped around it[9,10]. Each nucleosome contains two copies of the histones H3, H4, H2A and H2B. Several versions of these histones exist that can be exchanged based on biological context, such as the presence of DNA damage and formation of constitutive heterochromatin[11]. Moreover, histones can be post-translationally modified, which is a prominent layer of epigenetic regulation. Histone PTMs play important roles in a variety of cellular processes, including gene expression regulation, DNA damage response, recombination, and replication[12]. They exert their influence by serving as a binding platform for downstream effector proteins ("readers") and by directly modulating the biophysical properties of the chromatin fiber[11]. The presence of these marks is regulated by histone modifying enzymes that lay down these marks ("writers") and remove them ("erasers"). Given the involvement of histone PTMs in such crucial events, it is not surprising that histones and histone modifying enzymes are tightly regulated during development[11,13] and are often found to be dysregulated or mutated in cancer[11,14].

Many different histone modifications have been discovered to date, including methylation, acetylation, ubiquitination, SUMOylation, and others[12]. These modifications can be laid down on the histone core and on the amino acid tails that extend out of the core. The most studied histone PTMs include the acetylation and mono-, di-, and trimethylation of lysine residues of the histone tail. Some examples of euchromatic histone PTMs in mouse and human include H3 Lysine-4 trimethylation (H3K4me3, active promoters), H3K9ac (active promoters and enhancers), and H3K36me3 (active gene bodies). Similarly, a range of different heterochromatic marks exist. Classically, these can be divided into facultative heterochromatin marks, which

play a role in repression of lineage-specific genes, and constitutive heterochromatin marks, which repress viral elements and safeguard the integrity of the genome.

H3K27me3 and H2AK119ub1, markers of facultative heterochromatin

The histone PTMs H3K27me3 and H2AK119ub1 play an important role in the repression of lineage-specific genes during development and differentiation[15,16]. As such, they display cell type-specific patterns across the genome and are considered hallmarks of facultative heterochromatin. H3K27me3 and H2AK119ub1 are established by Polycomb Repressive Complex 2 (PRC2) and PRC1, respectively. The PRC2 catalytic subunit responsible for depositing H3K27me3 is Ezh1 or Ezh2, while Ring1a or Ring1b catalyzes H2AK119ub1 in PRC1. Different versions of both complexes exist that are defined by the presence of specific subcomponents, but all variants contain these catalytic subunits. There is extensive crosstalk between PRC1 and PRC2, as different versions can recognize both their own associated histone PTM and that of the other[17-21]. Consequently, the two marks overlap extensively in many systems. Despite a plethora of research, the exact mechanisms by which the PRC complexes and their associated histone marks mediate gene repression are still incompletely understood. The evidence so far suggests that repression is achieved through multiple routes in a context-specific manner and that most of these mechanisms rely on the catalytic activity of the complexes[22]. The histone PTMs themselves thus play an important role in transcriptional silencing. For example, H2AK119ub1 may directly limit transcription due to its bulky size, resulting in steric hindrance and exclusion of the transcriptional machinery[22]. H3K27me3, on the other hand, may in part contribute to repression via reader proteins that recruit transcriptional silencers[23,24]. See Blackledge & Klose (2021)[22] for an extensive review on the topic.

Given the roles of PRC1 and PRC2 in regulating gene expression, H3K27me3 and H2AK119ub1 show extensive changes in their distribution over the course of development. During oocyte maturation and the earliest stages of embryonic development in mouse, both marks are extensively and distinctly reprogrammed, which is discussed in detail in **Chapter 5**. In pluripotent cells, such as epiblast cells in the blastocyst and embryonic stem cells (ESCs), H3K27me3 and H2AK119ub1 are primarily found at the promoters of lineage-specific genes. Interestingly, the majority of these promoters are also decorated with H3K4me3, a histone PTM associated with active promoters. The co-occupancy of the Polycomb marks with H3K4me3 may poise the underlying genes for rapid activation upon removal of H3K27me3[25,26]. The combined presence of these opposing marks is referred to as a bivalent or, more recently, a bistable state[27,28]. During differentiation, lineage-appropriate genes lose the Polycomb marks and become active. Conversely, genes that are specific to different cell types lose H3K4me3 and the domains of H3K27me3/H2AK119ub1 broaden to ensure permanent repression[29-31]. As a consequence, committed cell types have broader domains of the Polycomb histone PTMs compared to pluripotent cells and cover a larger fraction of the genome.

In addition to their role in developmental gene expression, PRC1 and PRC2 are early regulators of chromosome X inactivation in mammals[32]. The silencing of one chromosome X allele happens in female embryos as dosage compensation for the double presence of genetic material compared to male embryos[32]. This process is controlled via the expression of Xist, a non-coding RNA that is transcribed from the X chromosome. Upregulation of Xist results in the recruitment of proteins that silence gene expression, remove active marks, and deposit heterochromatic marks[32]. Notably, Xist does not diffuse throughout the nucleus and specifically coats the X allele from which it is expressed[32]. Complex mechanisms are in place that ensure that exactly one allele is silenced only in female cells[32]. After initiation of X inactivation, active marks are rapidly depleted and gene expression is silenced[33]. H2AK119ub1 is the first heterochromatic mark to be laid down due to the early recruitment of PRC1 via Xist[33-36]. Subsequently, PRC2 is recruited via binding of subunit Jarid2 to H2AK119ub1 and deposits H3K27me3[33-38]. In line with the temporal order of epigenetic events, the initiation of gene silencing seems to be largely independent of PRC1 and PRC2[34,39]. However, the recruitment of these complexes is necessary to stabilize and maintain gene repression over time[34,39,40].

H3K9me2 and H3K9me3, markers of constitutive heterochromatin

The di- and tri-methylation states of H3K9 are markers of constitutive heterochromatin, which has a similar distribution across many cell types and is enriched in repeat-rich regions like the telomeres and centromeres[41-43]. Although H3K9me2 and 3 are generally considered to be de facto markers of constitutive heterochromatin, research over the past few years has revealed that they also play a role in the regulation of cell type-specific gene expression[44,45], indicating that in some instances this mark acts as a form of facultative heterochromatin. As such, H3K9 methylation is involved in several important cellular processes: It safeguards genome stability by facilitating proper chromosome segregation, preventing recombination between repetitive regions of the genome, and restricting the transposition of retroviral elements in the genome[46]. In addition, H3K9 methylation plays a role in the maintenance of cell-type specific gene programs by preventing aberrant binding of transcription factors and consequently erroneous activation of genes[44,45,47,48].

In mammals, six enzymes have been identified that catalyze H3K9 methylation: Suv39h1, Suv39h2, Setdb1, Setdb2, G9a and Glp. These enzymes are recruited to a variety of different targets and their activity greatly depends on sequence- and chromatin-specific interactions (extensively reviewed by Padeken et al.[49]). Suv39h1 and 2 mediate H3K9me2/3 in regions of constitutive heterochromatin, such as satellite repeats in pericentromeric regions and retroviral elements[50,51]. Setdb1 catalyzes all three methylation states of H3K9 and has similar types of targets as Suv39h1/2, although their exact localizations are only partially overlapping[45,51-54]. Setdb2, on the other hand, only establishes H3K9me3 from H3K9me1/2 and not much is known about its targets[49,55]. Finally, G9a and Glp lay down H3K9me1/2 and are recruited to various repeat classes to ensure their repression[56,57]. In addition, they are considered the main H3K9 methyltransferases involved in gene repression[58,59], although all

six enzymes have been implicated in the silencing of lineage-specific genes[44,45]. While the six methyltransferases have specialized roles, their activities partially overlap and loss of one or multiple of these enzymes can be compensated by the others[45,49]. H3K9 methylated chromatin is further stabilized due the fact that the different writers can bind to themselves, H3K9me1/2/3, and to readers of these marks[49]. The most well-characterized reader of H3K9me2/3 is heterochromatin protein 1 (Hp1)[60,61], which additionally forms dimers and interacts with other heterochromatic proteins[62]. Through these multivalent interactions, Hp1 enables the compaction and spreading of heterochromatin[62], in part through the formation of phase-separated droplets[63,64].

At sites of constitutive heterochromatin, such as centromeres and telomeres, H3K9me2/3 signatures are shared across most cell types. However, in other regions of the genome, these marks adopt cell-type specific patterns that change over the course of differentiation and development. Sequencing studies profiling the genome-wide distribution of H3K9me3 have revealed that this mark is extensively remodeled during mouse gametogenesis and early embryogenesis[65-67] (see **Chapter 5** for more details). In pluripotent systems, H3K9me3 is present at certain repeat classes and a set of developmental genes in relatively narrow domains[44,68,69]. Upon differentiation, these domains broaden to cover a substantially larger portion of the genome[29,30]. Compared to H3K9me3, less is known about the genome-wide distribution of H3K9me2. To date, H3K9me2 has not been profiled with whole-genome sequencing techniques in early mammalian embryos, but microscopy studies in mice do show that this mark is inherited from the oocyte and remains present in subsequent developmental stages[70]. The mark has been profiled in mouse ESCs (mESCs) and differentiated cells using ChIP and microarrays, but this has yielded conflicting results[71-73]. However, it appears that H3K9me2 tends to form broad domains in intergenic regions, which are enriched for repeats[71,72,74].

The increased prevalence of H3K9me2/3, as well as H3K27me3 and H2AK119ub1, in committed versus pluripotent cells suggests that heterochromatin expands over the course of differentiation. Indeed, microscopy experiments show an increase in the amount of compacted heterochromatin within the nucleus of differentiated cells[75]. These and other observations have led to the hypothesis that chromatin of pluripotent cells is in an open and dynamic state that becomes progressively more restricted upon lineage commitment[29,30,76]. It is likely through this expansion of heterochromatin domains that H3K9 methylation helps to maintain cell identity. In support of this hypothesis, the presence of these domains hampers the reprogramming of cells in the generation of induced pluripotent stem cells and of somatic cell nuclear transfer (SCNT) embryos [77,78].

Lamina-Associated Domains (LADs)

In addition to having distinct sets of epigenetic marks, euchromatin and heterochromatin localize to different regions within the nucleus. In nearly all cell types, there is a strong tendency for heterochromatin to be positioned at the periphery of the nucleus, whereas euchromatin is

located in the nuclear interior. The inner nuclear membrane is lined by a filamentous network of proteins called the nuclear lamina (NL) that provides a binding platform for chromatin[79]. The positioning of chromatin at the NL is an important facet of genome organization and is tightly linked to epigenetic state[80]. Regions contacting the NL are termed Lamina-Associated Domains (LADs). While the presence of heterochromatin along the NL was already observed by electron microscopy in the 1960's[81], the first genomic LAD profiles were obtained using DamID[82-84]. DamID is a genomics technique that relies on the fusion of the *E. coli* DNA adenine methyltransferase (Dam) to a protein of interest (POI). Upon expression of this fusion protein in the cell, Dam will methylate DNA in proximity to the POI, specifically the adenines in the palindromic sequence motif GATC. After cell harvesting, the methylated fraction of the genome can be recovered via digestion with the methylation-sensitive restriction enzyme DpnI, followed by adapter ligation and amplification.

DamID studies in mammalian cell lines have revealed that LADs span broad regions (median size ~500 kbp) that collectively cover ~40% of the genome[84,85]. LADs are generally gene poor, enriched for L1 LINE repeats and have a low GC content[84,86]. The genes that are located within LADs are typically silent or only transcribed at low levels[84,85]. A subset of LADs is shared across different cell types and are hence termed constitutive LADs (cLADs)[85,86]. Conversely, some regions change their radial positioning in the nucleus over the course of differentiation, which is frequently accompanied by changes in expression of the underlying genes[83,85,87,88]. These variable regions are therefore called facultative LADs (fLADs). The majority of LADs is enriched for H3K9me2/3[71,74,84,89-91], with especially H3K9me2 showing near exclusive localization at the NL in several cell types[74,92,93]. In some studies, a subset of LADs has also been shown to be enriched for H3K27me3, particularly at their borders[84,94]. However, a recent study has found that H3K27me3 may in fact constrain NL association[95].

The coincidence of LADs with silent heterochromatin and the existence of fLADs suggest a role for NL association in regulating gene expression. However, the exact contexts and mechanisms by which NL association contributes to gene repression are still under active investigation. Several studies have been performed in which a reporter locus was artificially tethered to the NL to observe its transcriptional response, but the results have been mixed with some reporters being repressed and others maintaining gene activity[96-98]. A study employing Thousands of Reporters Integrated in Parallel (TRIP) investigated the transcriptional activity of reporters integrated in thousands of loci in a mESC line and found that integrations in LADs showed considerably lower expression[99]. The suppressive role of LADs was stronger than could be explained by the presence of H3K9me2, local transcriptional activity, and chromatin compaction. Another comprehensive study confirmed the repressive effect of integration in LADs and, in addition, found that the extent of this repression is strongly influenced by the promoter sequence and chromatin context[100]. Together, these studies suggest that NL association itself can modulate gene expression on top of other factors such as regulatory sequences and chromatin state. While these results point toward a regulatory role for the NL in transcriptional regulation, the mechanisms by which this is achieved remain unclear.

Potentially, this involves the presence of transcriptional repressors and/or the exclusion of transcriptional activators[80].

Although LADs are a prominent feature of genome organization, the means by which chromatin is targeted to the NL are still unclear. Lamina association may in part be explained by the exclusion of heterochromatin from the nuclear interior, where the more decondensed and active euchromatin resides. However, heterochromatin has a preference to self-associate and several studies have shown that disruption of NL components and LADs results in increased interactions between these domains[101-105]. This demonstrates that LADs cannot be explained purely by passive processes and that there likely are active mechanisms that lock chromatin at the NL, countering the forces that promote homotypic interactions between heterochromatin. Several proteins have been shown to play a role in tethering LADs to the nuclear periphery. In mammalian systems, these include components of the NL (Lamin A, Lamin C, Lamin B1)[104-110], proteins spanning the nuclear membrane (Lbr, Emerin, Lap2b)[92,104,111-113], and proteins that associate with the NL (Prr14, Prdm16, Zkscan3, Hdac3)[92,114-118]. For several of these proteins, chromatin binding seems to depend on methylated H3K9, either directly or via Hp1[90,115-117,119]. While in mammals it appears that several proteins cooperate to target chromatin to the NL, the *C. elegans* protein Cec-4 forms a direct tether between the inner nuclear membrane and methylated H3K9 in embryos[120]. In addition to chromatin-mediated tethering, there is some evidence for sequence-specificity in NL contacts[94,113,121]. Despite the abundance of research in this area, it is still not completely understood how specific NL association is achieved[122]. Studying these mechanisms is complicated by the fact that they are likely redundant and not universal across cell types and genomic loci.

The single-cell perspective

The study of chromatin has been ongoing for over a century, starting in 1879 when Walther Flemming observed that nuclei contained threads that could be easily stained and coined this substance “chromatin”[123]. Since then, huge strides have been made in the field of epigenetics, assisted by the parallel development of increasingly more advanced scientific technologies. The latest revolution in the field of chromatin biology is the development of single-cell epigenomics techniques. While microscopy has allowed the study of individual cells for centuries (and was in fact the means by which cells were first observed by Robert Hooke and Antoni van Leeuwenhoek in the 17th century), genetic and transcriptomic studies were largely not possible until recently, except for a few cells or a few loci at a time using e.g. DNA/RNA FISH or single-cell qPCR[124-127]. With the advent of single-cell RNA sequencing methods, the single-cell biology field took off and since then a myriad of new techniques have been published. Initially, these largely assayed transcription, but soon single-cell versions of (epi)genomics techniques became available as well, including single-cell implementations of bisulfite sequencing[128,129] (for CpG methylation), Hi-C[130,131] (for 3D conformation of the DNA), ATAC-seq[132,133] (for chromatin accessibility), and ChIP-seq[134] (for protein-DNA interactions). By now, a wide range of single-cell techniques are available for essentially all aspects of genome regulation. Moreover, the latest

wave of innovations has enabled the combination of multiple techniques into one, resulting in so-called multi-omic or multi-modal techniques that provide a readout of two or more aspects of cellular state[135]. Single-cell techniques have facilitated numerous discoveries that would have been impossible to achieve with their bulk counterparts. However, the single-cell omics field is still relatively young and various experimental and analytical challenges remain. In the following sections, I will discuss the sources of cellular heterogeneity that warrant single-cell studies, the different experimental methods available to study protein-DNA interactions at a single-cell level, and the foundation of single-cell data analysis.

Heterogeneity between cells

Many valuable insights on gene regulation have been made using techniques assaying transcription or epigenetic features in samples of thousands to millions of cells. However, these so-called "bulk" or "population" techniques inherently provide a population-average view of the assayed feature, thus losing any information on variation within the sample. Broadly speaking, there are two sources of heterogeneity: differences in cell type and differences in cell state. While these definitions may vary per context and are an active topic of debate in the scientific community[136], cell type is used here to refer to the molecular properties shared between cells that are at the same point in a developmental trajectory and thus share the same developmental origin and potential (based on Domcke and Shendure (2023)[137]). Cell state, on the other hand, refers to the variation in these molecular properties due to processes that are not related to the developmental trajectory of the cell, such as the cell cycle, responses to the environment, or stochastic fluctuations. The differences in cell type and state can manifest in all aspects of the cell, including its genome, epigenome, transcriptome, proteome, metabolism, protein pathway activity, cell morphology, etcetera. However, some features may be more prone to variation than others: For example, in most circumstances, the genetic sequence remains constant between cells, while there may be big differences in transcription. In addition, the extent in which variation has an impact on cellular function, and thus cell type and state, also varies strongly depending on the feature and the context.

The extensive variability in complex biological systems thus poses a challenge to the study of gene regulation and chromatin using bulk technologies. However, using single-cell technologies, it is possible to assay this variability and get a more fine-grained view of the system, including the diversity of cell types and states. Moreover, the differences between individual cells also represent perturbations relative to the average cell state. If a readout of multiple cellular features is obtained for a large number of cells, these "perturbations" enable statistical analyses that can provide insight on mechanistic links between these features. In studying chromatin, single-cell information is thus key to understand the extent to which molecular features vary between cells and how they impact cellular function.

Single-cell techniques to profile DNA-protein interactions

To accomplish the goal of studying gene regulation at a single-cell level, various single-cell technologies have been developed over the past decade, with a particular acceleration in the past few years. These technologies now enable the measurement of many different aspects of the cell, including transcription, chromatin conformation, chromatin accessibility, DNA methylation, cell surface markers, and protein expression[135,138,139]. As epigenetic regulation is mainly accomplished through the interaction between proteins and DNA, I will here focus on the subset of single-cell epigenomics tools that provide genome-wide binding profiles of proteins.

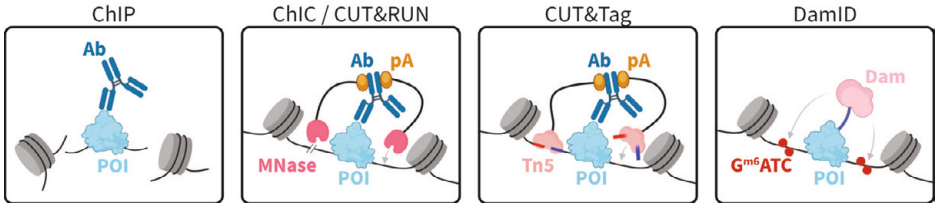


Figure 1: Strategies to capture protein-DNA interactions
The schematic representations show the step in the protocol that results in specific detection of DNA in proximity of the POI. **ChIP**: Cells or nuclei are fixed, after which the DNA is fragmented. Using antibodies (Ab) specific for the protein of interest (POI), the POI and associated DNA can be pulled down. **ChIC / CUT&RUN**: Nuclei are stained with an Ab specific for the POI. A fusion of protein A (pA) and MNase is added, which binds to the Ab. Upon addition of Ca^{2+}, MNase cleaves the DNA around the POI and the POI-bound fragments are released into the supernatent. **CUT&Tag**: Nuclei are stained with an Ab specific for the POI. A fusion of pA and Tn5 transposase is added, which binds to the Ab. The Tn5 is loaded with sequencing adapters and activated upon the addition of Mg^{2+}, resulting in the simultaneous cleavage of the DNA and integration of the adapters. **DamID**: A fusion between DNA adenine methyltransferase (Dam) and the POI is expressed in live cells. Dam methylates adenines in the GATC motif in proximity of the POI. After cell harvest, the methylated fraction of the DNA can be specifically digested and amplified. (Created with BioRender.com)

Among the range of single-cell techniques, there are currently five general strategies to capture protein-DNA interactions (Fig. 1). The first strategy is based on the principle of chromatin immunoprecipitation (ChIP), where chromatin is fixed, fragmented and stained with an antibody specific for the POI. The antibody-bound fraction of the chromatin can subsequently be pulled down and processed for sequencing. One of the first reported single-cell techniques profiling protein-DNA interactions was single-cell (sc)ChIP-seq[134], which made use of this principle. However, the pull-down step in ChIP approaches is quite inefficient, resulting in a loss of material and consequently very sparse profiles. As such, no further single-cell techniques have been based on this strategy, which has been abandoned in favor of approaches that retain more material. Among these alternatives is chromatin immuno-cleavage (ChIC), which had already been developed to investigate protein-DNA interactions of specific loci in bulk samples two decades ago[140] and has more recently been adapted for whole-genome sequencing in the form of CUT&RUN[141]. The basic premise of ChIC and CUT&RUN methods is that nuclei are stained with an antibody for the POI, after which a fusion of protein A and MNase (pA-MNase) is added. Protein A binds specifically to Immunoglobulin

G (IgG) antibodies, thus targeting the MNase to regions of the genome at which the POI is present. Upon addition of Ca^{2+}, MNase is activated, chromatin is cleaved, and the resulting DNA fragments are released into the supernatant. Several single-cell techniques make use of this principle, including scCUT&RUN[142], scChIC-seq[143], and sortChIC[144]. While these techniques all use protein A-based recruitment of MNase, they differ in the timing of antibody and pA-MNase incubation (before or after single-cell sorting) and the method by which digested fragments are enriched. A third approach to profile protein-DNA interactions also makes use of the antibody-binding properties of protein A, but uses it to recruit an adapter-loaded Tn5 transposase to the sites of antibody binding. Upon addition of Mg^{2+}, the Tn5 is activated and the adapters are inserted into the genome. The use of Tn5 thus combines the steps of digestion and adapter integration into one. This strategy is often referred to as CUT&Tag and was developed as an improvement on CUT&RUN[145]. Single-cell protocols based on the CUT&Tag principle include scCUT&Tag[145], antibody-guided chromatin tagmentation (ACT-seq)[146], and combinatorial barcoding and targeted chromatin release (CoBATCH)[147]. Among techniques using Tn5-based approaches, chromatin integration labeling (ChIL-seq)[148] employs a slight variation to the workflow: In this technique, the chromatin is stained with an antibody that has been conjugated to a DNA adapter, after which the Tn5 is assembled stepwise onto the adapter. A fourth and final method is DamID, thus labeling the DNA directly in live cells using a Dam-POI fusion. Of the four strategies, DamID is the only method that does not require an antibody specific for the POI. The single-cell implementation of this technique, scDamID[89], follows a similar principle as its bulk counterpart, but includes sorting cells into individual wells prior to digestion of methylated sites.

Whereas the techniques described above solely assay the interactions of a protein with the DNA, a new generation of single-cell techniques combines this with a readout of transcription. These so-called multi-omic or multi-modal techniques provide the chance to study the relationship between epigenetic control and transcriptional output at unprecedented resolution in complex biological systems. In **Chapter 2**, we describe the development of scDam&T-seq, the first technique to combine these two readouts. This technique is an adaptation of the original scDamID protocol to enable increased throughput and incorporate the molecular steps necessary to process mRNA. In the past years, additional techniques have been published that combine protein-DNA interaction and transcriptional readouts (scPCOR-seq[149], CoTECH[150], Paired-Tag[151], scSET-seq[152]). Notably, all these use Tn5 tagmentation-based strategies to capture the interactions of the POI with the genome. However, the means by which the transcriptional readout is captured and separated from the genomic readout differs per technique (reviewed in Vandereyken et al.[135]). In addition to the development of these multi-omic techniques, several methods have been published in the last year that capture the binding profile of multiple proteins from the same cell[153-160]. Such multifactorial techniques can help to further disentangle the regulatory relationships between various epigenetic mechanisms.

The basis of single-cell data analysis

The main strength of single-cell techniques is that they allow us to investigate how cellular heterogeneity is reflected in a certain molecular aspect of the cells, such as histone PTMs or transcription. The ability to observe heterogeneity opens the door to many powerful analyses that can infer information about the biological system, such as the diversity of cell types and states, the presence of cell state transitions, the transcriptomic and/or epigenetic features of these distinct states, and co-regulation of loci or genes. Such analyses rely on a number of key features of single-cell experiments (Fig. 2): **(1)** No a priori knowledge of the cell types and states in the biological sample is necessary. As long as cells can be dissociated and individually processed, the different populations will be represented in the data set. **(2)** Correspondingly, it is not necessary to identify and separate cell populations of interest experimentally, which requires reliable biomarkers and an a priori decision on which populations are relevant. Instead, cells can be grouped in silico in any desired configuration based on their single-cell readout. This is especially valuable for systems with dynamic transitions between cell states or types, precluding an easy definition of distinct groups. A notable prerequisite for this strategy is that the assayed feature indeed reflects the different cell types or states of interest. **(3)** The single-cell resolution provides a more fine-grained view of the relationship between different genomic features or regulatory layers. In the case of a single molecular readout, this allows us to relate the behavior of one genomic feature to that of others across all cells, for example the expression of a transcription factor with its potential target genes. In the case of multiple readouts, we can also apply this concept to study the relationship between two molecular layers, for example, to relate the presence of a histone PTM at the promoter of a gene to its expression. The fine-grained view of these relationships provides more direct mechanistic insight into their relationship, especially when coupled with perturbations[161,162]. **(4)** Due to the large number of single-cell samples that are typically obtained, analyses of single-cell data gain a lot of statistical power. Moreover, the increasingly high cell numbers open the door to machine learning methods, including artificial intelligence (AI), that can mitigate the sparsity of the single-cell data and have the potential to give more comprehensive insight in the highly complex biological relationships contained within it[163]. **(5)** Single-cell datasets are very versatile, as they can be reused and combined for various purposes by using the right data integration tools. The integration of multiple datasets can be useful for increasing cell numbers and thus improving sensitivity in cell clustering and the detection of rare cell types. For example, large single-cell atlases are available for a number of biological systems[141,164-169], which allows other studies with smaller datasets to study these systems at the same resolution of cell-type clustering. With the development of multi-omics tools, data integration can also be employed to incorporate additional readouts into an existing dataset. These five key features form the basis of single-cell data analysis and are utilized in an ever-expanding range of computational tools.

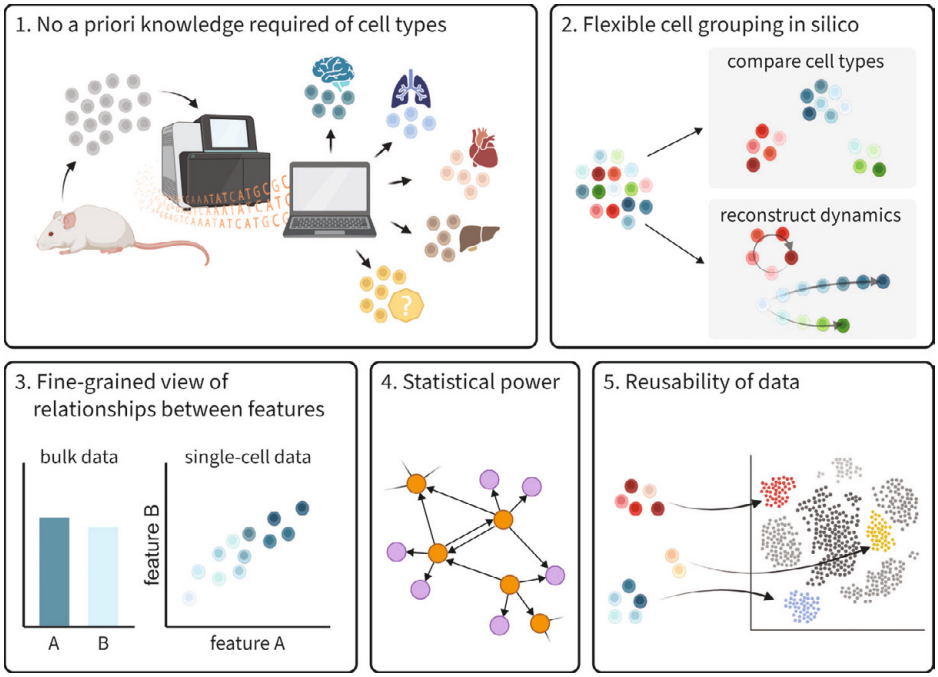


Figure 2: Five key advantages of single-cell data
1) No a priori knowledge of the cellular composition of the system is required; 2) Assayed cells can be endlessly grouped and arranged in silico, eliminating the need to specifically choose and isolate populations of interest experimentally; 3) The data provides a more fine-grained view of the relationship between features; 4) Large numbers of cells boost the statistical power of analyses, allowing the development and testing of complex models; 5) Single-cell datasets can be combined to improve cell type classification or share information across modalities. (Created with BioRender.com)

While single-cell experiments have unique advantages, they also suffer from some shortcomings. Due to incomplete capture and sequencing of all molecules in the cell, the resulting data is undersampled and typically very sparse. Consequently, some features that were actually expressed (or in contact with the POI) are not detected, resulting in zeros in the data due to dropout[170]. These zeros caused by dropout represent false negatives that are unfortunately indistinguishable from the true negative zero values of features that were indeed not expressed (or not in contact with the POI). Depending on the method used, dropout does not affect all features equally. For example, in the case of scRNA-seq experiments, genes with lower expression levels, such as TFs, naturally tend to drop out more. This can obscure interesting biological information, as TFs are important regulators of gene expression programs. One way to limit the impact of dropout is by averaging data over groups of similar cells. This provides a more robust data representation, but the single-cell resolution will be lost. Alternatively, several methods have been developed to combat dropout by imputing the missing data. These methods typically work by using statistical models to identify and impute dropouts, smoothing data over similar cells, or mapping cells to a latent space from which an estimate of the true values can be computed[163,171]. While data imputation strategies hold

strong promise, they have to be used with some caution, as their use can result in circularity in the data and false-positive relationships between features[171,172].

Another notable challenge for single-cell methods are batch effects between experiments. While batch effects also affect bulk techniques[173], they are especially problematic for single-cell experiments, as the downstream analyses often aim to perform data-driven clustering or ordering of cells into biologically relevant categories. The presence of batch effects can greatly impact this step, leading to technical rather than biological differences between categories. For data analysis involving multiple experiments, some action is thus required to limit the influence of technical factors. As for the sparsity problem, multiple computational methods have been developed to mitigate batch effects. The first popular batch correction methods are based on the principle of identifying similar cells across batches and mapping them to a common space in which technical variation is removed[174-176]. Recently, a new set of methods have been developed that make use of deep learning approaches to perform multiple tasks simultaneously, including batch correction, data imputation, and clustering[163]. While sparsity and batch effects represent two of the main technical limitations of single-cell data, there are additional challenges and opportunities in the field of single-cell data analysis that, when addressed, could reveal more of the biological information captured within the data[163].

Scope of this thesis

The overarching goal of the research presented in my dissertation is to understand the complex interplay between genome organization, chromatin, gene expression and cell state/type. Specifically, the focus has been on characterizing and relating several aspects of heterochromatin to each other and to gene expression programs in the context of development. The study of these relationships benefits tremendously from the use of single-cell technologies, particularly those providing a multi-omic readout of both epigenetic regulation and transcription. As few such techniques were available at the start of my PhD research, this thesis also has the technological objective of developing tools to study chromatin and expression at single-cell resolution.

In **Chapter 2**, we develop scDam&T-seq, the first reported method to jointly measure protein-DNA interactions and transcription in single cells. This combined readout is achieved by combining the single-cell implementation of DamID (scDamID)[89] with a single-cell protocol for RNA-seq (CEL-Seq2)[177]. We demonstrate the potential of scDam&T-seq by studying the effect of NL association on transcription within a culture of mESCs and recruitment of PRC1 component Ring1b to one allele of chromosome X during differentiation-associated random X inactivation.

To make scDam&T-seq more accessible to the scientific community, we provide an in-depth experimental and computational protocol in **Chapter 3**. Moreover, we elaborate on important factors to consider when designing experiments and provide an example of the expected output.

While DamID-based techniques can be implemented to study the genome-wide binding profiles of proteins that can be genetically encoded, it excludes the specific capture of

post-translationally modified versions of proteins, including histone PTMs. To extend the applicability of the DamID toolkit to histone PTMs, we developed EpiDamID, which is presented in **Chapter 4**. Using EpiDamID, we performed scDam&T-seq experiments to study the changing distribution of H3K27me3 over a differentiation trajectory from mESC to embryoid bodies and cell-type specific profiles of H3K9me3 in the zebrafish embryo.

Chapters 2-4 involve the development of tools necessary to study epigenetic regulation at a single-cell level. With these tools at our disposal, we next turn to the study of the preimplantation mouse embryo, a system in which chromatin state and expression patterns are highly dynamic. The major epigenetic changes that take place within the embryo and the limited availability of material make the implementation of single-cell techniques especially beneficial.

Chapter 5 provides an in-depth overview of our knowledge on heterochromatin and genome organization during mouse preimplantation development. The earliest stages of the mouse embryo present a fascinating system in which the maternal and paternal genomes are epigenetically reprogrammed to give rise to the totipotent zygote. During these stages, many epigenetic features adopt highly atypical distributions across the genome and the canonical 3D organization of the genome is largely lost. We discuss each of these features and discuss evidence for potential relationships between them.

Building on this existing body of work, we next employ scDam&T-seq and EpiDamID in **Chapter 6** to study LADs and heterochromatic histone PTMs over the first few stages of mouse development. We observe an unprecedented level of heterogeneity in genome-lamina contacts between cells and embryos at the 2-cell stage. The increase in LAD variability between the zygote and 2-cell stage coincides with an extensive reorganization of genome-lamina contact profiles, during which many canonical LADs are relocated to the nuclear interior. We discover that maternal PRC2 deposits H3K27me3 in canonical LADs and that this mark shows a strong negative correlation with NL association in the early embryo. Through various experiments and analyses, we demonstrate that the atypical and variable genome-lamina contacts are the result of a tug of war between two opposing forces: On the one hand, intrinsic NL affinity of the DNA sequence promotes peripheral localization, whereas, on the other hand, the presence of H3K27me3 hampers NL association. The extensive overlap between H3K27me3 and regions of high NL affinity is unique to the early embryo and likely contributes to the unusual organization characteristic of the totipotent genome.

Finally, in **Chapter 7**, I reflect on various topics that feature prominently in this thesis. First, I consider the recent and future developments in the field of single-cell omics. In addition, I discuss the advantages and limitations of scDam&T-seq, highlighting possible adaptations of the method to tackle new biological questions and remain competitive. Next, I discuss to which extent stochastic NL contacts may affect gene expression, incorporating findings from this thesis and literature. Furthermore, I attempt to consolidate the available and seemingly contradictive evidence on the effect of H3K27me3 on NL association. Finally, I speculate on what the role of the atypical NL association could be in the early embryo.

References

1 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

2 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

3 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022).

4 Bansal, A. K. & Meyer, T. E. Evolutionary analysis by whole-genome comparisons. *J Bacteriol* **184**, 2260-2272 (2002).

5 Wildman, D. E. *et al.* Genomics, biogeography, and the diversification of placental mammals. *Proc Natl Acad Sci U S A* **104**, 14395-14400 (2007).

6 Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741-748 (2020).

7 Mondal, M. *et al.* Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat Genet* **48**, 1066-1070 (2016).

8 Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* **17**, 355-370 (2019).

9 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871 (1974).

10 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260 (1997).

11 Martire, S. & Banaszynski, L. A. The roles of histone variants in fine-tuning chromatin organization and function. *Nat Rev Mol Cell Biol* **21**, 522-541 (2020).

12 Millan-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications - cause and consequence of genome function. *Nat Rev Genet* **23**, 563-580 (2022).

13 Chen, T. & Dent, S. Y. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet* **15**, 93-106 (2014).

14 Nacev, B. A. *et al.* The expanding landscape of 'oncohistone' mutations in human cancers. *Nature* **567**, 473-478 (2019).

15 Loh, C. H. & Veenstra, G. J. C. The Role of Polycomb Proteins in Cell Lineage Commitment and Embryonic Development. *Epigenomes* **6** (2022).

16 Piunti, A. & Shilatifard, A. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nat Rev Mol Cell Biol* **22**, 326-345 (2021).

17 Blackledge, N. P. *et al.* Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell* **157**, 1445-1459 (2014).

18 Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039-1043 (2002).

19 Cooper, S. *et al.* Targeting polycomb to pericentric heterochromatin in embryonic stem cells reveals a role for H2AK119u1 in PRC2 recruitment. *Cell Rep* **7**, 1456-1470 (2014).

20 Kalb, R. *et al.* Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nat Struct Mol Biol* **21**, 569-571 (2014).

21 Wang, L. *et al.* Hierarchical recruitment of polycomb group silencing complexes. *Mol Cell* **14**, 637-646 (2004).

22 Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat Rev Mol Cell Biol* **22**, 815-833 (2021).

23 Fan, H. *et al.* A conserved BAH module within mammalian BAHD1 connects H3K27me3 to Polycomb gene silencing. *Nucleic Acids Res* **49**, 4441-4455 (2021).

24 Xu, P. *et al.* FBXO11-mediated proteolysis of BAHD1 relieves PRC2-dependent transcriptional repression in erythropoiesis. *Blood* **137**, 155-167 (2021).

25 Zhang, J. *et al.* Highly enriched BEND3 prevents the premature activation of bivalent genes during differentiation. *Science* **375**, 1053-1058 (2022).

26 Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome research* **31**, 2170-2184 (2021).

27 Sneppen, K. & Ringrose, L. Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent. *Nat Commun* **10**, 2133 (2019).

28 Berry, S., Dean, C. & Howard, M. Slow Chromatin Dynamics Allow Polycomb Target Genes to Filter Fluctuations in Transcription Factor Activity. *Cell Syst* **4**, 445-457 e448 (2017).

29 Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642-654 (2013).

30 Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-491 (2010).

31 Pauler, F. M. *et al.* H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res* **19**, 221-233 (2009).

32 Loda, A., Collombet, S. & Heard, E. Gene regulation in time and space during X-chromosome inactivation. *Nat Rev Mol Cell Biol* **23**, 231-249 (2022).

33 Zylicz, J. J. *et al.* The Implication of Early Chromatin Changes in X Chromosome Inactivation. *Cell* **176**, 182-197 e123 (2019).

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

35 Almeida, M. *et al.* PCGF3/5-PRC1 initiates Polycomb recruitment in X chromosome inactivation. *Science* **356**, 1081-1084 (2017).

36 Pintacuda, G. *et al.* hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. *Mol Cell* **68**, 955-969 e910 (2017).

37 Cooper, S. *et al.* Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2. *Nat Commun* **7**, 13661 (2016).

38 da Rocha, S. T. *et al.* Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Mol Cell* **53**, 301-316 (2014).

39 Bousard, A. *et al.* The role of Xist-mediated Polycomb recruitment in the initiation of X-chromosome inactivation. *EMBO Rep* **20**, e48019 (2019).

40 Colognori, D., Sunwoo, H., Wang, D., Wang, C. Y. & Lee, J. T. Xist Repeats A and B Account for Two Distinct Phases of X Inactivation Establishment. *Dev Cell* **54**, 21-32 e25 (2020).

41 Martens, J. H. *et al.* The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J* **24**, 800-812 (2005).

42 Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113 (2001).

43 Garcia-Cao, M., O'Sullivan, R., Peters, A. H., Jenuwein, T. & Blasco, M. A. Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases. *Nat Genet* **36**, 94-99 (2004).

44 Bilodeau, S., Kagey, M. H., Frampton, G. M., Rahl, P. B. & Young, R. A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-2489 (2009).

45 Nicetto, D. *et al.* H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. *Science* **363**, 294-297 (2019).

46 Janssen, A., Colmenares, S. U. & Karpen, G. H. Heterochromatin: Guardian of the Genome. *Annu Rev Cell Dev Biol* **34**, 265-288 (2018).

47 Methot, S. P. *et al.* H3K9me selectively blocks transcription factor activity and ensures differentiated tissue integrity. *Nat Cell Biol* **23**, 1163-1175 (2021).

48 McCarthy, R. L. *et al.* Diverse heterochromatin-associated proteins repress distinct classes of genes and repetitive elements. *Nat Cell Biol* **23**, 905-914 (2021).

49 Padeken, J., Methot, S. P. & Gasser, S. M. Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. *Nat Rev Mol Cell Biol* (2022).

50 Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593-599 (2000).

51 Loyola, A. *et al.* The HP1alpha-CAF1-SetDB1-containing complex provides H3K9me1 for Suv39-mediated K9me3 in pericentric heterochromatin. *EMBO Rep* **10**, 769-775 (2009).

52 Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J., 3rd. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev* **16**, 919-932 (2002).

53 Yang, L. *et al.* Molecular cloning of ESET, a novel histone H3-specific methyltransferase that interacts with ERG transcription factor. *Oncogene* **21**, 148-152 (2002).

54 Wang, H. *et al.* mAM facilitates conversion by ESET of dimethyl to trimethyl lysine 9 of histone H3 to cause transcriptional repression. *Mol Cell* **12**, 475-487 (2003).

55 Falandry, C., Campone, M., Cartron, G., Guerin, D. & Freyer, G. Trends in G-CSF use in 990 patients after EORTC and ASCO guidelines. *Eur J Cancer* **46**, 2389-2398 (2010).

56 Jiang, Q. *et al.* G9a Plays Distinct Roles in Maintaining DNA Methylation, Retrotransposon Silencing, and Chromatin Looping. *Cell Rep* **33**, 108315 (2020).

57 Maksakova, I. A. *et al.* Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics Chromatin* **6**, 15 (2013).

58 Tachibana, M. *et al.* G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes Dev* **16**, 1779-1791 (2002).

59 Tachibana, M. *et al.* Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. *Genes Dev* **19**, 815-826 (2005).

60 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124 (2001).

61 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-120 (2001).

62 Canzio, D., Larson, A. & Narlikar, G. J. Mechanisms of functional promiscuity by HP1 proteins. *Trends Cell Biol* **24**, 377-386 (2014).

63 Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241-245 (2017).

64 Larson, A. G. *et al.* Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin. *Nature* **547**, 236-240 (2017).

65 Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620-631 (2018).

66 Xu, R. *et al.* Stage-specific H3K9me3 occupancy ensures retrotransposon silencing in human pre-implantation embryos. *Cell Stem Cell* **29**, 1051-1066 e1058 (2022).

67 Yu, H. *et al.* Dynamic reprogramming of H3K9me3 at hominoid-specific retrotransposons during human preimplantation development. *Cell Stem Cell* **29**, 1031-1050 e1012 (2022).

68 Bulut-Karslioglu, A. *et al.* Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Mol Cell* **55**, 277-290 (2014).

69 Lehnertz, B. *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* **13**, 1192-1200 (2003).

70 Liu, H., Kim, J. M. & Aoki, F. Regulation of histone H3 lysine 9 methylation in oocytes and early pre-implantation embryos. *Development* **131**, 2269-2280 (2004).

71 Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* **41**, 246-250 (2009).

72 Lienert, F. *et al.* Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS Genet* **7**, e1002090 (2011).

73 Filion, G. J. & van Steensel, B. Reassessing the abundance of H3K9me2 chromatin domains in embryonic stem cells. *Nat Genet* **42**, 4; author reply 5-6 (2010).

74 Poleshko, A. *et al.* Genome-Nuclear Lamina Interactions Regulate Cardiac Stem Cell Lineage Restriction. *Cell* **171**, 573-587 e514 (2017).

75 Ahmed, K. *et al.* Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS One* **5**, e10531 (2010).

76 Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol* **7**, 540-546 (2006).

77 Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994-1004 (2012).

78 Matoba, S. *et al.* Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. *Cell* **159**, 884-895 (2014).

79 Burke, B. & Stewart, C. L. The nuclear lamins: flexibility in function. *Nat Rev Mol Cell Biol* **14**, 13-24 (2013).

80 van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780-791 (2017).

81 Fawcett, D. W. On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates. *Am J Anat* **119**, 129-145 (1966).

82 van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* **18**, 424-428 (2000).

83 Pickersgill, H. *et al.* Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet* **38**, 1005-1014 (2006).

84 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).

85 Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-613 (2010).

86 Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* **23**, 270-280 (2013).

87 Madsen-Osterbye, J., Abdelhalim, M., Baudement, M. O. & Collas, P. Local euchromatin enrichment in lamina-associated domains anticipates their repositioning in the adipogenic lineage. *Genome Biol* **23**, 91 (2022).

88 Pindyurin, A. V. *et al.* The large fraction of heterochromatin in Drosophila neurons is bound by both B-type lamin and HP1a. *Epigenetics Chromatin* **11**, 65 (2018).

89 Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134-147 (2015).

90 Kind, J. *et al.* Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178-192 (2013).

91 Zheng, X., Kim, Y. & Zheng, Y. Identification of lamin B-regulated chromatin regions based on chromatin landscapes. *Mol Biol Cell* **26**, 2685-2697 (2015).

92 Poleshko, A. *et al.* H3K9me2 orchestrates inheritance of spatial positioning of peripheral heterochromatin through mitosis. *Elife* **8** (2019).

93 Yokochi, T. *et al.* G9a selectively represses a class of late-replicating genes at the nuclear periphery. *Proc Natl Acad Sci U S A* **106**, 19363-19368 (2009).

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

95 Siegenfeld, A. P. *et al.* Polycomb-lamina antagonism partitions heterochromatin at the nuclear periphery. *Nature Communications* **13**, 4199 (2022).

96 Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243-247 (2008).

97 Kumaran, R. I., Thakar, R. & Spector, D. L. Chromatin dynamics and gene positioning. *Cell* **132**, 929-934 (2008).

98 Finlan, L. E. *et al.* Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet* **4**, e1000039 (2008).

99 Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914-927 (2013).

100 Leemans, C. *et al.* Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**, 852-864 e814 (2019).

101 Bian, Q., Anderson, E. C., Yang, Q. & Meyer, B. J. Histone H3K9 methylation promotes formation of genome compartments in Caenorhabditis elegans via chromosome compaction and perinuclear anchoring. *Proc Natl Acad Sci U S A* **117**, 11459-11470 (2020).

102 Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* **570**, 395-399 (2019).

103 Sawh, A. N. *et al.* Lamina-Dependent Stretching and Unconventional Chromosome Compartments in Early C. elegans Embryos. *Mol Cell* **78**, 96-111 e116 (2020).

104 Solovei, I. *et al.* LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584-598 (2013).

105 Zheng, X. *et al.* Lamins Organize the Global Three-Dimensional Genome from the Nuclear Periphery. *Mol Cell* **71**, 802-815 e807 (2018).

106 Amendola, M. & van Steensel, B. Nuclear lamins are not required for lamina-associated domain organization in mouse embryonic stem cells. *EMBO Rep* **16**, 610-617 (2015).

107 Chang, L. *et al.* Nuclear peripheral chromatin-lamin B1 interaction is required for global integrity of chromatin architecture and dynamics in human cells. *Protein Cell* **13**, 258-280 (2022).

108 Kychygina, A. *et al.* Progerin impairs 3D genome organization and induces fragile telomeres by limiting the dNTP pools. *Sci Rep* **11**, 13195 (2021).

109 Shah, P. P. *et al.* Pathogenic LMNA variants disrupt cardiac lamina-chromatin interactions and de-repress alternative fate genes. *Cell Stem Cell* **28**, 938-954 e939 (2021).

110 Wong, X. *et al.* Lamin C is required to establish genome organization after mitosis. *Genome Biol* **22**, 305 (2021).

111 Clowney, E. J. *et al.* Nuclear aggregation of olfactory receptor genes governs their monogenic expression. *Cell* **151**, 724-737 (2012).

112 Demmerle, J., Koch, A. J. & Holaska, J. M. The nuclear envelope protein emerin binds directly to histone deacetylase 3 (HDAC3) and activates HDAC3 activity. *J Biol Chem* **287**, 22080-22088 (2012).

113 Zullo, J. M. *et al.* DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell* **149**, 1474-1487 (2012).

114 Dunlevy, K. L. *et al.* The PRR14 heterochromatin tether encodes modular domains that mediate and regulate nuclear lamina targeting. *J Cell Sci* **133** (2020).

115 Kiseleva, A. A., Cheng, Y. C., Smith, C. L., Katz, R. A. & Poleshko, A. PRR14 organizes H3K9me3-modified heterochromatin at the nuclear lamina. *Nucleus* **14**, 2165602 (2023).

116 Poleshko, A. *et al.* The human protein PRR14 tethers heterochromatin to the nuclear lamina during interphase and mitotic exit. *Cell Rep* **5**, 292-301 (2013).

117 Biferali, B. *et al.* Prdm16-mediated H3K9 methylation controls fibro-adipogenic progenitors identity during skeletal muscle repair. *Sci Adv* **7** (2021).

118 Hu, H. *et al.* ZKSCAN3 counteracts cellular senescence by stabilizing heterochromatin. *Nucleic Acids Res* **48**, 6001-6018 (2020).

119 See, K. *et al.* Histone methyltransferase activity programs nuclear peripheral genome positioning. *Dev Biol* **466**, 90-98 (2020).

120 Gonzalez-Sandoval, A. *et al.* Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. *Cell* **163**, 1333-1347 (2015).

121 Luderus, M. E., den Blaauwen, J. L., de Smit, O. J., Compton, D. A. & van Driel, R. Binding of matrix attachment regions to lamin polymers involves single-stranded regions and the minor groove. *Mol Cell Biol* **14**, 6297-6305 (1994).

122 Manzo, S. G., Dauban, L. & van Steensel, B. Lamina-associated domains: Tethers and looseners. *Curr Opin Cell Biol* **74**, 80-87 (2022).

123 Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nat Rev Mol Cell Biol* **4**, 809-814 (2003).

124 Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010-3014 (1992).

125 Klein, C. A. *et al.* Combined transcriptome and genome analysis of single micrometastatic cells. *Nat Biotechnol* **20**, 387-392 (2002).

126 Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res* **34**, e42 (2006).

127 Lennon, G. G. & Lehrach, H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet* **7**, 314-317 (1991).

128 Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **23**, 2126-2135 (2013).

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

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

131 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).

132 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

133 Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).

134 Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33**, 1165-1172 (2015).

135 Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* **24**, 494-515 (2023).

136 Fleck, J. S., Camp, J. G. & Treutlein, B. What is a cell type? *Science* **381**, 733-734 (2023).

137 Domcke, S. & Shendure, J. A reference cell tree will serve science better than a reference cell atlas. *Cell* **186**, 1103-1114 (2023).

138 Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* **52**, 1419-1427 (2020).

139 Shema, E., Bernstein, B. E. & Buenrostro, J. D. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat Genet* **51**, 19-25 (2019).

140 Schmid, M., Durussel, T. & Laemmli, U. K. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell* **16**, 147-157 (2004).

141 Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6** (2017).

142 Hainer, S. J., Boskovic, A., McCannell, K. N., Rando, O. J. & Fazzio, T. G. Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* **177**, 1319-1329 e1311 (2019).

143 Ku, W. L. *et al.* Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nat Methods* **16**, 323-325 (2019).

144 Zeller, P. *et al.* Single-cell sortChIC identifies hierarchical chromatin dynamics during hematopoiesis. *Nat Genet* **55**, 333-345 (2023).

145 Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).

146 Carter, B. *et al.* Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat Commun* **10**, 3747 (2019).

147 Wang, Q. *et al.* CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Mol Cell* **76**, 206-216 e207 (2019).

148 Harada, A. *et al.* A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat Cell Biol* **21**, 287-296 (2019).

149 Pan, L., Ku, W. L., Tang, Q., Cao, Y. & Zhao, K. scPCOR-seq enables co-profiling of chromatin occupancy and RNAs in single cells. *Commun Biol* **5**, 678 (2022).

150 Xiong, H., Luo, Y., Wang, Q., Yu, X. & He, A. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat Methods* **18**, 652-660 (2021).

151 Zhu, C. *et al.* Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat Methods* **18**, 283-292 (2021).

152 Sun, Z. *et al.* Joint single-cell multiomic analysis in Wnt3a induced asymmetric stem cell division. *Nat Commun* **12**, 5941 (2021).

153 Bartosovic, M. & Castelo-Branco, G. Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. *Nat Biotechnol* **41**, 794-805 (2023).

154 Gopalan, S., Wang, Y., Harper, N. W., Garber, M. & Fazzio, T. G. Simultaneous profiling of multiple chromatin proteins in the same cells. *Mol Cell* **81**, 4736-4746 e4735 (2021).

155 Handa, T. *et al.* Chromatin integration labeling for mapping DNA-binding proteins and modifications with low input. *Nat Protoc* **15**, 3334-3360 (2020).

156 Meers, M. P., Llagas, G., Janssens, D. H., Codomo, C. A. & Henikoff, S. Multifactorial profiling of epigenetic landscapes at single-cell resolution using MulTI-Tag. *Nat Biotechnol* **41**, 708-716 (2023).

157 Stuart, T. *et al.* Nanobody-tethered transposition enables multifactorial chromatin profiling at single-cell resolution. *Nat Biotechnol* **41**, 806-812 (2023).

158 Xiong, H., Wang, Q., Li, C. C. & He, A. Single-cell joint profiling of multiple epigenetic proteins and gene transcription. *Sci Adv* **10**, eadi3664 (2024).

159 Lochs, S. J. A. *et al.* Combinatorial single-cell profiling of major chromatin types with MAbID. *Nat Methods* **21**, 72-82 (2024).

160 Kefalopoulou, S. *et al.* Time-resolved and multifactorial profiling in single cells resolves the order of heterochromatin formation events during X-chromosome inactivation. *bioRxiv*, 2023.2012. 2015.571749 (2023).

161 Saunders, L. M. *et al.* Embryo-scale reverse genetics at single-cell resolution. *Nature* **623**, 782-791 (2023).

162 Huang, X. *et al.* Single-cell, whole-embryo phenotyping of mammalian developmental disorders. *Nature* **623**, 772-781 (2023).

163 Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020).

164 Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495 (2019).

165 Jiang, S. *et al.* Single-cell chromatin accessibility and transcriptome atlas of mouse embryos. *Cell Rep* **42**, 112210 (2023).

166 Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595 (2020).

167 Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360** (2018).

168 Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360** (2018).

169 Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987 (2018).

170 Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562-578 (2018).

171 Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol* **21**, 218 (2020).

172 Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Res* **7**, 1740 (2018).

173 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739 (2010).

174 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

175 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427 (2018).

176 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289-1296 (2019).

177 Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 77 (2016).

CHAPTER 2

Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells

Koos Rooijers[1,5], Corina M. Markodimitraki[1,5], Franka J. Rang[1], Sandra S. de Vries[1], Alex Chialastri[2,3], Kim L. de Luca[1], Dylan Mooijman[1,4], Siddharth S. Dey[2,3]* and Jop Kind[1]*

1: Oncode Institute, Hubrecht Institute–KNAW and University Medical Center Utrecht, Utrecht, the Netherlands.
2: Department of Chemical Engineering,
University of California Santa Barbara, Santa Barbara, CA, USA.
3: Center for Bioengineering, University of California Santa Barbara, Santa Barbara, CA, USA.
4: Present address: Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.
5: These authors contributed equally: Koos Rooijers, Corina M. Markodimitraki.
*Correspondence: S.S.D. (sdey@ucsb.edu) and J.K. (j.kind@hubrecht.eu).

Nature Biotechnology, 2019

Abstract

Protein–DNA interactions are critical to the regulation of gene expression, but it remains challenging to define how cell-to-cell heterogeneity in protein–DNA binding influences gene expression variability. Here we report a method for the simultaneous quantification of protein–DNA contacts by combining single-cell DNA adenine methyltransferase identification (DamID) with messenger RNA sequencing of the same cell (scDam&T-seq). We apply scDam&T-seq to reveal how genome–lamina contacts or chromatin accessibility correlate with gene expression in individual cells. Furthermore, we provide single-cell genome-wide interaction data on a polycomb-group protein, RING1B, and the associated transcriptome. Our results show that scDam&T-seq is sensitive enough to distinguish mouse embryonic stem cells cultured under different conditions and their different chromatin landscapes. Our method will enable the analysis of protein-mediated mechanisms that regulate cell-type-specific transcriptional programs in heterogeneous tissues.

Main

Recent advances in measuring genome architecture (Hi-C and DamID)[1,2,3,4], chromatin accessibility (ATAC-seq and DNase-I-seq)[5,6,7], various DNA modifications[8,9,10,11,12,13] and histone post-translational modifications (chromatin immunoprecipitation (ChIP)-seq)[14] in single cells have enabled characterization of cell-to-cell heterogeneity in gene regulation. More recently, multi-omics methods to study single-cell associations between genomic or epigenetic variations and transcriptional heterogeneity[15,16,17,18,19] have allowed researchers to link upstream regulatory elements to transcriptional output from the same cell. At all gene-regulatory levels, protein–DNA interactions play a critical role in determining transcriptional outcomes; however, no method exists to obtain combined measurements of protein–DNA contacts and transcriptomes in single cells. We have therefore developed scDam&T-seq, a multi-omics method that harnesses DamID to map genomic protein localizations together with mRNA sequencing from the same cell.

The DamID technology involves the expression of a protein of interest tethered to *Escherichia coli* DNA adenine methyltransferase (Dam)[20]. This enables detection of protein–DNA interactions through exclusive adenine methylation at GATC motifs. In vivo expression of the DamID-constructs requires transient or stable expression at low to moderate levels[21]. An important distinction between DamID and ChIP is the cumulative nature of the adenine methylation in living cells, allowing interactions to be measured over varying time windows. This property can be exploited to uncover protein–DNA contact histories[22]. For single-cell applications, a major advantage of DamID is the minimal sample handling, which reduces biological losses and enables the amplification of different molecules in the same reaction mixture. To make DamID compatible with transcriptomics, we adapted the method for linear amplification, which allows simultaneous processing of DamID and mRNA by in vitro transcription (Fig. 1a) without nucleotide separation.

As a proof-of-principle, we first benchmarked scDam&T-seq to the previously reported single-cell DamID (scDamID) method. Single KBM7 cells expressing either untethered Dam or Dam-LMNB1 were sorted into 384-well plates by fluorescence-activated cell sorting (FACS) as described previously[2]. For scDam&T-seq, polyadenylated mRNA is reverse transcribed into complementary DNA followed by second strand synthesis to create double-stranded cDNA molecules (Fig. 1a and Methods). Next, the DamID-labeled DNA is digested with the restriction enzyme DpnI, followed by adapter ligation to digested genomic DNA (gDNA; Fig. 1a), cells are pooled, and cDNA and ligated gDNA molecules are simultaneously amplified by in vitro transcription. Finally, the amplified RNA molecules are processed into Illumina libraries, as described previously[23] (Fig. 1a and Methods).

The crucial modification compared to the original scDamID protocol is the linear amplification of the m^{6}A-marked genome. The advantages of linear amplification include (1) compatibility with mRNA sequencing, (2) unbiased genomic recovery due to the amplification of single

ligation events, (3) a >100-fold increase in throughput due to combined sample amplification and library preparation and (4) a resulting substantial cost reduction. Additional improvements of scDam&T-seq involve the inclusion of unique molecule identifiers (UMI) for both gDNA- and mRNA-derived reads and the use of liquid-handling robots to increase throughput and obtain more consistent sample quality (Fig. 1a and Methods).

We qualitatively and quantitatively compared scDam&T-seq to previously published scDamID data in KBM7 cells[2]. As illustrated for chromosome 17, observed over expected (OE) scores[2] captured the same lamina-associated domains (LADs) and cell-to-cell heterogeneity in genome–nuclear lamina (NL) interactions as previously described (Fig. 1b and Fig. S1a). This is also illustrated by the high concordance (r = 0.97) in the contact frequencies (CFs), that is, the fraction of cells in contact (OE ≥ 1) with the NL for 100-kb genomic windows (Fig. S1b).

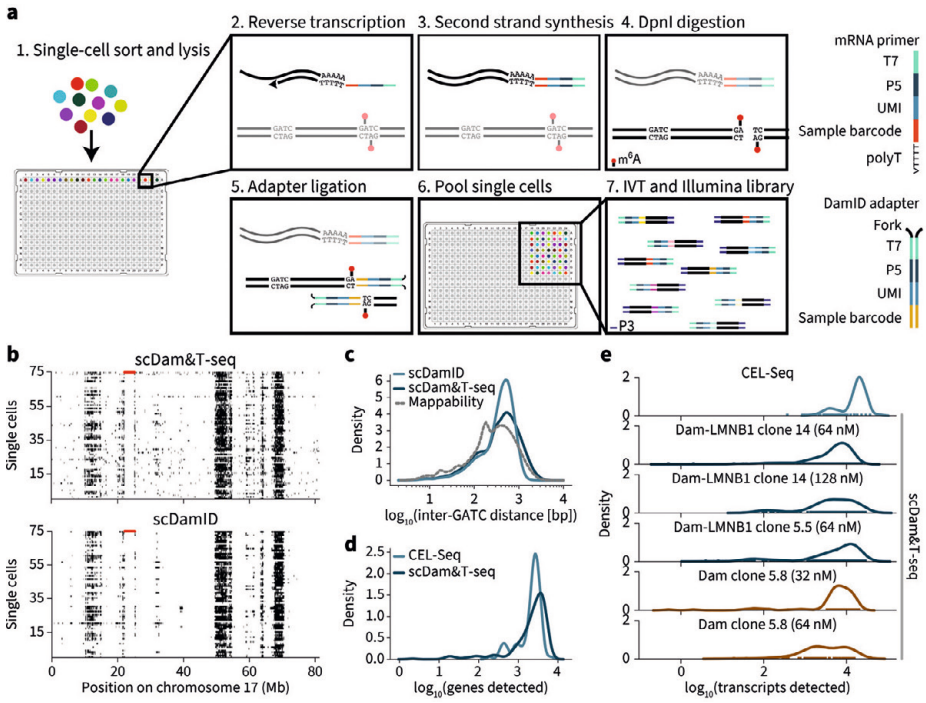


Figure 1: Quantitative comparison of scDamID, CEL-Seq and scDam&T-seq applied to KBM7 cells
a, Schematic overview of scDam&T-seq. **b**, Binarized OE values (black: OE ≥ 1) of Dam-LMNB1 signal on chromosome 17, measured with scDam&T-seq and scDamID[2] in 75 single cells with the highest sequencing depth. Each row represents a single cell and each column a 100-kb bin along the genome. Unmappable genomic regions are indicated in red along the top of the track. IVT, in vitro transcription. **c**, Distribution of inter-GATC distances of mappable GATC fragments genome-wide (dotted line), and observed in experimental data with scDamID and scDam&T-seq for Dam-LMNB1. **d**, Distributions of the number of unique genes detected using CEL-Seq[2] and scDam&T-seq on the same Dam-LMNB1 clone. **e**, Distribution of the number of unique transcripts detected by CEL-Seq (top) and scDam&T-seq for Dam and Dam-LMNB1 clones with varying DamID adapter concentrations.

In addition, scDam&T-seq and scDamID are similarly enriched on LADs in HT1080 cells[24] (Fig. S1c) and run-length analysis show similar prevalence of long stretches of genome–NL contacts in single cells (Fig. S1d). Finally, comparison of autocorrelation of in silico population samples show similar underlying genomic structures, with Dam-LMNB1 measuring larger structures than untethered Dam, as indicated by the lower rate of autocorrelation decay (Fig. S1e). Altogether these results show that scDam&T-seq successfully captures the distribution and variability of genome–NL interactions in single cells. The median scDam&T-seq complexity of 42,192 unique DamID reads per cell, is approximately four-fold reduced compared to scDamID (Fig. S1f). This difference may be attributed to greater sequencing depth in combination with selection and manual library preparation of single cells with the highest methylation levels for scDamID, as opposed to unbiased high-throughput preparation of scDam&T-seq libraries (Fig. S1f). Besides increased throughput, linear amplification of the DamID-products reduced the loss of reads resulting from incorrect adapter sequences (Fig. S1g) and a more accurate genome-wide distribution of GATC fragments (Fig. 1c).

Next, we benchmarked the transcriptomic measurements from scDam&T-seq to previously obtained CEL-Seq data for KBM7 cells[2]. Both methods detected the expression of a comparable number of genes (median: CEL-Seq = 2,508.5, scDam&T-seq = 2,282.5) (Fig. 1d) and unique transcripts (median: CEL-Seq = 4,920, scDam&T-seq = 4,009.5) (Fig. S2a). Transcriptomes measured by scDam&T-seq and CEL-Seq show a high degree of correlation (Fig. S2b, left panel) and display comparable single-cell variations indicated by the fraction of cells with detected genes (Fig. S2c, left panel), as well as by the relationship between mean gene expression and the coefficient of variation (Fig. S2d). These correlations are similar when comparing independent scDam&T-seq libraries (Fig. S2b,c, right panels). We observe batch effects between clones, libraries and methods (Fig. S2e). Principle component analysis to quantify batch effects in CEL-Seq and scDam&T-seq libraries showed that 16% of the total variance in transcriptional profiles can be attributed to differences between methods (scDam&T-seq and CEL-Seq), 9.7% is explained by clonal origin (Dam versus Dam-LMNB1) and 2.2% can be ascribed to differences between libraries (see Methods for details). Lastly, the overall efficiency and characteristics of mRNA detection are very similar to those of CEL-Seq (Fig. 1e and Fig. S2f,g), yet appear to reduce with increasing gDNA adapter concentrations (Fig. 1e). However, no correlations were found between the DamID and mRNA detection efficiencies within each condition (Fig. S2h). Since lowering the double-stranded adapter concentrations does not affect DamID complexity (Fig. S1f), mRNA detection may be further improved with reduced double-stranded adapter concentrations. In conclusion, scDam&T-seq produces single-cell data that are qualitatively and quantitatively comparable to its uncombined counterparts.

We also established scDam&T-seq in hybrid (129/Sv:CAST/EiJ) mouse embryonic stem cells (mESCs)[25] with auxin-inducible conditional DamID expression[26] (Fig. S3a). The median complexity of the scDam&T-seq libraries in mESCs is comparable to KBM7 cells (Supplementary Table 1) and strong overlap of DamID signal between the Dam-LMNB1 expressing mESCs and published Dam-LMNB1 bulk data[27] validates the applicability of scDam&T-seq to different cell types (Fig. S3b).

The untethered Dam enzyme was previously reported to accurately mark accessible chromatin[28]. We therefore wished to test the applicability of scDam&T-seq to quantify DNA accessibility and transcriptomes in single cells. We first quantified the levels of Dam methylation at transcription start sites (TSSs) and observed a sharp peak of Dam signal that scaled in accordance with increasing gene expression levels (Fig. 2a). Similar experiments with AluI digestions did not show signatures of accessibility around TSSs of actively transcribed genes (Fig. 2b), indicating that the observed Dam accessibility patterns are the result of in vivo Dam methylation at accessible regions of the genome and not restriction enzyme accessibility. We also observed strong Dam enrichment at active enhancers (Fig. 2c). Nucleosomes are regularly spaced around genomic elements like CTCF sites, which is a feature also observed in the scDam&T-seq data obtained with untethered Dam (Fig. 2d). The observed periodicity of 174 base pairs (bp) is in agreement with the reported spacing of nucleosomes in human cells[29,30] (Fig. S4a). Remarkably, the same periodicity is also apparent in single-cell samples (Fig. 2e), indicating that Dam can serve to determine nucleosome positioning in single cells in vivo.

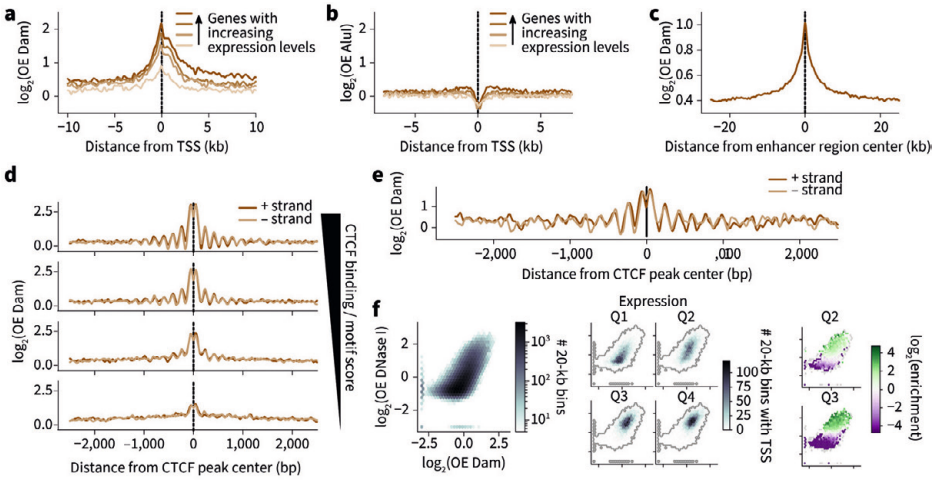


Figure 2: Untethered Dam marks accessible chromatin in single cells
a, Log-transformed OE values (log_2OE) of the Dam signal from an in silico population sample on TSSs of genes grouped into four equal-sized categories with increasing expression levels (ordered light to dark). **b**, log_2OE values obtained from AluI-derived fragments for identical TSSs as in **a**. **c**, log_2OE values of the Dam signal from an in silico population sample at active enhancers (see Methods for more details defining active enhancers). **d**, log_2OE values of the Dam signal from an in silico population sample at CTCF sites, stratified in four regimes of increasing CTCF binding activity (see Methods for details on stratification). **e**, Example of the log_2OE Dam signal of a single-cell sample at CTCF sites with the highest CTCF binding activity. **f**, Relation between DNase I (y axis) and in silico population Dam data (x axis): left, density of genomic 20-kb bins; middle, density of 20-kb bins with (one or more) TSSs of a gene, stratified in four gene expression quartiles from lowest (Q1) to highest (Q4) expression; right, significant enrichment (green) and depletion (purple) of transcribed 20-kb regions for the two expression quantiles (Q2 and Q3). Points in the plot with fewer than 10 20-kb bins were kept gray, as well as (statistically) insignificant enrichments/depletions (see Methods). The axes of the left panel also apply to plots in the middle and right panels.

scDam&T data correlate strongly with DNase I at open chromatin, but less at relatively condensed chromatin, where Dam distinguishes between a larger range of chromatin accessibilities (Fig. 2f, left). This increased sensitivity is functionally related to genes with low expression levels. Stratifying genes into four expression quantiles, shows a strong depletion of DNase I marked regions of the second expression quantile as opposed to moderate Dam signal for the same genomic regions (Fig. 2f, middle and right). This increased sensitivity of Dam in measuring lowly transcribed gene regions may be attributed to the ability of Dam to mark gene-units encompassing both active gene promoters (marked by H3K4me3) and gene bodies (marked by H3K36me3) (Fig. S4b), whereas DNase I has been reported to primarily detect active promoters[31]. Finally, we compared scDam&T-seq in mESCs cells to scNMT-seq: a method for single-cell detection of 5-methylcytosine (5mC), chromatin accessibility and mRNA[19]. scDam&T-seq and scNMT-seq display similar nucleosome positioning characteristics at DNase I hypersensitivity sites, with a 30-fold shallower sequencing depth for scDam&T-seq (Fig. S4c). The numbers of detected genes are also very similar between methods at comparable sequencing depths (Fig. S4d). scDam&T-seq, therefore, provides data quality similar to scNMT-seq, yet at greatly reduced sequencing depth.

We next determined the single-cell associations of genome–NL contacts or chromatin accessibility with gene expression in mESCs. First, the scDamID profiles were converted into binary contact maps as previously described[2] (Fig. 3a, step 1). For the untethered Dam enzyme, regions of high CF indicate transcriptional active open chromatin configurations, while high CF regions for Dam-LMNB1 indicate an association with the NL and therefore a repressed chromatin state. Previously in KBM7 cells, the frequency with which genomic regions associate with the NL was shown to inversely correlate with gene activities[2]. Indeed, in mESCs, we observe that mean expression levels gradually drop with increased genome–NL CFs (Fig. 3b, left). In contrast and as expected, increased Dam CFs positively correlate with mean gene expression levels (Fig. 3b, right). To investigate the impact of genome–NL contacts and chromatin accessibility on gene expression in single cells, we determined the $\log_2$(fold change) in expression ($\log_2$FC) in cells showing contact and no-contact states per genomic bin (Fig. 3a, steps 2 and 3). Intriguingly, a genome-wide negative association between genome contact and expression was observed for Dam-LMNB1, and a positive association for the untethered Dam (Fig. 3c). Thus, cell-to-cell variations in genome–NL contacts impact on gene expression; regions are more likely to be active in those cells where they are detached from the NL. The positive association between $\log_2$FC in expression and contact with Dam indicates that, between single cells, a genomic region is more likely active when in an open chromatin state. These single-cell associations are largely independent of mean expression levels and expression variance (Fig. S5a–d). Interestingly, the negative relationship between genome–NL contact and gene expression is only observed for genomic regions that infrequently associate with the NL (Fig. 3d, left panel), while genes residing within medium to high open chromatin are transcriptionally most sensitive to changes in chromatin accessibility (Fig. 3d, right panel). The small effect size between the associations of Dam and Dam-LMNB1 with transcription could be resulting from the limited time resolution of these experiments (12 h) and/or the effect of

the relatively large 100-kb bins. A cell line with elevated Dam-expression levels combined with more rapid inducibility may improve this. These data suggest that genomic regions that typically reside in the nucleoplasm are most sensitive to occasional NL association, and that genes respond differently to changes in accessibility depending on their chromatin contexts. Interestingly, the LADs in the low CF range are relatively depleted of constitutive chromatin marked by H3K9me3 and enriched for the facultative heterochromatin modification H3K27me3 (Fig. S5e, top). Consistently, the chromatin state of the low CF regions is enriched for cell-type-specific (facultative) fLADs, as opposed to cell-type invariant (constitutive) cLADs (Fig. S5f, top). The opposite patterns can be observed for the Dam contact regions (Fig. S5e,f, bottom). Collectively, these observations suggest that fLADs are more susceptible to dissociation from the NL and subsequent transcriptional activation compared to the H3K9me3-enriched cLADs.

We next investigated how DNA accessibility relates to gene expression at an allelic resolution. First, to account for potential allelic copy number variations (CNVs) that would introduce biases in our analysis, we performed single-cell reduced-representation whole-genome sequencing by substituting DpnI with AluI in the scDam&T-seq protocol (Fig. S6a). Chromosomes 5, 8 and 12 were found frequently (partially) duplicated or lost and were excluded from our analyses (Fig. S6a). For the Dam data, approximately 45% of reads could be attributed to either allele, and the same CNVs were apparent in the resulting allelic single-cell chromatin accessibility tracks (Fig. S6b). Surprisingly, we also detected a small fraction of cells that displayed a reverse DNA accessibility bias on chromosome 12, and a corresponding allelic bias in transcription for one cell (Fig. S6c). After excluding chromosomes with frequent CNVs as well as samples showing a CNV on any other chromosome, we found a positive allelic single-cell association between chromatin accessibility and transcription (Fig. S6d). Therefore, scDam&T-seq can be employed to investigate single-cell allelic relationships between expression and chromatin states.

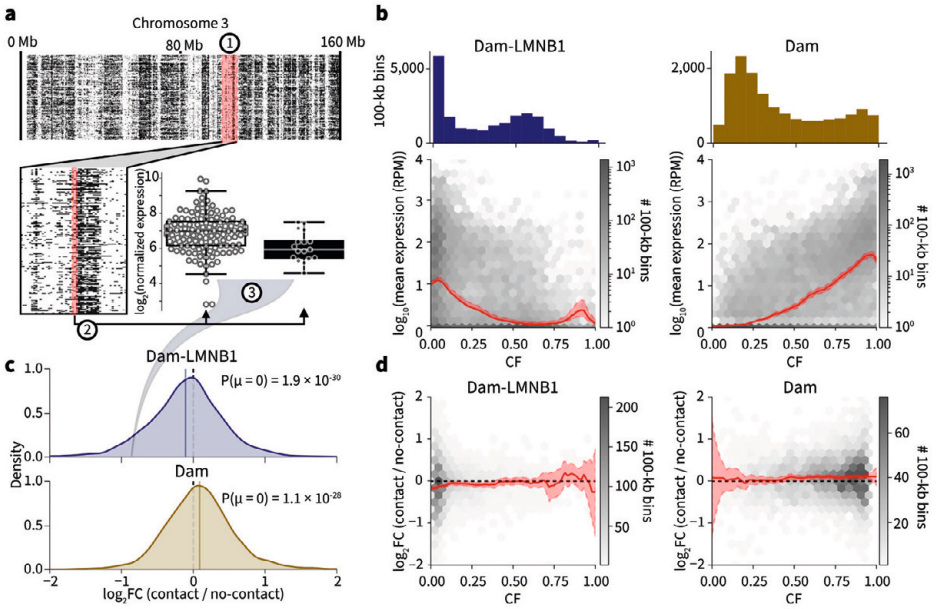


Figure 3: Parallel transcriptomic and DamID measurements link transcriptional dependencies to heterogeneity in DamID contacts

a, Schematic of analysis to determine the $\log_2$FC in transcription between contact and no-contact states. (1) Per genomic bin (of 100 kb), the single-cell samples are binarized into two groups, having either high (OE ≥ 1, black) or low (OE < 1, white) DamID signal, corresponding to a DamID contact and no contact, respectively. (2) The expression of the two groups of samples in that genomic bin is computed, and (3) a group-wise $\log_2$FC in expression is calculated. The example shows one bin on chromosome 3 where mESC Dam-LMNB1 contact is associated with a decrease in expression of about two-fold (−1 $\log_2$FC) compared to no Dam-LMNB1 contact. The example bin displays NL contacts in 17 out of 143 single cells and $\log_2$FC is determined based on the expression of genes in the 100-kb bin (containing two expressed genes). Box plots indicate the 25th and 75th percentile (box), median (line) and 1.5 times the interquartile range (IQR) past the 25th and 75th percentiles (whiskers). Data points are overlaid as circles. n = 126 and n = 17, in left and right box, respectively. **b**, Relation between expression (y axis) and CF (x axis) defined as the fraction of cells that show high DamID signals (OE ≥ 1) across 100-kb genomic bins. Dam-LMNB1 (left) and Dam (right) are shown, as well as the genome-wide distribution of CF values across mappable bins (histogram on top). The solid line indicates the mean, shaded area indicates 1.96 times the standard deviation around the mean. **c**, Distribution of expression $\log_2$FC values with Dam-LMNB1 (top) and Dam (bottom), genome-wide across 100-kb bins. Note that only 100-kb bins with at least three single-cell samples in both groups, and having expression in at least 20% of the single-cell samples were included in the analysis. P values of a two-sided one-sample t-test are indicated. **d**, Relation between expression $\log_2$FC values and CF, for Dam-LMNB1 (left) and Dam (right). The red shadings indicate 95% confidence intervals. The solid line indicates the mean, the shaded area indicates 1.96 times the standard deviation around the mean.

Next, we established scDam&T-seq as an in silico cell sorting strategy to identify and group cell types based on their transcriptomes and uncover the underlying cell-type-specific gene-regulatory landscapes from DamID data. We first performed a scDam&T-seq proof-of-principle experiment on mESCs cultured under 2i or serum conditions. scDam&T-seq derived transcriptomics were separated into two distinct clusters based on independent component analysis (ICA, Fig. 4a). Expression analysis showed signature genes differentially expressed between the two conditions (Fig. S7a). DNA accessibility profiles generated from the two in silico transcriptome clusters showed differential accessibility patterns on a genome-wide scale. *Peg10*, a gene strongly upregulated under serum conditions, showed increased accessibility at the TSS and along the gene body (Fig. 4b). Interestingly, this increased accessibility stretches beyond the *Peg10* gene locus, encompassing a large topologically associating domain (TAD). Genome-wide TAD analysis reveals that global changes in chromatin accessibility between 2i and serum conditions occur more within TAD domains than for randomized domains of the same size (Fig. S7b). Thus, chromatin relaxation of the TAD that encompasses *Peg10* in serum conditions is illustrative of a broader phenomenon occurring within the genome-wide TAD framework. At the gene level, differential upregulation in either 2i or serum conditions is also associated with increased DNA accessibility (Fig. 4c and Fig. S7c). Interestingly, the increased accessibility at the TSS extends into the gene body (Fig. S7d). The same increased accessibility is also observed in single cells for the top five differentially expressed genes between conditions (Fig. S7e). Together, these results demonstrate that scDam&T-seq can be used to effectively generate cell-type-specific DNA accessibility profiles from heterogeneous mixtures of cells, based on in silico identification and grouping of cell types.

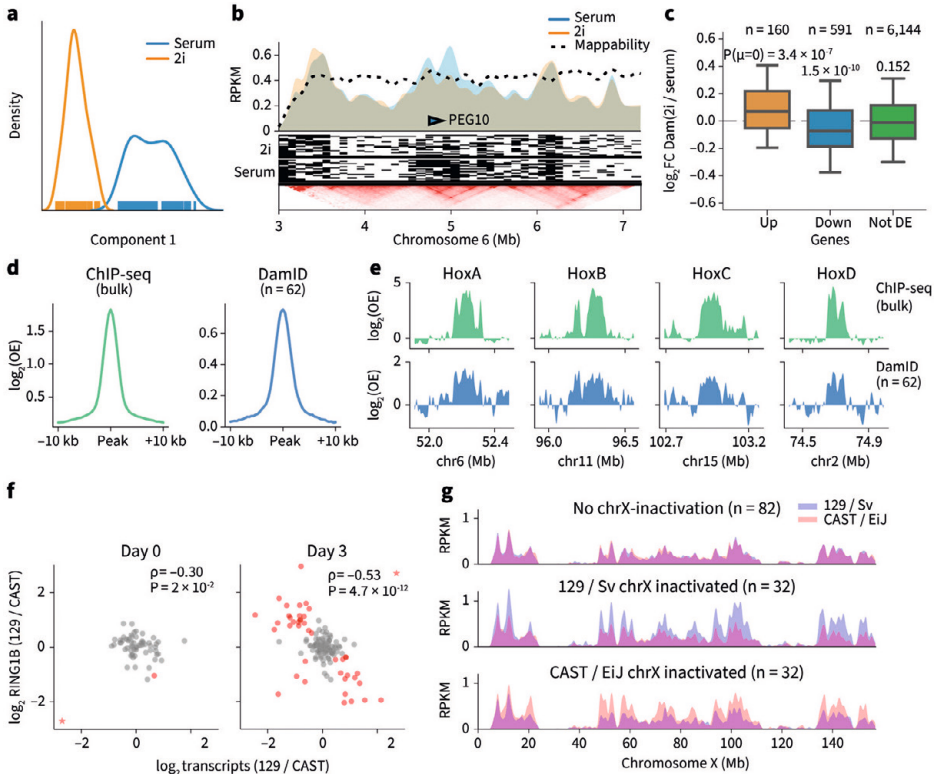


Figure 4: scDam&T-seq enables in silico cell sorting and reconstruction of corresponding cell-type-specific gene-regulatory landscapes

a, ICA on Dam-expressing mESCs cultured in 2i (orange) or serum (blue) conditions. **b**, DNA accessibility profiles in 2i and serum conditions of in silico populations (top track) and single cells (signal binarized to high (OE ≥ 1) as black and low (OE < 1) as white). The lower panel shows mESC HiC data[34] at the same locus, displayed with the 3D genome browser[35]. **c**, Fold change in the Dam signal (reads per million, RPM) between 2i and serum conditions for genes that show statistically significant upregulation (orange), downregulation (blue) or are unaffected (DE, green) in 2i conditions compared to serum. Box plots indicate the 25th and 75th percentile (box), median (line) and 1.5 times the IQR past the 25th and 75th percentiles (whiskers). P values indicate the result of a two-sided t-test against a mean of 0. n = 158, 577 and 6,056 genes, in boxes from left-to-right, respectively. **d**, Average $\log_2$OE signal over all RING1B ChIP-seq peaks obtained with ChIP-seq (left) and scDam&T-seq (right) in 2-kb bins. **e**, Signal ($\log_2$OE) over the four HOX gene clusters for RING1B ChIP-seq and RING1B DamID. In **d** and **e**, population ChIP-seq data were normalized for the corresponding input control; RING1B DamID data represent an in silico population of 62 single cells and were normalized with an in silico population Dam sample. **f**, Relationship between allelic bias in transcription and DamID on chromosome X. Spearman's ρ and P values (two-sided test, determined by bootstrap) are indicated. Cells are indicated in red when both the transcriptional and DamID allelic biases deviated more than expected based on the somatic chromosomes (see Methods). Cells marked as a star fell outside the shown data range; the cell marked as a star in the serum condition is suspected of having lost one chromosome X allele and was excluded from the Spearman correlation. **g**, Average allelic DamID profiles for cells that had a transcriptional bias on chromosome X toward neither allele (top), toward 129/Sv (middle) or toward CAST/EiJ (bottom) for chromosome X.

Finally, to further test the in silico sorting strategy to profile gene-regulatory landscapes, we chose the polycomb-repressive-complex 1 (PRC1) subunit RING1B (RNF2), which is responsible for the ubiquitination of histone H2AK119[32]. Because of the role of PRC1 and 2 complexes in the regulation of X chromosome inactivation, we tested whether scDam&T-seq can be employed to identify the randomly inactivated allele in combination with RING1B occupancy in single cells. In undifferentiated mESCs, the cumulative single-cell RING1B scDam&T-seq data are strongly enriched over RING1B binding sites detected by ChIP-seq (Fig. 4d). Similarly, the patterns of enrichment on *HOX* genes are very comparable (Fig. 4e) and genome-wide scDam&T-seq and ChIP-seq correlate well (Fig. S7f). At day 3 of differentiation, random X inactivation is apparent in a fraction of single cells based on the ratio of allelic expression on chromosome X, a pattern that is not observed for autosomal transcripts (Fig. S7g). The allelic bias in transcription correlates with increased RING1B levels on the transcriptionally repressed allele (Fig. 4f,g), a pattern that is not observed for autosomes of the same cells (Fig. S7h). The observed increased levels of RING1B on the inactive X chromosome are consistent with the identification of H2AK119 ubiquitination as one of the earliest events during X inactivation[33] (Fig. S7i). These results demonstrate that scDam&T-seq can be employed to systematically dissect the regulatory mechanisms underlying X chromosome inactivation in single cells.

In summary, scDam&T-seq allows simultaneous quantifications of DNA–protein interactions and transcription from single cells. We have shown that scDam&T-seq enables measuring the impact of spatial genome organization and chromatin states on gene expression and it can be applied to sort cell types in silico and obtain their associated gene-regulatory landscapes. Applied to dynamic biological processes, scDam&T-seq should prove especially powerful to identify protein-mediated mechanisms that regulate cell-type-specific transcriptional programs in dynamic processes and heterogeneous tissues.

Acknowledgements

We would like to thank the members of the Kind, Dey and van Oudenaarden laboratories for their comments on the manuscript and J. Gribnau (Erasmus UMC) for kindly providing the 129/Sv:CAST/EiJ mESCs and for advice on differentiation. We would like to thank B. de Barbanson and J. Yeung for suggestions regarding computational analyses and statistics, R. van der Linden for FACS and M. Muraro and L. Kester for input on the scDam&T-seq technique. S.S.D and A.C. received computational support from the Center of Scientific Computing at UCSB based on funding from NSF MRSEC (DMR-1720256) and NSF CNS 1725797. This work was funded by a European Research Council Starting grant (no. ERC-StG 678423-EpiID) and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Open grant (no. 824.15.019) and ALW/VENI grant (no. 016. Veni.181.013). The Oncode Institute is supported by the KWF Dutch Cancer Society

Author contributions

K.R. and C.M.M. contributed equally as first authors. F.J.R. and S.S.d.V. contributed equally as second authors. K.R., S.S.D. and J.K. designed the study and wrote the manuscript. S.S.D. developed the method. C.M.M. optimized the method and performed all the experiments unless stated otherwise. S.S.d.V. and K.L.d.L. assisted with experiments. S.S.d.V. created the mESC lines. K.R. performed all analyses except when stated otherwise. F.J.R. performed cloning and all analyses pertaining to the RING1B data. A.C. performed the analysis of Fig. S4d and exploratory analyses together with S.S.D. D.M. provided input during initial technology development. J.K. and S.S.D conceived and supervised the study.

Methods

Cell culture

Haploid KBM7 cells were cultured in suspension in IMDM (Gibco) supplemented with 10% FBS (Sigma) and 1% Pen/Strep (Gibco). Shield1-inducible Dam-LMNB1 and Dam stable clonal KBM7 cell lines were used as described previously[2]. Cells were split every 3 d. F1 hybrid 129/Sv:Cast/EiJ mESCs[25] were cultured on irradiated primary mouse embryonic fibroblasts (MEFs), in ES cell culture media; G-MEM (Gibco) supplemented with 10% FBS (Sigma), 1% Pen/Strep (Gibco), 1× GlutaMAX (Gibco), 1× non-essential amino acids (Gibco), 1× sodium pyruvate (Gibco), 0.1 mM β-mercaptoethanol (Sigma) and 10^3 U ml^{-1} ESGROmLIF (EMD Millipore, ESG1107). Cells were split every 3 d. Expression of constructs was suppressed by the addition of 1 mM indole-3-acetic acid (IAA; Sigma, I5148). 2i F1 hybrid 129/Sv:Cast/EiJ mESCs were cultured for 2 weeks on primary MEFs in 2i ES cell culture media; 48% DMEM/F12 (Gibco) and 48% Neurobasal (Gibco), supplemented with 1× N2 (Gibco), 1× B27 supplement (Gibco), 1× non-essential amino acids, 1% Pen/Strep, 0.1 mM β-mercaptoethanol, 0.5% bovine serum albumin (Sigma), 1 µM PD0325901 (Axon Medchem, 1408), 3 µM CHIR99021 (Axon Medchem, 1386) and 10^3 U ml^{-1} ESGROmLIF. Cells were split every 3 d. Expression of constructs was suppressed by addition of 1 mM IAA (Sigma). The stable mESC clones were differentiated by culturing on gelatin-coated six-well plates after MEF depletion, in monolayer differentiation media; IMDM supplemented with 15% FBS, 1% Pen/Strep, 1× GlutaMAX, 1× non-essential amino acids (Gibco), 50 µg ml^{-1} ascorbic acid (Sigma, A4544) and 37.8 µl l^{-1} monothioglycerol (Sigma, M1753). Expression of constructs was suppressed by addition of 1 mM IAA. After MEF depletion, one confluent six well of mESCs was split 1:15 on six gelatin-coated wells of a six-well plate in differentiation media for 3 d. The medium was changed every other day.

Generating cell lines

Stable clonal Dam and Dam-LMNB1 F1 hybrid mESC lines were created by co-transfection of the EF1α-Tir1-IRES-neo and hPGK-AID-Dam-mLMNB1 or hPGK-AID-Dam plasmids in a ratio of 1:5. Cells were trypsinized and 0.5×10^6 cells were plated directly with Effectene transfection mixture (Qiagen) in 60% buffalo rat liver (BRL)-conditioned medium; 120 ml of BRL medium (in-house production), 80 ml of G-MEM (Gibco) supplemented with 10% FBS, 1% Pen/Strep,

1× GlutaMAX, 1× non-essential amino acids, 1× sodium pyruvate, 0.1 mM β-mercaptoethanol and 10^3 U ml^{-1} ESGROmLIF on gelatin-coated wells of a six-well plate. The transfection was performed according to the Effectene protocol (Qiagen). Cells were selected for 10 d with 250 µg ml^{-1} G418 (Thermofischer) and selection of the clones was based on methylation levels, determined by DpnII-qPCR assays as described previously[2]. To reduce the background methylation levels in the presence of 1 mM IAA, we transduced the selected clones of both AID-Dam-LMNB1 and Dam-only with extra hPGK-Tir1-puro lentivirus followed by selection with 0.8 µg ml^{-1} puromycin. Positive clones were screened for IAA induction in the presence and absence of IAA by DpnII-qPCR assays and DamID PCR products as previously described[2]. Stable clonal AID-Dam-RING1B F1 hybrid mESCs were created by lentiviral co-transduction of pCCL-EF1α-Tir1-IRES-puroR and pCCL-hPGK-HA-AID-Dam-RING1B virus in a 4:1 ratio, after which the cells were selected for 10 d on gelatin-coated 10 cm dishes in BRL-conditioned medium containing 0.8 µg ml^{-1} puromycin (Sigma) and 0.5 mM IAA. Individual puromycin-resistant colonies were tested for the presence of the constructs by PCR using primers fw-ttcaacaaaagccaggatcc and rev-gacagcggtgcataaggcgg. Positive clones were screened further for their level of induction on IAA removal by DamID PCR products.

DamID induction

Expression of Dam-LMNB1 and Dam constructs was induced in the KBM7 cells with 0.5 nM Shield1 (Glixx laboratories, 02939) 15 h before harvesting as described previously[2]. Expression of Dam-LMNB1 or Dam constructs was induced in the F1 mESCs by IAA washout with PBS (in-house production) 12 h before harvesting. Based on the growth curve of cells counted at time points 12, 24, 30, 36, 42, 48, 54, 60, 72 and 84 h after plating, the generation time of both the Dam-LMNB1 and Dam cell lines was estimated at 12 h (data not shown). Considering that 55% of the cells are in G1 and early S phase, the estimated time these cells reside in G1 and early S phase is 6.75 h.

Cell harvesting and sorting

KBM7 cells were harvested in PBS (in-house production), stained with 0.5 µg ml^{-1} 4,6-diamidino-2-phenylindole (DAPI, Sigma) for live/dead selection. Single cells were sorted based on small forward and side-scatter values (30% of total population) and selected for double positive Fucci profile as described previously[2,36]. F1 mESCs expressing Dam-LMNB1, Dam or Dam-RING1B were collected in plain or 2i ES cell culture media and stained with 30 µg ml^{-1} Hoechst 34580 (Sigma, 63493) for 45 min at 37 °C. mESC singlets were sorted based on forward and side scatter properties, and in mid-S phase of the cell cycle based on DNA content histogram. Differentiated F1 mESCs expressing Dam-RING1B were collected in differentiation media and stained with 30 µg ml^{-1} Hoechst 34580 for 45 min at 37 °C. The same cells were stained with 1 µg ml^{-1} propidium iodide (Sigma) for live/dead selection. Differentiated mESCs singlets were sorted based on forward and side scatter properties, and in G1, S and G2/M phase of the cell cycle based on DNA content histogram. One cell was sorted per well of 384-well plates (Biorad, HSP3801) using the BD FACSJazz cell sorter. Wells contained 4 µl of mineral oil (Sigma) and 100 nl of 15 ng $µl^{-1}$ unique CEL-Seq2 primer[23].

Robotic preparation of scDam&T-seq

Mineral oil (4 µl) was dispensed manually into each well of a 384-well plate using a multichannel pipette and 100 nl of unique CEL-Seq primer was dispensed per well using a Mosquito HTS robot (TTP Labtech). The NanodropII robot (BioNex) was used for all subsequent dispensing steps at 12 psi pressure. After sorting, 100 nl of lysis mix was added (0.8 U RNase inhibitor (Clontech, 2313 A), 0.07% Igepal, 1 mM dNTPs, 1:500,000 ERCC RNA spike-in mix (Ambion, 4456740)). Each single cell was lysed at 65 °C for 5 min and 150 nl of reverse transcription mix was added (1× First Strand Buffer (Invitrogen, 18064-014), 10 mM DTT (Invitrogen, 18064-014), 2 U RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, 10777019), 10 U SuperscriptII (Invitrogen, 18064-014)) and the plate was incubated at 42 °C for 1 h, 4 °C for 5 min and 70 °C for 10 min. Next, 1.92 µl of second strand synthesis mix was added (1× second strand buffer (Invitrogen, 10812014), 192 µM dNTPs, 0.006 U *E. coli* DNA ligase (Invitrogen, 18052019), 0.013 U RNAseH (Invitrogen, 18021071)) and the plate was incubated at 16 °C for 2 h. 500 nl of protease mix was added (1× NEB CutSmart buffer, 1.21 mg ml^{-1} ProteinaseK (Roche, 000000003115836001)) and the plate was incubated at 50 °C for 10 h and 80 °C for 20 min. Next, 230 nl of DpnI mix was added (1× NEB CutSmart buffer, 0.2 U NEB DpnI) and the plate was incubated at 37 °C for 4 h and 80 °C for 20 min. Finally, 50 nl of DamID2 adapters were dispensed (final concentrations varied between 32 and 128 nM), together with 450 nl of ligation mix (1× T4 Ligase buffer (Roche, 10799009001), 0.14 U T4 Ligase (Roche, 10799009001)) and the plate was incubated at 16 °C for 12 h and 65 °C for 10 min. Contents of all wells with different primers and adapters were pooled and incubated with 0.8 volume magnetic beads (CleanNA, CPCR-0050) diluted 1:4 or 1:8 with bead binding buffer (20% PEG8000, 2.5 M NaCl) for 10 min, washed twice with 80% ethanol and resuspended in 7 µl of nuclease-free water before in vitro transcription at 37 °C for 14 h using the MEGAScript T7 kit (Invitrogen, AM1334). Library preparation was done as described in the CEL-Seq protocol with minor adjustments[23]. Amplified RNA (aRNA) was cleaned and size-selected by incubating with 0.8 volume magnetic beads (CleanNA) for 10 min, washed twice with 80% ethanol and resuspended in 22 µl of nuclease-free water, and fragmented at 94 °C for 2 min in 0.2 volume fragmentation buffer (200 mM Tris-acetate pH 8.1, 500 mM KOAc, 150 mM MgOAc). Fragmentation was stopped by addition of 0.1 volume fragmentation STOP buffer (0.5 M EDTA pH 8) and quenched on ice. Fragmented aRNA was incubated with 0.8 volume magnetic beads (CleanNA) for 10 min, washed twice with 80% ethanol and resuspended in 12 µl of nuclease-free water. Thereafter, library preparation was done as previously described[23] using 5 µl of aRNA and PCR cycles varied between 8 and 10. Libraries were run on the Illumina NextSeq platform with high output 75 bp paired-end sequencing.

DamID adapters

The adapter was designed (5′ to 3′) with a 4 nucleotide (nt) fork, a T7 promoter, the 5′ Illumina adapter (as used in the Illumina small RNA kit), a 3 nt UMI, an 8 nt unique barcode followed by CA. The Dam-RING1B mESCs were processed with different adapters. These contained a 6 nt fork, a 6 nt unique barcode followed by GA. The barcodes were designed with a Hamming distance of at least 2 between them. Bottom sequences contained a phosphorylation site at the 5′ end. Adapters were produced as standard desalted oligos. Top and bottom sequences were

annealed at a 1:1 volume ratio in annealing buffer (10 mM Tris pH 7.5–8.0, 50 mM NaCl, 1 mM EDTA) by immersing tubes in boiling water, then allowing them to cool to room temperature. The oligo sequences can be found in Supplementary Table 2.

CEL-Seq primers

The RT primer was designed according to the Yanai protocol[23] with an anchored polyT, an 8 nt unique barcode, a 6 nt UMI, the 5′ Illumina adapter (as used in the Illumina small RNA kit) and a T7 promoter. The barcodes were designed with a Hamming distance of at least 2 between them. Primers are desalted at the lowest possible scale, stock solution 1 µg/µL. The oligo sequences can be found in Supplementary Table 3.

Raw data preprocessing

First mates in the raw read pairs (that is, 'R1' or 'read1') conform to a layout of either: 5′-[3 nt UMI][8 nt barcode]CA[gDNA]-3′ in the case of gDNA (DamID and AluI restriction) reads, or 5′-[6 nt UMI][8 nt barcode][unalignable sequence]-3′ in the case of transcriptomic reads. In the case of transcriptomic reads, the second mate in the read pair contains the mRNA sequence. Raw reads were processed by demultiplexing on barcodes (simultaneously using the DamID and transcriptomic barcodes), allowing no mismatches. The UMI sequences were extracted and stored alongside the names of the reads for downstream processing.

Sequence alignments

After demultiplexing of the read pairs using the first mate and removal of the UMI and barcode sequences, the reads were aligned. In the case of gDNA-derived reads, a 'GA' dinucleotide was prepended to the sequences of read1 ('AG' in the case of AluI), and the gDNA sequence of read1 was then aligned to a reference genome using bowtie2 (v.2.3.2) with the parameters: seed 42, very-sensitive-N 1. For transcriptome-derived reads, read2 was aligned using tophat2 (v.2.1.1) with the parameters: segment-length, 22; read-mismatches, 4; read-edit-dist, 4; min-anchor, 6; min-intron-length, 25; max-intron-length, 25,000; no-novel-juncs; no-novel-indels; no-coverage-search; b2-very-sensitive; b2-N 1; b2-gbar 200 and using transcriptome-guiding (options GTF and transcriptome-index). Human data were aligned to hg19 (GRCh37) including the mitochondrial genome, the sex chromosomes and unassembled contigs. Transcriptomic reads were aligned using transcript models from GENCODE (v.26) (https://www.gencodegenes.org/human/grch37_mapped_releases.html). mESC data were aligned to reference genomes generated by imputing 129S1/SvImJ and CAST/EiJ single nucleotide polymorphisms obtained from the Sanger Mouse Genomes project[37] onto the mm10 reference genome. The mitochondrial genome, sex chromosome and unassembled contigs were included during the alignments. Transcriptomic reads were aligned using a GTF file with transcript annotations obtained from ENSEMBL (release 89) (ftp://ftp.ensembl.org/pub/release-89/gtf/mus_musculus/Mus_musculus.GRCm38.89.gtf.gz). Both human and mouse transcriptome references were supplemented with ERCC mRNA spike-in sequences (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_095047.txt). For both genomic and transcriptomic data, reads that yielded an alignment with mapping quality (BAM field 'MAPQ') lower than 10 were discarded,

as well as reads aligning to the mitochondrial genome or unassembled contigs. For the genomic data, reads not aligning exactly at the expected position (5′ of the motif, either GATC in the case of DpnI restriction or AGCT in the case of AluI restriction) were discarded. For the transcriptomic data, reads not aligning to an exon of a single gene (unambiguously) were discarded. The mESC reads were assigned to the 129S1/SvImJ or CAST/EiJ genotype by aligning reads to both references. Reads that aligned with lower edit distance (SAM tag 'NM') or higher alignment score (SAM tag 'AS') in case of equal edit distance to one of the genotypes were assigned to that genotype. Reads aligning with equal edit distance and alignment score to both genotypes were considered of 'ambiguous' genotype. For analyses comparing allelic signals, counts with 'ambiguous' genotype were discarded (Fig. 4f,g, Fig. S6, Fig. S7g,h). For all other figures concerning mESC data, UMI-unique data of the two alleles were summed together with the ambiguously assigned data.

PCR duplicate filtering

For the genomic data (DamID and AluI-WGS), the number of reads per motif, strand and UMI were counted. Read counts were collapsed using the UMIs (that is, multiple reads with the same UMI count as 1) after an iterative filtering step where the most abundant UMI causes every other UMI sequence with a Hamming distance of 1 to be filtered out. for example, observing the three UMIs 'AAA', 'GCG' and 'AAT' in decreasing order would count as two unique events (with UMIs 'AAA' and 'GCG', since 'AAT' is within 1 Hamming distance from 'AAA'). The number of observed unique UMIs was taken as the number of unique methylation events (for DamID) or unique transcripts (for the transcriptomics). For the data from KBM7 (a near-complete haploid cell line) at most one unique event per GATC position and strand was kept. For the mESC data at most one unique event per GATC position, strand and allele were kept, or two unique events, in the case of 'ambiguous' allelic assignment.

Filtering of samples

We observed that the number of unique methylation events and unique transcripts per single-cell sample followed a bimodal distribution in most data sets. To discard samples that clearly failed, we applied the following cutoffs: only single-cell samples with at least $10^{3.7}$ unique DamID events and at least 10^{3} unique transcripts were taken into consideration for the analyses. These cutoffs were applied jointly for all analyses, regardless of whether genomic and/or transcriptomic signals were used. These numbers were established on our earliest (human and mouse) data sets, by fitting a two-component Gaussian mixture model to the observed unique counts (with all samples across the data sets).

Normalization of expression values

UMI-unique transcript counts per gene were further normalized using scran[38,39]. We used computeSumFactors with reduced sizes parameter where our sample sizes were too small for default parameters and using only genes expressed in at least 1% of all samples, and other parameters were left to their default values. Expression values were then converted to log-transformed counts per million (TPM, transcripts per million reads) using logcounts.

Binning and calculation of OE values

DamID and WGS data were binned using consecutive non-overlapping 100-kb bins. For analyses at TSS, enhancer and CTCF sites, data were binned with high resolution (a bin size of 10 bp was used). To calculate OE values, the mappability of each motif (GATC or AGCT) was determined by generating sequences of 65 nt (in both orientations) from the reference genome(s) and aligning and processing them identically to the data. By binning the in silico generated reads, the maximum amount of mappable unique events per bin was determined.

OE values were calculated using

$$OE = \frac{O + \psi}{E + \psi} \times \frac{T_E + B\psi}{T_O + B\psi}$$

where O is the number of observed unique methylation events per bin, E is the number of mappable unique events per bin, ψ is the pseudocount (1, unless otherwise stated), T_O and T_E are the total number of unique methylation events observed and mappable, respectively in the sample and B is the number of bins. For analyses across multiple windows, for example, windows around TSSs or CTCF sites, O and E were summed across the windows, before calculating the OE values.

For the definition of 'contact', regions with OE values ≥1 were considered as 'in contact'. Further details and justification can be found in a previous report[2], in particular its Extended Experimental Procedures (section "Processing of single-cell DamID sequencing reads") and its Fig. S2a. CF was defined as the fraction of samples (passing cutoffs) showing 'contact' (OE ≥ 1) and is expressed as fraction in [0, 1] per genomic bin.

Comparison scDam&T-seq to Kind Cell 2015 data

For the comparisons with individual measurements of scDamID and single-cell transcriptomics (CEL-Seq)[2] with scDam&T-seq (Fig. 1), the scDam&T-seq data were made comparable to the published data by truncating the reads at the 3′ end such that gDNA and mRNA sequence lengths were identical to the published data, which were sequenced with shorter reads. Furthermore, UMIs were completely left out of the consideration for the DamID measurements. For the transcriptional measurements, the UMIs were truncated to 4 nt to make the data comparable to the published CEL-Seq data.

Signal of scDam&T-seq LMNB1 data on microarray-defined LADs

Comparisons of LMNB1 data obtained with scDam&T-seq to independently identified LADs (Fig. S1c for human data and Fig. S3b for mouse data) were made using published HT1080 (ref. 24) and mESC27 data. We used the LAD coordinates available from the processed data corresponding to ref. 24 at GEO (GSE22428) and Table S2 from ref. 27. We remapped LAD coordinates using liftOver (from mm9 to mm10 and from hg18 to hg19, for mouse and human data, respectively) and discarded LADs that spanned less than 500 kb after the liftOver procedure.

Run-length analysis

Run-length analysis was done as described previously[2] with the exception that we did not remove bins from the analysis with a CF of 0. Random shuffling with preservation of marginal distributions was done as described previously[4].

Autocorrelation analysis

Autocorrelation of raw signals was analyzed with a maximum resolution limited by a bin size of 100 bp. In silico population profiles were generated for each indicated condition and downsampled to 50 times the DamID methylation count cutoff of $10^{3.7}$. Only chromosomes larger than 100 Mb were considered in the analysis, as autocorrelation of large distances cannot be measured on shorter chromosomes. Furthermore, sex chromosomes were discarded. We used a FFT approach to determine the statistical autocorrelation of the signal at each chromosome, then summed the autocorrelation profiles to arrive at the genome-wide autocorrelation profiles.

Assessment of technical batch effects on variance in transcriptomics data

Principal component analysis on the transcriptome data shows that batch effects always appear in the first, or first few principal components. This is unsurprising since these single-cell samples are biologically homogeneous (for instance, clonal cells, FACS-sorted in the same cell phase). To assess to which degree technical effects influence variance in the transcriptomics data, we employed an approach analogous to Bushel 2008 (pvca: principal variance component analysis, R package v.1.22.0)[40], with the exception that we fitted simple ordinary least-squares models (with one factor) rather than mixed linear models. Weighing the coefficient of determination for the batch effect of each principal component with variance explained by the principal component a total of 16% of data variance can be explained by the method, between scDam&T-seq and CEL-Seq (Fig. S2d). For reference, 2.2% of total data variance can be explained by batch when contrasting two scDam&T-seq libraries, and 9.7% of total variance in expression data can be explained by clonal origin when contrasting Dam-LMNB1 and Dam transcriptomes measured by scDam&T-seq. Finally, we also used ComBat[41] to estimate the amount of data variance explained by these technical variables, by comparing the amount of data variance before and after removing 'batch effects'. We obtained similar ratios of variance explained but in general observe lower amounts of total data variance explained by batch (8.9% explained when using CEL-Seq versus scDam&T-seq as batch, 3.6% by clonal origin, 3.0% when contrasting two Dam-LMNB1 batches).

Using principal component analysis on our mESC 2i versus serum transcriptomics data, a high degree of separation was shown between 2i and serum samples on the first principal component, but also a strong association with sample depths (despite using best practices to normalize our single-cell transcriptomics data). We, therefore, employed a two-component ICA to deconvolve sample depth effects from the 2i/serum effects on the (normalized) transcriptomics data. The ICA separating 2i and serum samples is shown in Fig. 4a.

TSS, CTCF and enhancer locations

For the analyses at TSSs, one isoform per gene was chosen from the gene annotations, by preferentially taking isoforms that carry the GENCODE 'basic' tag, have a valid, annotated CDS (start and stop codon, and CDS length being a multiple of 3 nt), with ties broken by the isoform with the longest CDS, and shortest gene length (distance from 5′ nucleotide of first exon to 3′ nucleotide of last exon). As TSS, the most 5′ position of the first exon was taken. CTCF sites were obtained by integrating ENCODE ChIP-seq data (wgEncodeRegTfbsCellsV3, K562 CTCF ChIP-seq tracks) with CTCF motif sites (factorbookMotifPos obtained via the UCSC genome browser, http://genome.ucsc.edu)[42]. Only CTCF ChIP-seq peaks that contained a CTCF binding motif with a score of at least 1.0 within 500 bp of the center of the ChIP-seq peak were considered. The ChIP-seq peaks were subdivided by ChIP-seq binding score (reported in the ENCODE processed data file) and the group of peaks with maximum score (of 1,000) was subdivided into three groups by the motif score, such that four approximately equal-sized groups of CTCF-bound loci were obtained. Enhancer locations were given by the ENCODE HMM chromatin segmentation for K562 cells[43]. The centers of segments annotated as '4/Strong enhancer' and '5/Strong enhancer' were used in our analysis.

H3K4me3, H3K36me3, RING1B and DNase data (external data sets)

H3K4me3 ChIP-seq, H3K36me3 ChIP-seq and DNase data were obtained from ENCODE (sample IDs GSM788087, GSM733714 and GSE90334_ENCFF038VUM, respectively) as processed bigWig files. To calculate OE values for these data sets, whole-genome mappability as determined by the ENCODE project was used (wgEncodeCrgMapabilityAlign36mer). RING1B ChIP-seq data and corresponding input control were obtained from the Gene Expression Omnibus (GSM2393579, GSM2393592) and aligned to the GRCm28 mouse reference index with bowtie2 (v.2.3.3.1) using parameters: seed, 42; very-sensitive-N 1. Genome-wide coverage was obtained with bamCoverage from the DeepTools toolkit (v.3.1.2) using parameters: ignoreDuplicates; minMappingQuality, 10. ChIP-seq domains were called with the callpeak tool of MACS2 (v.2.1.1.20160309) using parameters: keep-dup, 1; seed, 42; broad; broad-cutoff, 0.005.

Comparison DNase and scDam&T-seq Dam stratified by expression

For Fig. 2f, we used an independent microarray expression data set (GSE56465, only the haploid KBM7 samples). Microarray probes that had no gene ID assigned to them were discarded. For gene IDs with multiple assigned probes, the median value was taken. Only gene IDs present in GENCODE v.26 were used in our analysis. We stratified all genes with at least one expression datum (microarray probe) into four expression quantiles. Figure 2f, middle, shows the density of TSSs of genes with the indicated expression quantiles, according to the scDam&T-seq Dam and DNase OE value of the 20-kb bin in which those TSS lie. To determine whether a point in the scDam&T-seq-DNase space was enriched for 20-kb bins contained a TSS of the indicated expression quantile, we used the 'significant fold change' approach, reported previously[44]. Briefly, a normal-approximation using the expected value np, with $p = T/(4N)$, where n is the number of 20-kb bins with given scDam&T-seq Dam and DNase value, T is the total number of 20-kb bins with a TSS and N is the total number of (mappable) 20-kb bins, and a variance

of $n^*p(1-p)$, where n^* is max(25,n) is used to define a confidence interval (we used a critical value of $\alpha = 20\%$) to determine whether the actual number of observed 20-kb bins with a TSS of a gene in the quantile constitutes enrichment or depletion.

Comparison of scDam&T-seq to scNMT-seq

Trancriptomics data from scDam&T-seq (mESC serum) and scNMT-seq were downsampled to 1.5×10^5 raw reads per single cell. Single-cell samples with fewer reads were left out of the transcriptomics comparison. The detected number of genes per cell for both methods is shown in Figure S4d. GpC accessibility data from scNMT-seq were obtained from the processed data of GSE109262.

$\log_2$FC between contact/no-contact groups of samples

$\log_2$FCs between single-cell samples that showed contact and those that showed no contact (see Fig. 3a) were computed as follows. In 100-kb bins across the genome, the $\log_2$FC in gene expression was calculated between samples that have a DamID OE value ≥1 versus samples that have a DamID OE value lower than 1. The expression per bin was determined by the sum over all genes that have their TSS in that bin. Genomic bins that were considered unmappable (fewer than two GATCs per kb) were excluded, as well as bins where either group of samples (high OE, low OE) contained fewer than three samples, or where fewer than 7.5% of all samples showed any expression. Finally, an additional cutoff on samples was used (besides the manuscript-wide cutoffs on DamID event and transcript counts) to exclude samples with anomalous genome-wide DamID patterns (judging by their high-OE bins). The distributions of total fraction of high-OE bins across the genome (bins meeting the mappability and expression cutoffs described above) over all the samples (for Dam-LMNB1 and Dam separately) was modeled as a Gaussian mixture with $k = 1,2,...,5$ Gaussian components with independent means and variances. Using a 25-fold randomized 50% split of samples, we fitted the Gaussian mixture on one half and measured the goodness-of-fit using the other half (using the Akaike information criterion, AIC, which penalizes goodness-of-fit for the number of model parameters). We took the mean of each cross-validation and repeated this process ten times, for each k. We then took the number of Gaussian components k that minimized the mean AIC, which was 2 for both Dam-LMNB1 and untethered Dam. Samples assigned to the Gaussian component with the majority of samples, with a probability of at least 67%, were used further in the analysis of $\log_2$FC in expression.

Rolling mean and standard deviations as a function of CF

In Fig. 3 and related supplementary figures, a rolling mean is shown together with the confidence interval for the mean. To obtain these measurements we calculated the mean and standard deviations of the metric on the y axis for each point on the x axis using a local linear regression approach where data points are weighted according to an exponential decay, that is, $\exp(-d/\tau)$. Here d is the distance between the point at the x axis where the mean is being determined and the data point, and τ is a 'decay factor' (or effective radius). For regressions against CF (Fig. 3b,d and Fig. S5a) a radius of 0.025 (CF units) was used. The shadings indicate

a 95% confidence interval for the means and are determined by 1.96 times the standard deviations, measured using the same exponentially weighted approach as the means.

Variance-to-mean ratios

In our expression data, we observed a variance-to-mean ratio (VMR) that increased with increasing mean expression, indicative of overdispersion (with respect to Poisson-distributed counts). We de-trended the VMR from the (log_2-normalized) mean expression using local linear regression with exponentially decaying weights (see the above paragraph). Fig. S5b shows this 'de-trended' VMR on the *x* axis. Note that, since the log_2FC between high-OE and low-OE samples is largely independent on mean expression (Fig. S5a), raw VMR values show very similar results. The rolling mean and confidence interval in Fig. S5b uses local linear regression with a radius of 0.25 (log_{10}(VMR) units).

Relationship between TAD structure and differential accessibility in 2i versus serum

TADs were obtained from ref.[34] and converted to a 100-kb resolution. Specifically, TAD boundaries were taken to be the midpoint between TADs and rounded to the nearest 100-kb point. The variance in log_2FC serum/2i accessibility (DamID) data in 100-kb bins within each TAD was calculated for all TADs that contained at least three 100-kb bins with at least two mappable GATC motifs per kb. Subsequently, the order of the TADs was randomized per chromosome and the new TAD coordinates were used to calculate a control variance distribution. This process was repeated 50 times. *P* values between the distributions corresponding to the original and randomized TAD structure were calculated using a two-sided Mann–Whitney *U*-test with continuity correction.

Testing for differential gene expression between 2i and serum in mESCs

To determine genes differentially expressed between 2i and serum conditions we employed egdeR[45], using the exactTest function with sample totals determined by scran (computeSumFactors) rather than edgeR's internal sample normalization routines. Panels in Fig. 4a consider genes with a false discovery rate smaller than 5% and an absolute log_2FC greater than 2.0 as either up- or downregulated. For Fig. 4c, genes with absolute log_2FC smaller than 1.3 and unadjusted *P* value greater than 0.5 were considered as 'not differentially expressed'. For Fig. S7c, where all genes (regardless of statistically significant differential expression) are shown, we removed weakly expressed genes by setting a threshold such that 95% of the differentially expressed genes met that threshold.

Detecting chrX allelic biases in DamID and transcription data during differentiation

Allelic coverage in undifferentiated mESCs indicated a CAST/EiJ duplication of the final ~20 Mb of chromosome X. The analyses described below, therefore, include only the first 150 Mb of chromosome X. To detect allelic biases on chromosome X in DamID and transcription data, the log_2(FC) of 129/Sv over CAST/EiJ was calculated for the total number of DamID counts and

transcripts on chrX (with a pseudocount of 1). Subsequently, allelic DamID and transcripts counts on the somatic chromosomes were subsampled such that the combined depth of both alleles corresponded to that of chromosome X. The allelic counts were then used to calculate $\log_2$FC values. One cell in the serum condition showed high CAST/EiJ DamID counts (134) and transcript number (47) while showing no data for 129/Sv (0 counts, 0 transcripts). No such discrepancy was seen for the somatic chromosomes, suggesting that this cell lost its maternal chromosome X. Therefore, the cell was excluded in the calculation of Spearman's correlation coefficient. For differentiation day 3, cells that had a transcriptional chrX allelic bias that exceeded the mean ± 1 s.d. of the somatic chromosome allelic biases were marked as having 129/Sv or CAST/EiJ X inactivation, while the remaining cells were labeled as not showing X inactivation. For the cells in these three categories, the average reads per kilobase per million mapped (RPKM) values on chrX and chr6 were calculated for the two alleles.

Details regarding statistical tests can be found in Supplementary Table 4.

Data Availability

The sequencing data from this study are available from the Gene Expression Omnibus, accession number GSE108639.

Code Availability

All computational code used for this study is available upon request.

References

1. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64 (2013).
2. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134–147 (2015).
3. Flyamer, I. M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 544, 110–114 (2017).
4. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64 (2017).
5. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914 (2015).
6. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).
7. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature 528, 142–146 (2015).
8. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 23, 2126–2135 (2013).
9. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods 11, 817–820 (2014).
10. Farlik, M. et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. 10, 1386–1397 (2015).
11. Mooijman, D. et al. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. Nat. Biotechnol. 34, 852–856 (2016).
12. Zhu, C. et al. Single-cell 5-formylcytosine landscapes of mammalian early embryos and ESCs at single-base resolution. Cell Stem Cell 20, 720–731 (2017).
13. Wu, X. et al. Simultaneous mapping of active DNA demethylation and sister chromatid exchange in single cells. Genes Dev. 31, 511–523 (2017).
14. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol. 33, 1165–1172 (2015).
15. Dey, S. et al. Integrated genome and transcriptome sequencing of the same cell. Nat. Biotechnol. 33, 285–289 (2015).
16. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods 12, 519–522 (2015).
17. Hou, Y. et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 26, 304–319 (2016).
18. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat. Methods 13, 229–232 (2016).
19. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. 9, 781 (2018).
20. Steensel van, B. et al. Chromatin profiling using targeted DNA adenine methyltransferase. Nat. Genet. 27, 304–308 (2001).
21. Vogel, M. J. et al. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nat. Protoc. 2, 1467–1478 (2007).
22. Kind, J. et al. Single-cell dynamics of genome–nuclear lamina interactions. Cell 153, 178–192 (2013).
23. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 17, 77 (2016).
24. Meuleman, W. et al. Constitutive nuclear lamina–genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 23, 270–281 (2013).
25. Monkhorst, K. et al. X inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator. Cell 132, 410–421 (2008).

26. Nishimura, K. et al. An auxin-based degron system for the rapid depletion of proteins in nonplant cells. Nat. Methods 6, 917–922 (2009).

27. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome–nuclear lamina interactions during differentiation. Mol. Cell 38, 603–613 (2010).

28. Aughey, G. N. et al. CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. eLife 7, e32341 (2018).

29. Schones, D. E. et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 132, 887–898 (2008).

30. Valouev, A. et al. Determinants of nucleosome organization in primary human cells. Nature 474, 516–520 (2011).

31. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311–322 (2008).

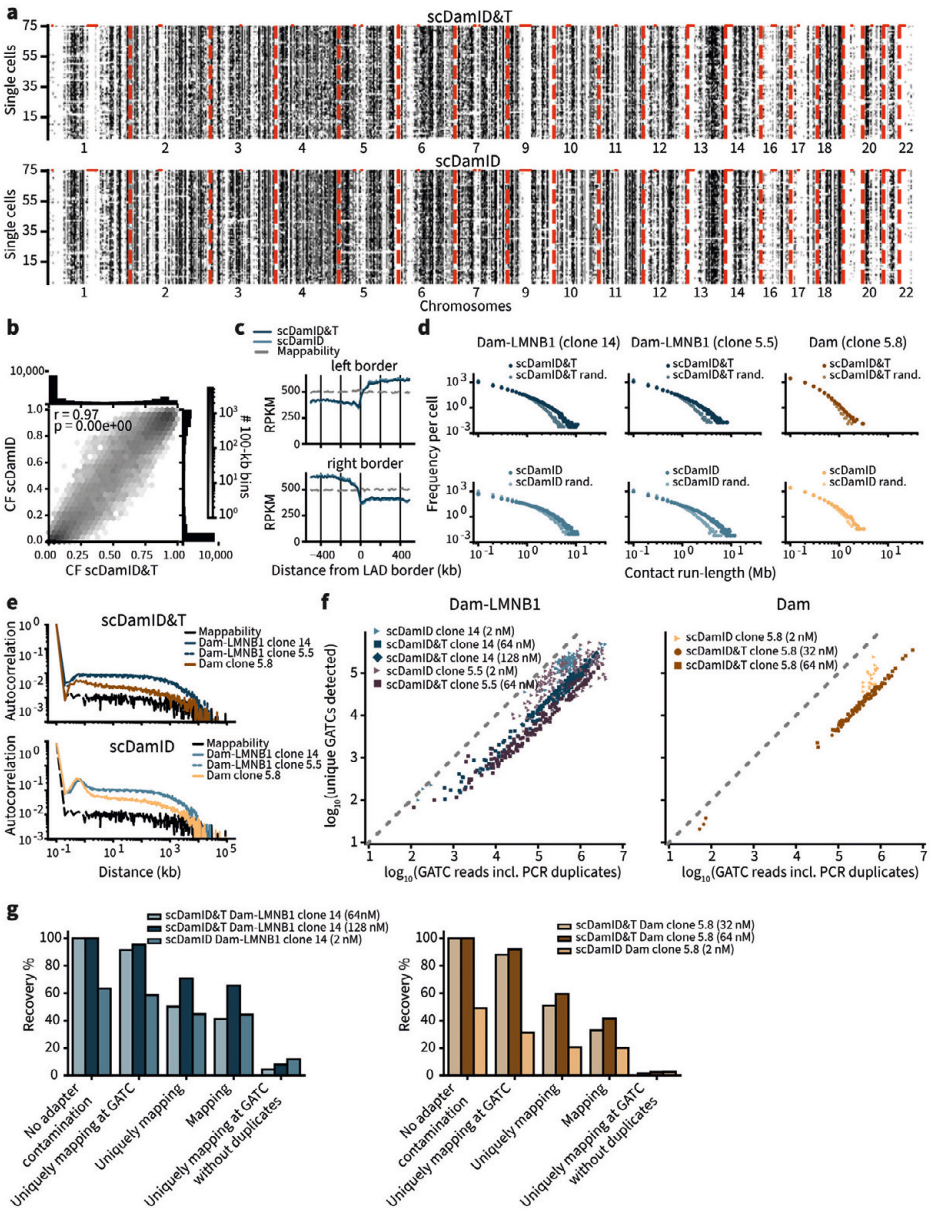
32. Wang, H. et al. Role of histone H2A ubiquitination in polycomb silencing. Nature 431, 873–878 (2004).

33. Zylicz, J. J. et al. The implication of early chromatin changes in X chromosome inactivation. Cell 176, 182–197 (2019).

34. Bonev, B. et al. Multiscale 3D genome rewiring during mouse neural development. Cell 171, 557–572.e524 (2017).

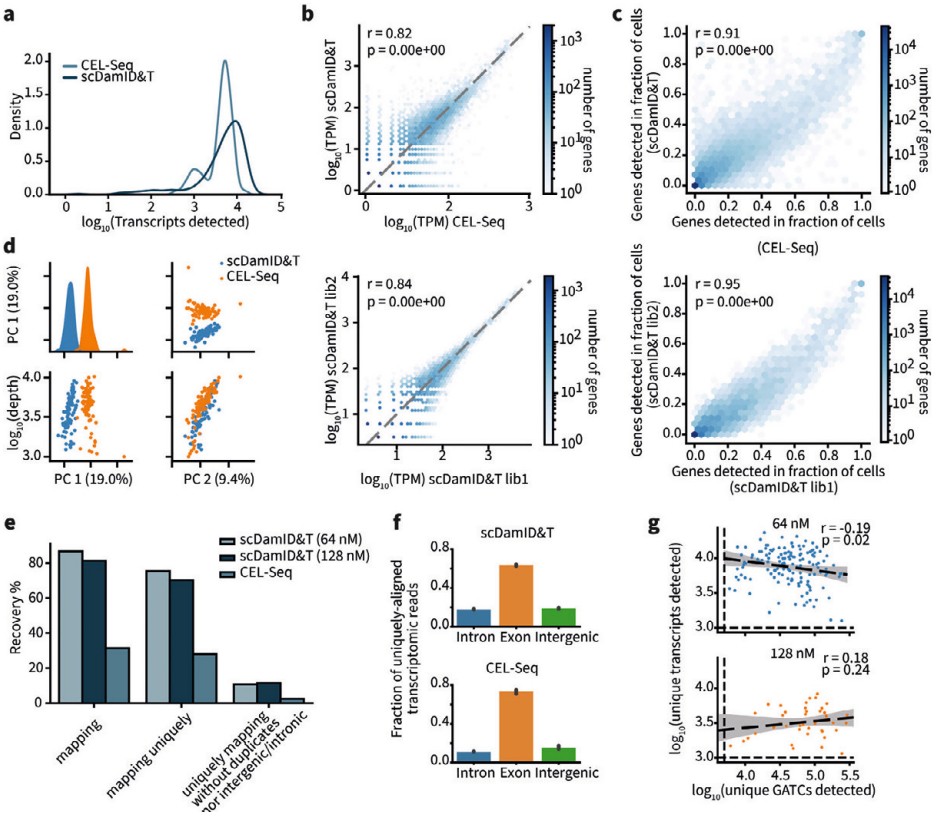
35. Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. 19, 151 (2018).

Supplementary Figures



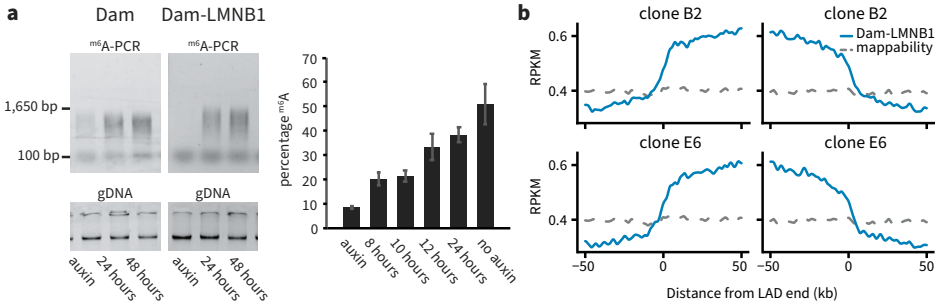
Supplementary Figure 1: Quantitative comparison between scDamID and scDam&T-seq
a, Comparison between the binarized (OE >= 1) single cell (horizontal tracks) maps for scDamID and scDam&T-seq (horizontal tracks; both panels show 75 single-cell samples with highest sample depths). Each row represents a single cell; each column a 100-kb bin along the genome. Unmappable genomic regions are indicated in red along the top of the track**. b**, Comparison of scDamID and scDam&T-seq CFs. CF distributions are depicted in the margins. Pearson's r and p-value are indicated. P-value indicates the

result of a two-sided test (0 indicates a value smaller than 32-bit floating point precision, that is 1.18e-38). **c**, Raw signal (RPKM values) on LAD-boundaries, for both scDamID and scDam&T-seq. LAD positions were defined independently based on HT1080 cells. **d**, Run-length frequencies of uninterrupted 'OE >= 1' runs for two Dam-LMNB1 clones (#14 and #5-5) and one Dam clone (#5–8) for both scDam&T-seq (top) and scDamID (bottom). Run-length frequencies of randomized matrices with preserved marginals (Nature communications 5, 4114, 2014) are shown in light colors. **e**, Pearson autocorrelation of raw signal (y-axis) vs genomic distance (x-axis) of in silico population samples for two Dam-LMNB1 clones and one Dam clone, measured with scDam&T-seq (top) and scDamID (bottom). **f**, Comparison of sample complexities obtained with scDam&T-seq (dark markers) and scDamID (light markers) for Dam-LMNB1 clones and one Dam clone. Unique detected GATCs are depicted on the y-axis vs. GATC-aligning reads (including duplicates) on the x-axis. **g**, Overview of losses during processing of raw sequencing data in scDamID and scDam&T-seq. Bars from left-to-right follow the order of the processing pipeline, where raw reads are first filtered on the correct adapter structure, then aligned to the human genome, where reads not yielding a unique alignment are filtered out, as well as reads not aligning immediately adjacent to GATCs. Finally, duplicate reads are removed, on account of the haploid nature of the KBM7 cell line.



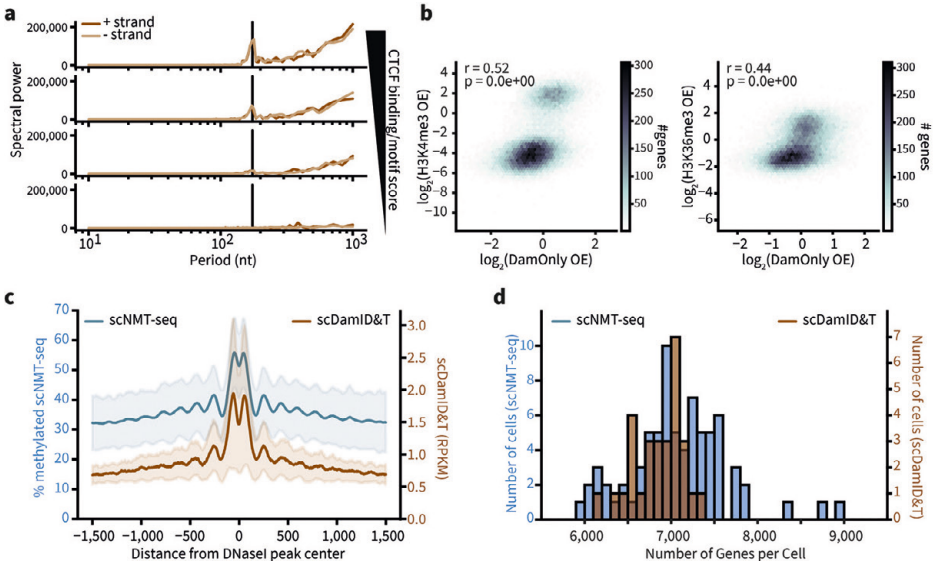
Supplementary Figure 2: Quantitative comparison between CEL-Seq and scDam&T-seq

a, Distributions of the number of unique transcripts detected using CEL-Seq and scDam&T-seq. **b**, Correlation in expression values (TPM, transcripts per million reads) between scDam&T-seq and CEL-Seq (left panel) and two scDam&T-seq libraries processed in parallel (right panel). Pearson's correlation coefficient is indicated. P-value indicate the result of a two-sided test (0 indicates a value smaller than 32-bit floating point precision, ie. 1.18e-38). **c**, Correlation of fraction of cells (passing cutoffs) in which a gene was detected between scDam&T-seq and CEL-Seq (left panel) and two scDam&T-seq libraries processed in parallel (right panel). Pearson's correlation. P-value indicates the result of a two-sided test (0 indicates a value smaller than 32-bit floating point precision, ie. 1.18e-38). **d**, Coefficient of variation (CV) of gene expression (y-axis) vs, mean expression values (x-axis), as measured by scDam&T-seq (4 libraries across 2 KBM7 clones) and CEL-Seq. The dotted line indicates the CV of Poisson-distributed data ($CV = \lambda^{-1/2}$). **e**, Principal component analysis on normalized expression data obtained from CEL-Seq and scDam&T-seq (Dam-LMNB1 clone #14), where the first three principal components are shown, as well as correlation of PC1 with sample depth (number of unique transcripts detected). Numbers in parentheses indicate the fraction of data variance explained by the principal component. **f**, Overview of losses during processing of transcriptomic data obtained with CEL-Seq and scDam&T-seq. Bars from left-to-right follow the order of the processing pipeline, where raw reads are aligned to the human genome, reads that do not yield unique alignments are filtered, as well as reads that do not match exons. Finally, duplicate reads are removed based on the UMIs. **g**, Fraction of transcriptomic reads mapping uniquely to either gene introns, exons or to intergenic loci for scDam&T-seq (top) and CEL-Seq (bottom) in KBM7 samples. Error bars indicate a 95% confidence interval for the mean. Error bars indicate a 95% confidence interval of the mean (calculated by bootstrap procedure). n = 315 for scDam&T-seq, n = 87 for CELseq. **h**, Relation between number of unique transcripts detected (y-axis) and number of unique GATCs detected (x-axis) with scDam&T-seq for two DamID adapter concentrations. Pearson's r and p-values (two-sided test) are indicated. The dotted line indicates a linear regression estimate, the shaded area indicates a 95% confidence interval of regression estimates (determined by bootstrap procedure).



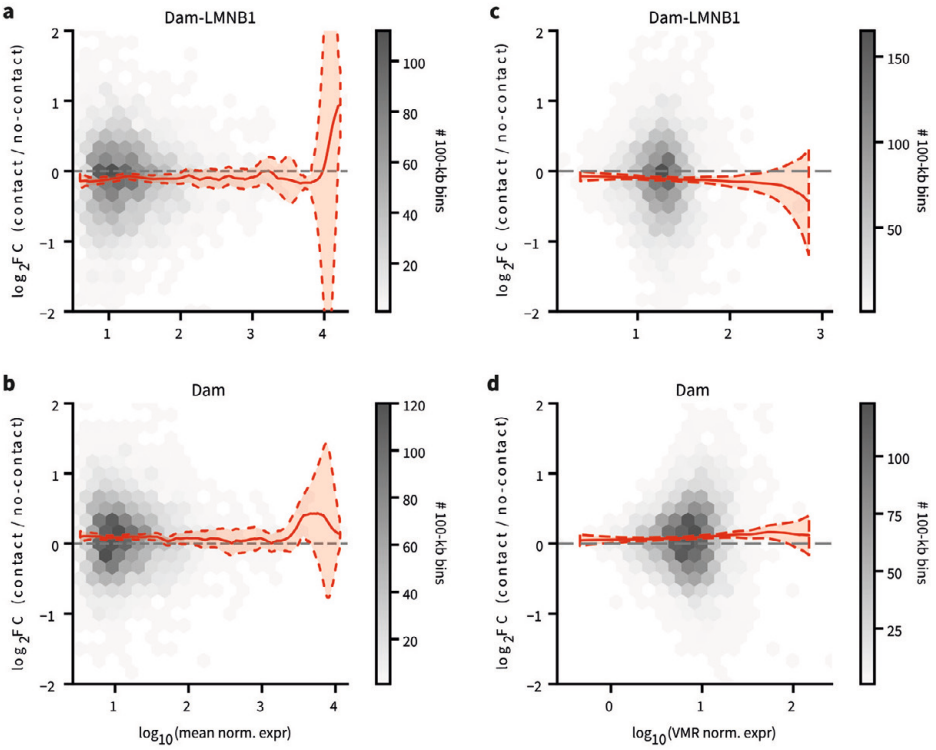
Supplementary Figure 3: scDam&T-seq in hybrid mESCs

a, Auxin-mediated control of AID-Dam (clone #c8) and AID-Dam-LMNB1(clone #b2) cell lines. DamID PCR products of cells 24 and 48 hours after auxin washout (left). Time course and quantitative PCR analysis of auxin induction for a locus within a LAD, 0-, 8-, 10-, 12- and 24 hours after auxin washout (right). Quantification of the m6A levels as described for the DpnII assay. Dot-plot depicts the mean value of n = 2 independent experiments. **b**, mESC in silico population Dam-LMNB1 RPKM values projected on the starts and ends of LAD boundaries defined previously, for two different Dam-LMNB1 clones (#b2 and #e6) and a total of 166 single-cell samples.



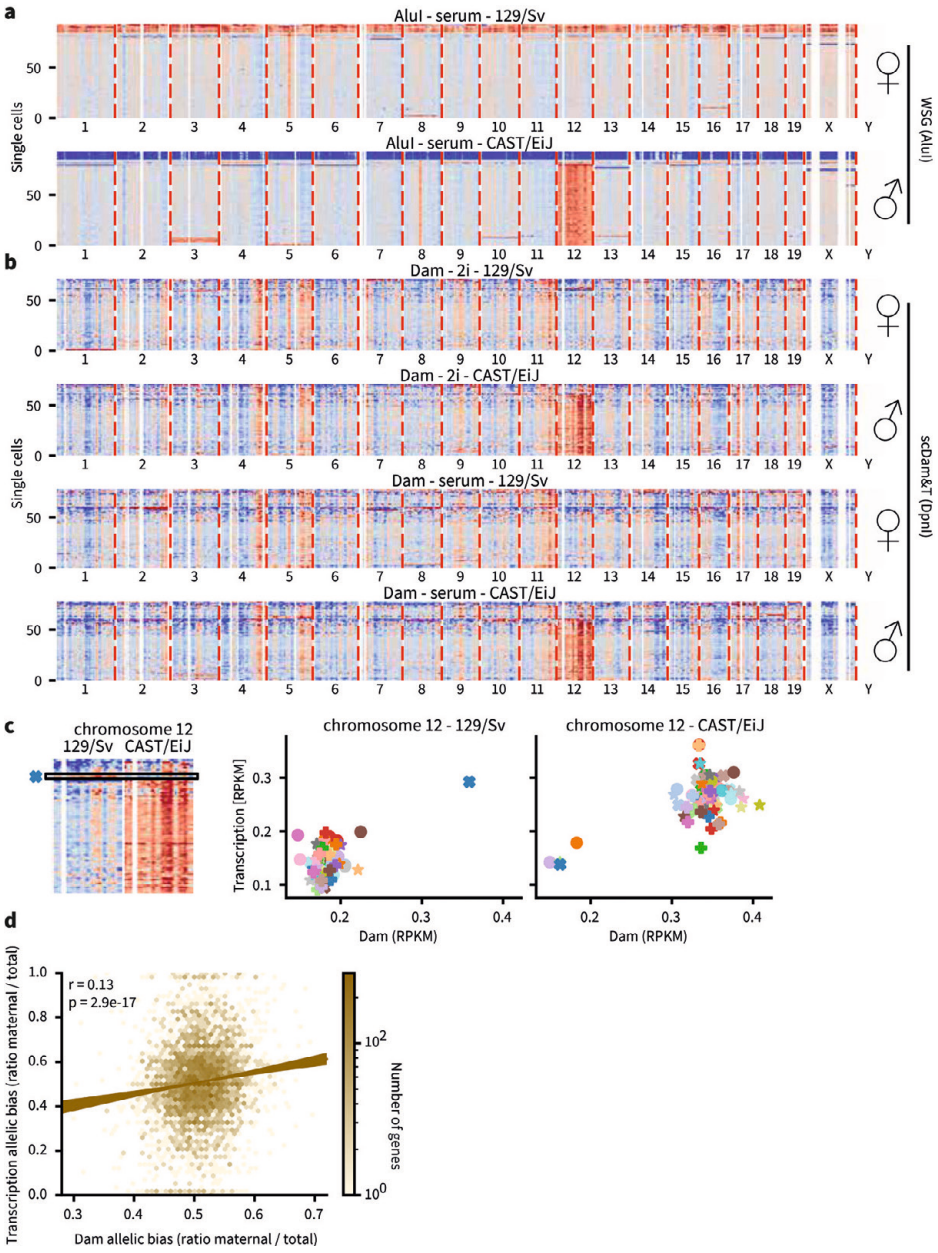
Supplementary Figure 4: Untethered Dam enzyme marks accessible chromatin in single cells

a, 10 bp resolution frequency spectrum of in silico population Dam signal stratified in four regimes of increasing CTCF binding activities (corresponding to Fig. 2d). The black vertical lines indicate 174 bp. **b**, Distribution of 20-kb bins as a function of bulk H3K4me3 (y-axis, left) or bulk H3K36me3 (y-axis, right) ChIP-seq and in silico population Dam data (x-axis). Pearson's r and p-values are indicated. **c**, Percentage of methylation (for scNMT-seq) and RPKM values (for scDam&T-seq Dam signal) at DNaseI hypersensitivity sites, relates to Fig. 1d from Clark et al. (Nature communications 9, 781, 2018). Solid lines indicate the mean across single-cell samples while the shaded areas indicate the standard deviation of signals observed across single-cell samples. n = 95 Dam-only single-cell samples, n = 72 scNMT-seq single-samples. **d**, Number of unique genes observed with scNMT-seq and scDam&T-seq, using single-cell samples down sampled to 150,000 reads. Samples below cutoff were not considered for this analysis.



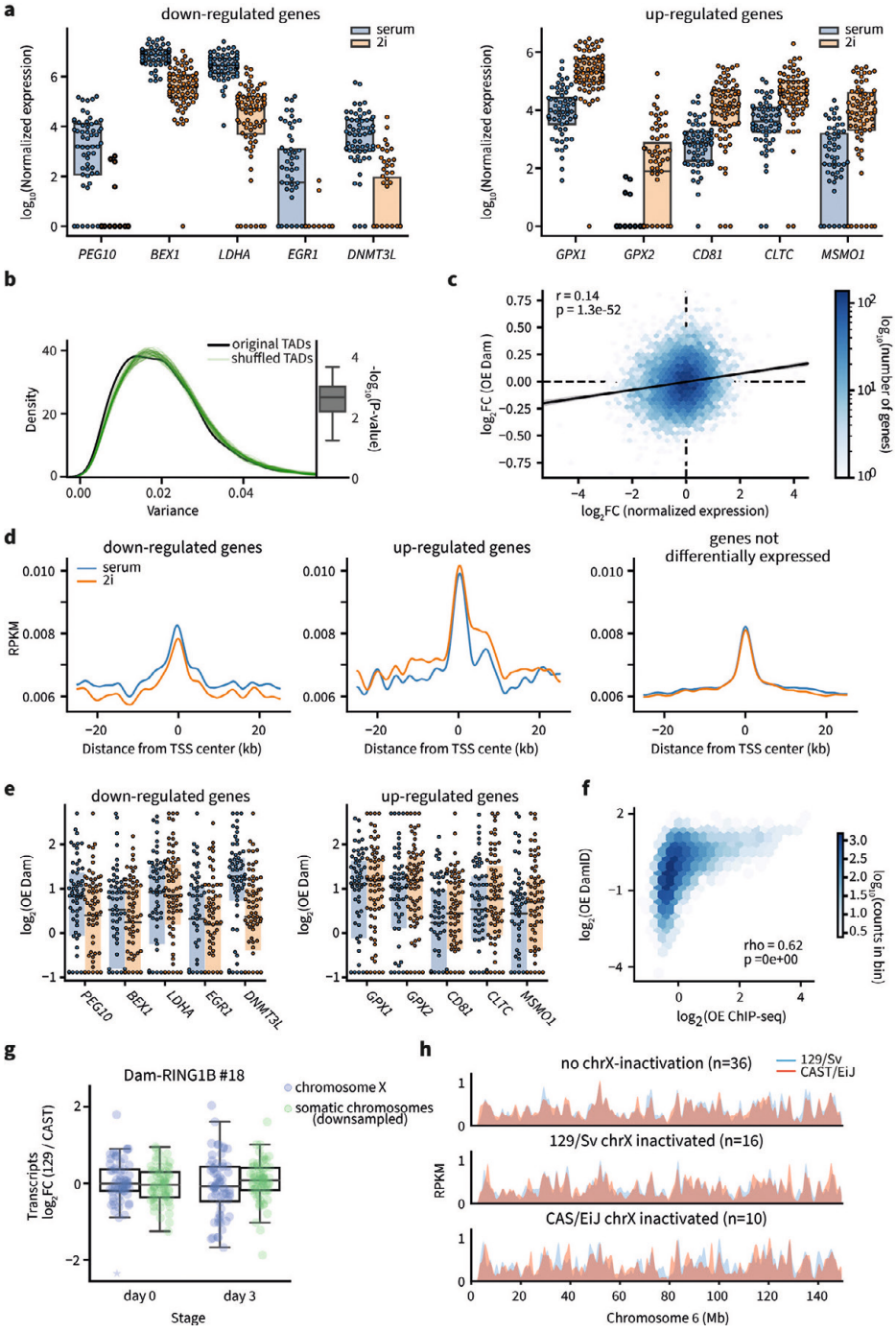
Supplementary Figure 5: Single-cell associations between transcription levels and variance, and Dam or Dam-LMNB1 contacts

a, Relation between expression log_2FC values and mean expression levels for Dam-LMNB1. See Fig. 3a for analysis of expression log_2FC values. Solid line indicates the mean, shaded area indicates 1.96 times the standard deviation around the mean. **b**, As (a), but for Dam. **c**, Relation between expression log_2FC values and expression variance-to-mean ratio for Dam-LMNB1. The variance-to-mean ratios were adjusted by controlling for mean expression levels, since the raw variance-to-mean values were not constant (nor linearly correlating) with mean expression levels (see methods for details). Solid line indicates the mean, shaded area indicates 1.96 times the standard deviation around the mean. **d**, As (c), but for Dam. **e**, Relative enrichment (min-max normalized) of several histone post-translational modifications (PTMs) in genomic regions with different CF values. **f**, The fraction of the genome coverage for (constitutive) cLADs, (facultative) fLADs, ciLADs and fiLADs over a range of CF values.



Supplementary Figure 6: Allelic associations between single-cell transcription and Dam contacts
a, AluI signal obtained from 129/Sv:CAST/EiJ mESCs. Each row represents a single cell; each column a 100-kb bin along the genome. Red colors indicate an enrichment of signal compared to expected (OE, based on AluI-motif density) whereas blue colors indicate a depletion. The top cells with exclusive 129/Sv genomic annotations are likely a contamination of feeder cells (mouse embryonic fibroblasts). **b**, Dam signal from the same clone as in (a), on the maternal (129/Sv) and paternal (CAST/EiJ) alleles, and for 2i and serum, respectively. **c**, Example of a cell (marked with a blue X) which has no duplication of the

paternal chromosome 12 (unlike the majority of the population), but harbors a duplication of the maternal chromosome 12 instead, observable in the Dam signals. This reciprocally corresponds to allele-specific transcription with approximately double the maternal level and half the paternal level, compared to the majority. **d**, Relation between allelic imbalance ('bias') of Dam signals (x-axis) and transcription (y-axis). Note that chromosomes 5, 8 and 12 (and sex chromosomes) seem frequently (partially) duplicated and were excluded from this analysis, as well as single-cell samples for which there was evidence of any CNV on any autosome (see methods). Pearson's correlation is indicated. P-value indicate the result of a two-sided test. The solid line indicates a linear regression estimate, the shaded area indicates a 95% confidence interval of regression estimates (determined by bootstrap procedure).



Supplementary Figure 7: In silico identification of cell identities and corresponding regulatory landscapes with scDam&T-seq

a, Normalized expression values for the top five down-regulated (left) and up-regulated (right) genes in

2i compared to serum. Box plots indicate the 25th and 75th percentile (box), and the median (line). Data points are overlaid as circles. n = 61 and n = 71 for serum and 2i conditions, respectively. **b**, Distribution of variances of differential 2i/serum accessibility between 100-kb bins within TADs (black line) and within randomly reordered TADs (50 iterations, green lines). P-values between variance distributions of the original and randomized TADs were calculated using a two-sided Mann-Whitney U test. Box plots indicate the 25th and 75th percentile (box), median (line) and 1.5 times the inter-quartile range (IQR) past the 25th and 75th percentiles (whiskers). **c**, Relation between $\log_2$ fold-change in Dam signal (RPM) values of in silico population samples (y-axis) and normalized transcription (x-axis) between 2i and serum conditions. Pearson's correlation coefficient and p-value (two-sided test) are indicated. The solid line indicates a linear regression estimate, the shaded area indicates a 95% confidence interval of regression estimates (determined by bootstrap procedure). **d**, Dam OE values of in silico population samples at TSSs of down-regulated (left), upregulated (middle), or unaffected (right) genes in 2i (orange) compared to serum (blue). **e**, Dam OE values measured in single cells at TSSs of the top 5 down-regulated (left) and up-regulated (right) genes in 2i compared to serum. Box plots indicate the 25th and 75th percentile (box), and the median (line). Data points are overlaid as circles. **f**, Genome-wide correlation between RING1B ChIP-seq (bulk, input normalized) and DamID (62 cells, Dam normalized) in 100-kb bins. Spearman's rho and p-value (two-sided test, determined by bootstrap) are indicated. Chromosomes 12, X and Y were excluded from the analysis. **g**, Transcriptional allelic bias on chromosome X (blue) compared to the allelic bias on the somatic chromosomes (green, downsampled to chromosome X coverage). Box plots indicate the 25th and 75th percentile (box), median (line) and 1.5 times the inter-quartile range (IQR) past the 25th and 75th percentiles (whiskers). Data points are overlaid as circles. **h**, Average allelic profiles of RING1B scDam&T-seq signal on chromosome 6 for cells that show a transcriptional bias on chromosome X towards neither allele (top), towards 129/Sv (middle), or towards CAST/EiJ (bottom). **i**, H2AK119Ub ChIP-seq data from GSM3267034 showing allele-specific signals upon X-inactivation (left panel). Comparison of allelic imbalance (ratio maternal over total) between DamID signals measured in cells showing either maternal or paternal X-inactivation, to H2AK119Ub allelic imbalance, per 100-kb bin. Pearson's correlation coefficient and p-value (two-sided test) are indicated in the right panels. The solid line indicates a linear regression estimate, the shaded area indicates a 99% confidence interval of regression estimates (determined by bootstrap procedure).

Supplementary Table 1: **Sample statistics of KBM7 and mESC experiments**

KBM7						Unique DamID events			Unique transcripts			Samples		
Restriction enzyme	Fusion construct	clone id	phase	condition	DamID2 adapter conc. [nM]	Q1	median	Q3	Q1	median	Q3	Total	Passing cutoffs	Passing cutoffs (%)
DpnI	Dam-Lmnb1	clone_14	G1_S	NA	64	9368.5	34608	67011	4416.5	7979.5	11275.5	192	141	73.4
					128	42878	83476	138439	1925	3014.5	4479.5	48	43	89.6
		clone_5.5	G1_S	NA	64	1301.5	4434	24378.5	5005.5	11190	15504	144	56	38.9
	Dam-only	clone_5.8	G1_S	NA	32	6930.5	14404	34133	4033.5	5311	6796.5	48	38	79.2
					64	7659.5	19294.5	39884.5	1111.5	1749.5	2353	48	37	77.1

mESC						Unique DamID events			Unique transcripts			Samples		
Restriction enzyme	Fusion construct	clone id	phase	condition	DamID2 adapter conc. [nM]	Q1	median	Q3	Q1	median	Q3	Total	Passing cutoffs	Passing cutoffs (%)
DpnI	Dam-Lmnb1	clone_B2	G2_M	serum	64	5519.5	24625	77978	2110.5	4473	7545	144	88	61.1
			midS	serum	64	452	4154.5	23030	5778	12823.5	22582.5	48	21	43.8
		clone_E6	G2_M	serum	64	4289	22126	68768	2106	6910.5	10845	144	76	52.8
	Dam-Ring1B	clone_18	un-sorted	Differen-tiation day 3	32	6992.5	16255	33054	10904.5	17230	28475.5	184	145	78.8
				serum	32	3264	9633.5	17423.5	6501.5	9520.5	12049.5	92	62	67.4
	Dam-only	clone_C8	un-sorted	Differen-tiation day 3	32	1835	14395.5	39982	5357.5	9276.5	12856.5	92	62	67.4
			midS	2i	64	15011	61792	131903	2506.5	4167	5598	96	71	74.0
				serum	64	1948	35304.5	93598	7041	9793.5	13968	96	61	63.5

Supplementary Table 2: **8-basepair barcoded double-stranded DamID adapters**

Available in online version: https://doi.org/10.1038/s41587-019-0150-y

Supplementary Table 2: **8-nucleotide barcoded CEL-Seq2 primers**

Available in online version: https://doi.org/10.1038/s41587-019-0150-y

Supplementary Table 4: **Information on statistical tests**

Figure	test	test statistic	p value	n	df
Fig. S1b	Pearson correlation	657.01	0	24168	24166
Fig. S2b (left)	Pearson correlation	201.118	0	20276	20274
Fig. S2c (left)	Pearson correlation	531.72	0	60028	60026
Fig. S2b (right)	Pearson correlation	204.334	0	17132	17130
Fig. S2c (right)	Pearson correlation	721.246	0	60028	60026
Fig. S2h (top; 64nM)	Pearson correlation	-2.29452	0.02326	141	139
Fig. S2h (bottom; 128nM)	Pearson correlation	1.20471	0.23522	43	41
Fig. S4b (left)	Pearson correlation	146.18	0	56621	56619
Fig. S4b (right)	Pearson correlation	116.207	0	57238	57236
Fig. 3c (Dam-LMNB1)	One-sample t-test	-11.5798	1.90E-30	3497	3496
Fig. 3c (Dam)	One-sample t-test	11.2099	1.10E-28	3668	3667
Fig. S6d	Pearson correlation	8.48884	2.89E-17	4051	4049
Fig. 4c (DE up)	One-sample t-test	5.3248	3.40E-07	158	157
Fig. 4c (DE down)	One-sample t-test	-6.52035	1.50E-10	577	576
Fig. 4c (not DE)	One-sample t-test	-1.43223	0.15213	6056	6055
Fig. 4f (left. serum)	Spearman correlation	NA	0.01962543	59	
Fig. 4f (right. day3)	Spearman correlation	NA	4.70E-12	146	144
Fig. S7c	Pearson correlation	15.3439	1.30E-52	11221	11219
Fig. S7f	Pearson correlation	NA	0	22486	22484

CHAPTER 3

Simultaneous quantification of protein–DNA interactions and transcriptomes in single cells with scDam&T-seq

Corina M. Markodimitraki[1,5], Franka J. Rang[1,5], Koos Rooijers[1], Sandra S. de Vries[1], Alex Chialastri[2,3], Kim L. de Luca[1], Silke J. A. Lochs[1], Dylan Mooijman[1,4], Siddharth S. Dey[2,3] * and Jop Kind[1]*

1: Oncode Institute, Hubrecht Institute–KNAW and University Medical Center Utrecht, Utrecht, The Netherlands.
2: Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, CA 93106, USA.
3: Center for Bioengineering, University of California Santa Barbara, Santa Barbara, CA 93106, USA.
4: Present address: Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.
5: These authors contributed equally to this work
*Correspondence: S.S.D. (sdey@ucsb.edu) and J.K. (j.kind@hubrecht.eu).

Nature Protocols, 2020

Abstract

Protein–DNA interactions are essential for establishing cell type–specific chromatin architecture and gene expression. We recently developed scDam&T-seq, a multi-omics method that can simultaneously quantify protein–DNA interactions and the transcriptome in single cells. The method effectively combines two existing methods: DNA adenine methyltransferase identification (DamID) and CEL-Seq2. DamID works through the tethering of a protein of interest (POI) to the *Escherichia coli* DNA adenine methyltransferase (Dam). Upon expression of this fusion protein, DNA in proximity to the POI is methylated by Dam and can be selectively digested and amplified. CEL-Seq2, in contrast, makes use of poly-dT primers to reverse transcribe mRNA, followed by linear amplification through in vitro transcription. scDam&T-seq is the first technique capable of providing a combined readout of protein–DNA contact and transcription from single-cell samples. Once suitable cell lines have been established, the protocol can be completed in 5 d, with a throughput of hundreds to thousands of cells. The processing of raw sequencing data takes an additional 1–2 d. Our method can be used to understand the transcriptional changes a cell undergoes upon the DNA binding of a POI. It can be performed in any laboratory with access to FACS, robotic and high-throughput-sequencing facilities.

Introduction

A myriad of proteins cooperate to establish cell type–specific chromatin architecture and gene expression through their contact with DNA. Such proteins range from post-translationally modified histones to transcription factors, from nuclear lamina (NL) constituents to the transcriptional machinery. Methods to measure protein–DNA interactions (ChIP-seq[1] and DamID[2]) or their effect on chromatin organization (DNase-seq[3] and Hi-C[4]) have provided valuable insight into the link between epigenetic regulation and transcriptional output. However, these methods originally required thousands to millions of cells, and the resulting population-averaged data prohibited the study of diversity and heterogeneity within the sample. Recent technological advances have resulted in single-cell implementations of several methods to study genome architecture[5,6,7,8], chromatin accessibility[9,10,11], DNA modifications[12,13,14,15,16,17] and protein–DNA interactions[18,19,20,21]. The data generated by these single-cell techniques have revealed that there is heterogeneity between the epigenetic states of individual cells. Moreover, single-cell multi-omics methods combining accessibility or DNA methylation readouts with a transcriptional readout have been able to make a direct connection between epigenetic and transcriptional heterogeneity[22,23,24]. However, until recently, a single-cell multi-omics method to study protein–DNA interactions in conjunction with transcription had been lacking.

Here, we describe scDam&T-seq, a method we recently developed to measure protein–DNA interactions and transcripts from the same single cell[25]. scDam&T-seq essentially combines two single-cell methods: single-cell DamID (scDamID)[6] to measure protein–DNA interactions and CEL-Seq2[26,27] to determine the transcriptional state (Fig. 1). The DamID technique relies on the tethering of a POI to the *Escherichia coli* DNA adenine methyltransferase (Dam), an enzyme that exclusively methylates adenines in a GATC motif. Expression of the fusion protein in cells results in methylation of genomic regions where the POI is present. Methylated DNA is then specifically digested and amplified. Through the incorporation of a barcode and a T7 promoter in both the CEL-Seq2 primer and the DamID adapter, DamID and CEL-Seq2 products can be simultaneously amplified, and separation of material is not necessary. Once cell lines expressing the Dam-POI fusion have been established, the processing of single cells and library preparation can be completed in 5 d. Data processing takes 1–2 d. The protocol can be performed in any laboratory with access to FACS, robotic and high-throughput-sequencing facilities.

The dual readout of scDam&T-seq provides the possibility to link the transcriptional and epigenetic states in a way that is impossible when these measurements are performed separately. One possible application of the scDam&T-seq data is to compare the transcriptional output of loci in their bound and unbound state. Another possibility is to use the transcriptional information to assign a cell type or state to each cell and subsequently study how the underlying epigenetics differ between these different populations. We have successfully applied both strategies to study how contact of chromatin with the NL impacts transcription in mouse embryonic stem cells (mESCs) and how Ring1B is progressively enriched on the inactive X chromosome[25].

Overview of the protocol

Before the implementation of the protocol, a cell line (or other biological system) capable of expressing the Dam-POI protein needs to be established ('Experimental design: Selecting stable clones'). Importantly, a cell line expressing the untethered Dam protein needs to be established as well. The expression of the untethered Dam protein will result in the methylation of the accessible regions in the genome and can therefore be used as a control for the non-targeted GATC methylation by the Dam-POI in the experiment. The expression of the Dam-POI is induced before harvesting the cells to allow for the accumulation of GATC methylation. Most of the Dam-POI constructs in our laboratory are controlled by the auxin system, but any other induction system can be used[28]. With the auxin system, the Dam-POI construct is continuously degraded in the presence of auxin in the medium. For the stabilization of the protein, auxin-containing medium needs to be removed from the cells and replaced with fresh medium without auxin. Cells express the Dam-POI a certain amount of time, before they are harvested and prepared for FACS sorting (Fig. 1a,b). This is done by live-staining the cells for DNA content quantification with Hoechst dye before adding propidium iodide to stain and later exclude dead cells. Cells are then sorted as single cells or small populations in 384-well plates containing CEL-Seq2 primers and mineral oil and can be either frozen or used immediately for scDam&T-seq. CEL-Seq2 primers contain a T7 promoter, a P5 Illumina adapter, a unique molecular identifier (UMI), a sample barcode and a poly-dT tail (Fig. 2a, 'Reagents' and 'Experimental design: Design and concentration of DamID adapters and CEL-Seq2 primers'). Once the cell is lysed, the poly-A tail of the mRNA molecule is annealed by the CEL-Seq2 primer and is reverse transcribed (Figs. 1c and 2a). This is followed by second-strand synthesis (Fig. 1d), before proteinase K is added to remove all chromatin-associated proteins from the genomic DNA (Fig. 1e). In the next step, the restriction enzyme DpnI is added to specifically digest methylated GATC sites, generating blunt ends and leaving non-methylated GATC sites intact (Fig. 1f). The cell-specific DamID adapters are then added to the reaction and ligated to the digested genomic DNA (Fig. 1g). The DamID adapters contain a forked sequence that prevents adapter concatemer formation, a T7 promoter, a P5 Illumina adapter, a UMI and a sample barcode (Fig. 2b, 'Reagents'). Once the ligation has taken place, genomic DNA of each cell will contain unique barcodes that do not overlap with the mRNA barcodes, and this allows for the samples to be pooled together and cleaned through bead purification (Fig. 1h). The pooled single-cell material is subsequently linearly amplified by the T7 polymerase in an in vitro transcription reaction, generating multiple copies of each molecule (Fig. 1i). The amplified RNA (aRNA) is then bead-purified, fragmented by salt buffer at a high temperature and purified again before CEL-Seq2 library preparation with minor adjustments[27] (Fig. 1j). The concentration and the quality of the finished libraries are assessed before paired-end sequencing on a NextSeq 500 machine (Fig. 1k). The resulting sequencing reads are derived from the DamID and CEL-Seq2 products.

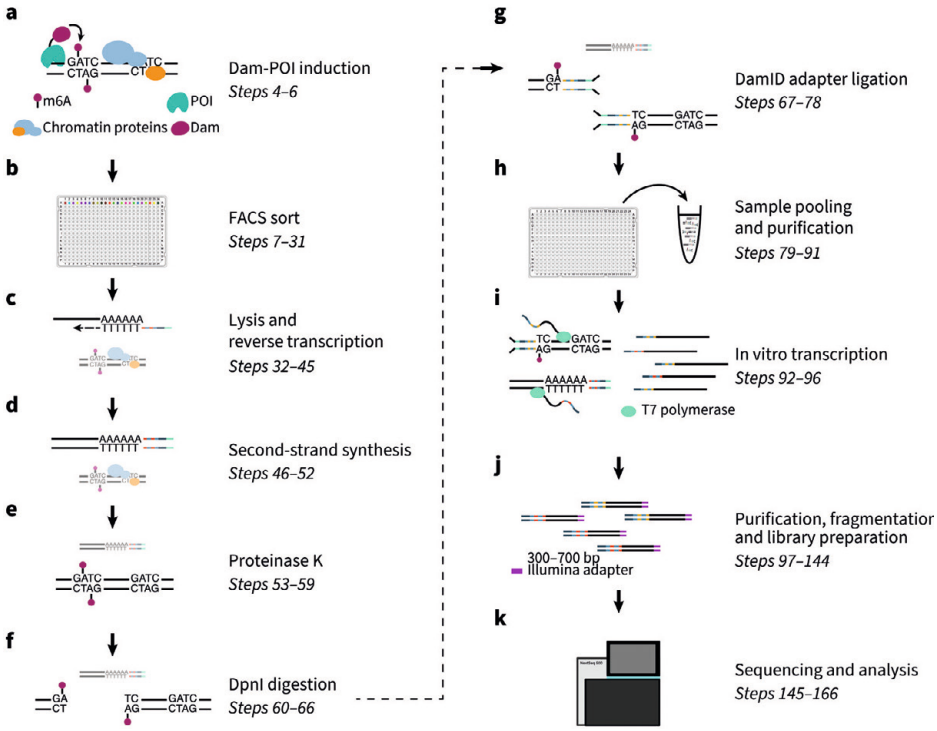


Figure 1: Overview of the method
Panels **a-k** describe the main parts of the protocol (Overview of the Method). The indicated steps refer to the relevant sections of the experimental procedure. In **c-g**, both transcript and gDNA-derived molecules are shown, with the relevant molecule in each step shown in the foreground.

The R1 reads of the DamID product contain a DamID barcode, a UMI and the genomic sequence, while the R2 reads solely contain the genomic sequence on the other side of the fragment. Similarly, the R2 reads of the CEL-Seq2 product contain the sequence of the mRNA molecule, whereas the R1 reads contain the CEL-Seq2 barcode, a UMI and a part of the poly-A tail. It is thus necessary to sequence paired-end to get the genomic sequence of the transcript-derived reads. The data are analyzed following the pipeline provided in this protocol. For a step-by-step detailed overview of the molecule structure throughout the protocol, please refer to the Supplementary Manual.

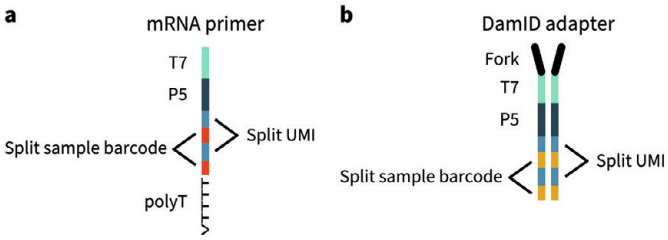


Figure 2: CEL-Seq2 primer and DamID adapter structure
a, CEL-Seq2 primer**. b**, DamID adapter (Overview of the Method and Experimental Design: Design and concentration of DamID adapters and CEL-Seq2 primers).

Extensions of the method

We developed scDam&T-seq as a method to simultaneously detect protein–DNA interactions and poly-adenylated transcripts. However, there are several ways in which this method may be extended. In the original publication, we showed that untethered Dam itself may be used as a chromatin accessibility readout in addition to its role as a control[25]. Consequently, the untethered Dam experiment provides extra insight into the system under study and can be used to link accessibility and transcriptional output. In addition, we applied the scDam&T-seq protocol using the methylation-insensitive restriction enzyme AluI instead of DpnI to digest all the genomic AluI motif occurrences and effectively obtain a reduced-representation whole-genome sequencing[25]. From the resulting data, regions that are enriched or depleted in signal, indicating duplicated and deleted regions, respectively, can be identified across the genome. Such an extension may be of particular interest in systems where frequent large-scale duplications and deletions are known to occur. Moreover, the successful application of AluI opens the door to experimentation with other restriction enzymes (e.g., those sensitive to DNA modifications such as AbaSI for 5-hydroxymethylcytosine or LpnPI and MspJI for 5-methylcytocine[15,29,30)]. Finally, in the original protocol, we employ barcoded poly-dT primers to selectively amplify polyadenylated transcripts. Conceivably, an unbiased sampling of the complete RNA pool could be obtained by substituting the poly-dT primers for random hexamer primers.

Next to these extensions of the protocol, we find that limiting the protocol to the DamID-specific steps generates improved results compared to the original scDamID protocol[6,31]. This adaptation leaves out all mRNA-processing steps and requires only a few minor adaptations of the DamID steps (Supplementary Methods: scDamID2). In addition, we have extended the DamID-only protocol to population samples (Supplementary Methods: DamID2 in bulk). These adaptations are especially helpful when the experimental question requires no transcriptional information or when performing trial experiments (e.g., to select a clone with ideal Dam-POI expression levels). Excluding the transcription-specific steps greatly reduces sample processing time, reagent costs and sequencing cost. The reduction in sequencing costs is twofold, since there is less material to sequence, and the library can be sequenced single-end.

Comparison with other methods

To our knowledge, scDam&T-seq is the first method capable of simultaneously assaying protein–DNA interactions and transcription. However, there are several technologies that probe transcription and/or epigenetic state in single cells. The most obvious comparison is to scDamID, the method from which scDam&T-seq was derived[2,31]. In the original scDamID protocol, methylated DNA is enriched through DpnI digestion followed by the ligation of adapters and PCR amplification. Since a sample-specific barcode is introduced during the PCR via the primer, samples can be pooled only after amplification. In scDam&T-seq, on the other hand, the adapters contain both a sample barcode and a UMI. Therefore, the samples can be pooled before amplification, enabling the multiplexing of high numbers of cells, and amplification biases can be minimized by means of the UMIs. In addition, the exponential PCR amplification is replaced by an in vitro transcription (IVT) reaction, which amplifies the material linearly and should result in fewer amplification biases[32]. It is worth noting that in both scDam&T-seq and scDamID, the obtained depth and resolution are much lower than for classic DamID protocols performed on millions of cells. Whereas bulk DamID may give a resolution of individual GATC fragments (< ~1 kb), single-cell-based methods typically reach a resolution of 50–100 kb.

At the moment, several other single-cell methods are available to probe protein-binding to the DNA[18,19,20,21,33,34]. However, these techniques suffer from low coverage due to precipitation steps[18] or low sample throughput[19]. Recently, a ChIP-based method was published in which the nucleosomes of single cells are barcoded in droplets before being pooled together for immunoprecipitation[33]. Such innovations could potentially improve the efficiency of single-cell ChIP methods. Another promising class of techniques are those based on chromatin immunocleavage–based methods[35], lately adapted for sequencing as CUT&RUN[36]. In these methods, nuclei are isolated, permeabilized and incubated with antibodies against the POI. Subsequently, protein A fused to micrococcal nuclease[37] is added. The protein A–micrococcal nuclease fusion localizes to the bound antibody and, upon calcium addition, cleaves proximal DNA. The resulting DNA fragments are isolated and processed for sequencing. Using this approach, high-quality data can be obtained from low-input (~1,000 cells) samples and even single cells[20,21,34]. With some further development, chromatin immunocleavage–based methods could provide a powerful tool for studying protein–DNA interactions in single cells. However, due to the required isolation of nuclei, these methods currently cannot be combined with transcriptome measurements.

Advantages of the method

One of the main advantages of the scDam&T-seq protocol is that it has been designed to minimize the loss of material and technical biases. Due to the addition of a T7 promoter in the DamID adapters and CEL-Seq2 primers, it is possible to amplify DNA and RNA simultaneously, and there is no need to perform a separation step in which material may be lost. In general, DamID-based methods preserve material well, since no pull-down is required. Meanwhile,

the use of linear amplification through IVT and the presence of UMIs in the adapters minimize the impact of amplification biases. The early ligation of the barcoded adapters allows many samples to be pooled at an early stage of the protocol, which minimizes technical variation between samples and enables the processing of hundreds of samples simultaneously. As a result, it is possible to process thousands of single cells in as little as 5 d. Finally, the technique does not rely on the availability of high-quality antibodies, which often require extensive testing and optimization of buffers.

In addition to these technical features, scDam&T-seq has a number of inherent advantages. The fact that methylation is accumulated over time in vivo means that even transient interactions can be recorded, which could be missed with techniques capturing protein–DNA interactions as a snapshot during sample collection. Since the methylation mark is laid down before sample collection, this also means that biases in the DamID signal due to cell stress are limited. Moreover, the cumulative nature of the DamID mark provides the possibility of tracking the history of protein–DNA contacts during the course of one cell cycle.

Limitations of the method

In most instances, the biggest limitation of scDam&T-seq is the fact that the Dam-POI protein needs to be expressed in the system of interest. This requires the design of a suitable construct, the cloning of the construct into a vector and integration into a cell line or other system of interest. Subsequently, different clones need to be screened to find one showing the best results, as the specificity of the methylation is dependent on the expression level of the Dam-POI. Consequently, scDam&T-seq is applied less readily in samples that are not easily cultured, such as in vivo settings and in clinically derived samples.

Another limitation of scDam&T-seq is its limited applicability to proteins that bind theDNA in very narrow domains, such as transcription factors. Although DamID has been applied successfully to transcription factors in low-input samples (1,000 cells[38]), the resolution obtained for single-cell samples is typically too low to study their localization in a meaningful manner. This problem is exacerbated for proteins that bind accessible chromatin, since regions of accessibility tend to accumulate unspecific methylation. Although untethered Dam can be used to control for the accessibility signature, the contributions of true Dam-POI localization and accessibility will be difficult to disentangle at the low resolution obtained for single cells.

Finally, there is an inherent limitation to the resolution that can be obtained with DamID-based technologies, since signal can only be recorded at GATC motifs. On average, GATC motifs occur at 256-bp intervals, which represents the theoretical upper limit of the resolution. However, in practice, the sparsity of the single-cell data is the true limitation of the resolution in a single-cell sample, with a typical bin size of 50–100 kb.

Experimental design

Necessary controls

During expression of the Dam-POI construct, the fusion protein will sometimes methylate regions of the DNA without proper localization of the POI, resulting in background signal. Such background methylation preferentially occurs at accessible regions of the chromatin. For that reason, a control experiment should always be performed in which untethered Dam is expressed. The extent to which an accessibility signature is present in the data depends strongly on the nature of the POI. Proteins such as LMNB1 are localized to the inaccessible NL, resulting in little to no accessibility signal. On the other hand, proteins that can freely diffuse throughout the nucleoplasm will have stronger accessibility signatures. The accessibility signature obtained from the unthethered Dam experiment can be used for direct normalization of the Dam-POI data or as a negative control. In which way the control data is used depends on the experiment and the specific research question. In general, we find that a normalization works well for population samples or averages of single-cell samples, while treating the untethered Dam as a negative control is more suitable to single-cell data.

In addition to an untethered Dam control, there are a number of technical controls that can be used to optimize the experiments. For instance, leaving a few empty wells in the plate is useful in assaying what the leakage of adapters between samples is. A few wild-type samples not expressing a Dam protein can be included in a library to determine how much of the DamID signal is the result of random genomic DNA (gDNA) breaks that ligate to DamID adapters. Finally, we recommend including up to four wells of small populations of 20 or 100 cells that can be used when performing the protocol on a new Dam-POI experiment, to optimize conditions if the single-cell data do not show anticipated results.

Construct design

The design of the construct depends on many factors that are specific to the POI and the biological system. Factors to consider include whether the Dam protein should be fused to the N or C terminus of the POI, what kind of induction system will be used and whether Dam will be inserted in a targeted manner into the endogenous POI locus or will be randomly integrated.

In our experience, some biological systems are more sensitive to expression levels and duration than others. The DpnI restriction enzyme does not recognize hemi-methylated GATCs and consequently will not digest DNA that has been replicated, because the newly synthesized DNA will contain only unmethylated GATC sites. Therefore, rapidly cycling cells are relatively insensitive to continuous Dam-POI expression, while the genome of senescent or slowly cycling cells may become entirely methylated if Dam-POI expression is not restricted by a proper induction method. In addition, the choice of generating a knock-in of Dam versus an exogenous Dam-POI integration may depend on the expression dynamics of the POI in the biological system.

We recommend that users consider these factors carefully during the design of their constructs and try multiple strategies if necessary. However, not all constructs or Dam-POI fusion proteins work as anticipated. In some cases, the tethering of Dam to the POI affects protein stability, function or localization. Since the generation of a stable expression system can be time consuming, we suggest that the constructs first be tested using DamID on populations of cells. This can be achieved by transducing or transfecting cells with the construct and performing DamID2 in bulk on these heterogeneous samples (Supplementary Methods: DamID2 in bulk).

Selecting stable clones

Once the construct has been introduced into a cell line (or other system[31]) via knock-in or random integration, multiple clones should be screened to select the clone with optimal expression levels. Random integration results in much more diversity than knock-in strategies and may require more clones to be screened to find one that performs well. There are many ways in which the clones can be compared, but we recommend performing at least the following three steps.

First, the expression levels of Dam-POI can be tested by performing a methyl-PCR on the clones[2]. In our experience, clones showing a smear of PCR product on a 1% (wt/vol) agarose gel at 14–24 cycles with 250 ng of input material typically have suitable expression levels for single-cell experiments. A subset of clones with varying expression levels can then be selected for further testing. In addition to expression levels, this experiment is ideal for testing the inducibility of the different clones. Samples can be collected for each clone before induction and at different times after induction to see whether clones are inducible and what their induction dynamics are.

Once several clones have been selected, an initial DamID2 experiment can be performed on bulk samples (Supplementary Methods: DamID2 in bulk). The main purpose of this experiment is to determine whether methylation enrichment is observed at expected regions or in domains of the expected size. Which analyses are most helpful in answering these questions depends entirely on the POI. In the case of very broad domains, such as observed with LMNB1, one suitable measure is the autocorrelation function, which measures the correlation of a signal with a shifted copy of itself. Since the expected domains are broad, the signal should show a higher correlation with itself over longer distances than in the case for untethered Dam. Another way to validate the signal is to evaluate its enrichment over genomic regions where the POI is known to bind. This could be, for example, on the promoters of active or inactive genes or over domains determined for the same mark by ChIP-seq, if data are available.

Although the analyses on bulk samples can give insight on whether or not a clone has successful Dam-POI binding, we caution against picking a clone solely on results obtained from population samples. The ideal Dam-POI expression levels for bulk samples tend to be too low for single-cell samples, where it is important that most cells have sufficient signal. As a final step, we therefore suggest doing an experiment comparing single-cell samples of

the different clones. In the case of a cell line, we find that 50–100 single-cell samples per clone are typically enough to perform the comparison. Since the transcriptional readout is not relevant for clone selection, the samples can be processed following the scDamID2 protocol (Supplementary Methods: scDamID2).

Induction of Dam-POI expression

The method of induction can differ between biological systems. We make use of the auxin-based degron system, which results in specific and fast degradation of the Dam-POI construct[28]. In the absence of auxin, degradation stops and the protein is stabilized. In cultured cells, this is achieved by auxin washout. To obtain a cell line expressing the Dam-POI in an auxin-inducible manner, first a vector containing the TIR1 sequence as well as a selection marker has to be introduced in the genome. This is achieved either by random integration or in a targeted locus-specific manner. Once a clonal cell line has been selected based on TIR1 protein levels, the Dam-POI sequence has to be introduced, C-terminally tagged with the Auxin-Inducible Degron box. The ideal clonal cell line will contain both the TIR1 sequence and the Auxin-Inducible Degron–Dam-POI sequence. Finally, the optimal concentration of auxin needs to be determined, so that the Dam-POI construct is sufficiently degraded without the auxin being inhibitory for cell functions.

The timing of induction (auxin washout) of the Dam-POI construct depends on the activity of the promoter and cell cycle duration. First, time course series are necessary to determine the ideal induction time by checking expression levels of the construct by qPCR or bulk/single cell DamID2. Second, the optimal time window for construct expression needs to be determined. For fast-cycling cells in which the m6A mark is lost upon passage through DNA replication, we recommend induction times that allow for the GATC methylation mark to be re-established after S-phase and sorting of the cells in G2/M. In mESCs, for instance, we recommend induction of Dam-POI constructs for 6–12 h. For slow-cycling or post-mitotic cells, we recommend titration of the induction timings to avoid the accumulation of background methylation.

Sample collection

Depending on the induction time, the research question and the nature of the POI studied, cells can be either collected in the same cell cycle phase or not. For the correct estimation of the cell cycle phase, we live-stain cells with Hoechst dye, which binds to the DNA and allows quantification of the cells' DNA content. Hoechst staining of the DNA has to be optimized per cell type. We recommend testing different concentrations and times of incubation. As an example, we stain 1×10^6 mESCs at a final concentration of 30 µg/ml with a 45-min incubation at 37 °C. Finally, when sorting single cells, we recommend saving the index information of each sorted cell (if this option is available), to be able to link the DamID and transcriptome information of a cell with its cell cycle phase.

Sample pooling

Once the plate has been processed, samples have to be pooled for IVT amplification. For a successful IVT reaction, we recommend pooling a minimum number of 48 single cells. The maximum number of single cells we have successfully pooled for IVT is 384. To avoid library preparation biases, pool as many samples without overlapping barcodes as possible in one library. In our experience, single-cell samples and up to four small population samples ($n \leq 20$) can be pooled in the same library. This has the advantage of eliminating library preparation bias between single cells and populations. On the other hand, by doing so, one loses the ability of choosing the amount of sequencing reads that will be assigned to either the populations or the single cells (sequencing weight). When the experiment contains multiple conditions, it is important to combine these conditions in the same library. In this way, batch effects can be properly assayed and corrected. Therefore, it is recommended to sort samples from different conditions in the same plate and pool samples together without overlapping barcodes, for both DamID and CEL-Seq2. If this is not possible, different conditions can be sorted in different plates and then pooled together, as long as the barcodes do not overlap.

Library preparation and sequencing

The aRNA product of each sample pool is used for the production of one Illumina library, by following the CEL-Seq2 library preparation protocol[27]. This way, each sample pool constitutes a library, barcoded with a unique Illumina index (P7). The indices will be used in downstream bioinformatic analysis, to assign reads to each library. We recommend submitting at least four libraries of 384 single cells or an equivalent thereof, with unique (non-overlapping) Illumina indices per sequencing run, to ensure run complexity and successful cluster formation.

Design and concentration of DamID adapters and CEL-Seq2 primers

The DamID adapters contain a 6-nt fork, a T7 promoter, a P5 Illumina adapter, an 8-nt sample barcode and a 6-nt UMI. The barcode and the UMI sequences are split and alternate in the sequence (Fig. 2b, 'Reagents'). The CEL-Seq2 primers contain a T7 promoter, a P5 Illumina adapter, an 8-nt sample barcode, a 6-nt UMI and a poly-dT tail to anneal to the poly-A tail of the transcripts. Again, the sample barcode and the UMI sequences are split and alternate in the sequence (Fig. 2a, 'Reagents').

The design of split barcode and UMI sequences was chosen primarily with the DamID adapters in mind. For the preparation of the double-stranded DamID adapters, the top and bottom oligo strands of each barcode-specific adapter are combined for annealing. During this procedure, the top and bottom strands with different UMI sequences might not anneal. This would result in the formation of 'bubbles' of non-annealed UMI sequences, which could interfere with the adapter ligation to the gDNA, since the UMI sequence is close to the 3′ end of the adapter. Therefore, the split barcode and UMI design minimizes the formation of such 'bubbles'. The DamID and CEL-Seq2 barcode sequences were designed according to the following four criteria: (i) GC content is between 35% and 65% in the barcode sequences; (ii) there are no homopolymers of ≥3 nt in the barcode sequences and no homopolymers of ≥2 nt in barcode

sequences bordering UMI sequences; (iii) the Hamming distance to all other DamID (CEL-Seq2) barcodes is ≥3; and (iv) the Hamming distance to all CEL-Seq2 (DamID) barcodes is ≥2. The first two criteria ensure that no barcodes of low complexity are generated. The third criterion ensures that each DamID (CEL-Seq2) barcode can be distinguished from all other DamID (CEL-Seq2) barcodes with high confidence. Similarly, the fourth criterion ensures that there is no overlap between DamID and CEL-Seq2 barcodes and that reads originating from the two different techniques can be distinguished. As a result, the DamID and the CEL-Seq2 sample barcodes are non-overlapping and can safely be combined in one experiment. Supplementary Tables 1 and 2 contain 384 CEL-Seq2 primer sequences and 384 DamID adapter sequences, respectively, as they are currently used in our experiments. These primers and adapters can be used to process 384 samples simultaneously within one sequencing library. CEL-Seq2 primers and DamID adapters can be matched in any combination.

The concentration of the DamID adapters can influence the number of obtained DamID reads. Depending on the nature of the POI binding (narrow or broad domains) and the expected intensity of the DamID readout ('Experimental design: Selecting stable clones'), the ideal adapter concentration can vary. To avoid excessive adapter in the samples, we recommend a range of 1.25–100 nM in the reaction. In our experience, increasing concentrations of the DamID adapters does not interfere with the CEL-Seq2 product while increasing DamID output. However, at an equal sequencing depth, an increase in DamID material may result in lower coverage of the CEL-Seq2 material.

Bioinformatic analysis

Raw sequencing data are demultiplexed on CEL-Seq2 and DamID barcodes. After demultiplexing, the reads are aligned and processed according to data type. In general, a high percentage of reads (~95%) can be attributed to a unique CEL-Seq2 or DamID barcode, with most reads being derived from the DamID product (~90%) and a smaller fraction from CEL-Seq2 product (~5%). The final number of unique DamID and CEL-Seq2 reads depends on the complexity of the libraries. For a high-quality library containing 96 single-cell samples, we expect 20–40% of the reads to be unique. The expected output of a scDam&T-seq experiment is discussed in greater detail in the 'Anticipated Results' section at the end of this paper. We have established a pipeline for the processing of the raw sequencing data (Fig. 3). The scripts necessary for the analyses are available on GitHub (https://github.com/KindLab/scDamAndTools), and the main functionalities are described in Table 1.

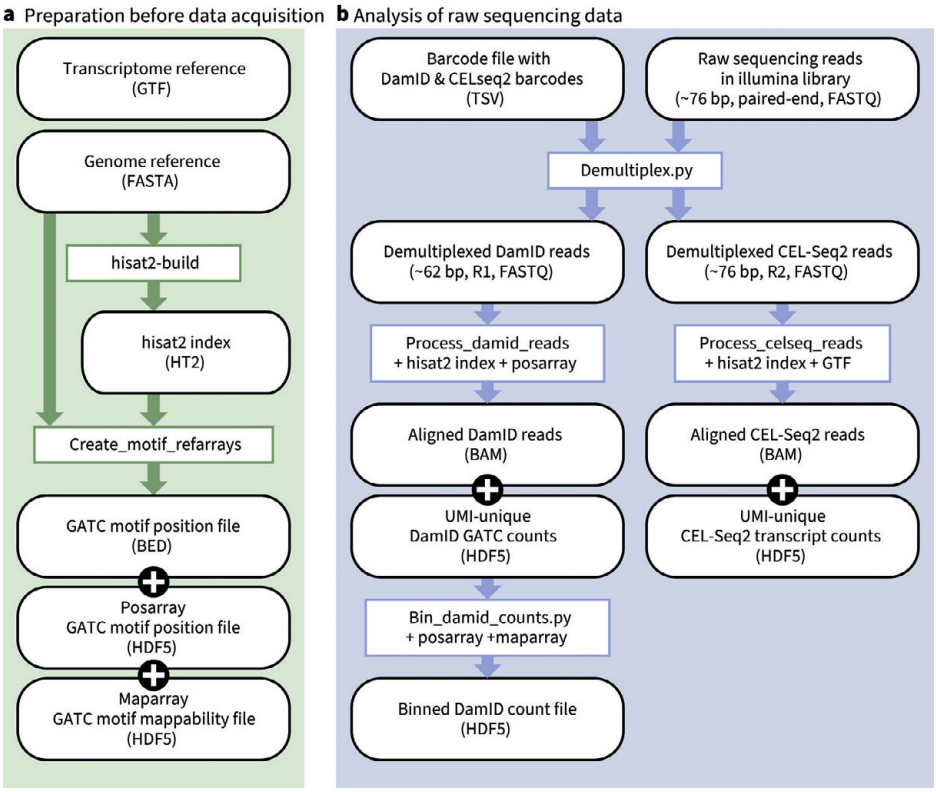


Figure 3: Bioinformatics workflow
a, Preparation of reference files, which only needs to be performed once per reference genome. The genome reference (FASTA) file is used as input to generate the HISAT2 index, as well as the motif arrays. **b**, Processing of raw sequencing data to tables of unique DamID and CEL-Seq2 counts. White, rounded boxes show (intermediate) files; grey, rectangular boxes show programs and necessary reference files. Arrows indicate which files are used as input for subsequent programs.

Table 1: **Description of important functions in the scDamAndT package**

Name	Important parameters	Output
bin_damid_counts.py	**--outfile** Name of the ouput file. **--binsize** Size of the bins to be generated. **--posfile** Location of the GATC position array. **--mapfile** Location of the GATC mappability array. **[input]** HDF5 file containing unique DamID counts.	An HDF5 file containing a dataset for each chromosome. Each dataset is a vector with the observed UMI-unique GATC counts per genomic bin. The bins all have an equal size. The first bin starts at chromosomal position 0.
create_motif_arrays	**-m** The sequence motif targeted by the RE, defaults to "GATC". **-o** The prefix, including path, for the generated files. **-r** The expected read length (R1) obtained from sequencing, excluding UMI and barcode. **-x** HISAT2 index. **[input]** FASTA file of the reference genome.	A BED file containing the strand-specific positions of the GATC motif throughout the genome; a HDF5 file containing the positions of all GATC motifs in the genome (position array); a HDF5 file containing the strand-specific mappability status of all motif occurrences in the genome (mappability array).
demultiplex.py	**-o --outfmt** Output file name format. Should contain a "{name}" and "{readname}" field. **-m --mismatches** Number of allowed mismatches between the sequenced barcode and true barcode sequence. **[bcfile]** File containing barcode names and sequences. **[input]** FASTQ files of the R1 and R2 reads of the library.	A FASTQ file for each DamID and CEL-Seq2 barcode present in the library with reads that had the corresponding barcode. Barcode sequences are removed during demultiplexing.
process_celseq_reads	**-o** The prefix, including path, for the generated files. **-g** GTF file. **-x** HISAT2 index. **[input]** FASTQ file containing CEL-Seq2 reads.	A BAM file containing all alignments to the reference genome; a HDF5 file containing the number of observed UMI-unique counts per gene, ordered by ENSEMBL ID.
process_damid_reads	**-o** The prefix, including path, for the generated files. **-p** Motif position array. **-x** HISAT2 index. **-m** Motif prefix to append to read. **-u** If set, UMI information is taken into account. **[input]** FASTQ file containing CEL-Seq2 reads.	A BAM file containing all alignments to the reference genome; a HDF5 file containing the number of observed UMI-unique DamID for all GATC occurrences in the genome.

Expertise needed to implement the protocol

This protocol requires a FACS sorting facility for the single-cell sorting in 384-well plates. Furthermore, this protocol requires a robot facility or knowledge of robotic operation. In case users do not have access to the robotic systems Mosquito HTS and NanoDrop II used in this protocol ('Equipment'), other robotic machines can be used. These should be able to dispense master mix volumes ranging between 100 and 1,920 nl. To avoid contaminations from previous dispensions, the tubing systems and needles should be able to be flushed or changed. The option for transferring volumes from a 96- or 384-well plate to a 384-well plate should be available as well, for CEL-Seq2 primer and DamID adapter dispension. If the user does not have access to a robotic facility, we recommend upscaling the reactions to the point that the volumes can be handled by hand pipetting. However, we lack experience in this and therefore cannot guarantee fail-proof execution of the protocol. Finally, for the successful sequencing of the libraries, a dedicated sequencing facility is needed.

In addition to these experimental facilities, a high-performance computing system with a Linux operating system is necessary for the analysis of the data. The bioinformatic procedures described in Steps 146–159 give an example of the processing of a single sample. A good understanding of command line usage is necessary to perform the processing of multiple samples in parallel. Finally, knowledge of a programming language such as Python or R is necessary for further downstream analyses and data interpretation.

Materials

Biological materials

- Cell lines. For Figures 3 and 5 of this protocol we used F1 mouse embryonic stem cells with a hybrid genetic background of 129/Sv and Cast/EiJ RRID CVCL_XY63[39]. We have also applied the protocol on the haploid human myeloid leukemia cell line KBM7[6]. Cells were negative for mycoplasma contamination and were not systematically authenticated. ! CAUTION Cell lines should be regularly checked for authenticity and mycoplasma contamination.

Reagents

- Wizard genomic DNA purification kit (Promega cat. no. A1120)
- DNAZap (Invitrogen cat. no. AM9890) **! CAUTION:** *Chronically toxic for the aquatic systems.*
- RNase ZAP (Invitrogen cat. no. AM9780) **! CAUTION:** *Aerosols or vapor can cause irritation to lungs* and mucous membranes. Work in a fume hood.
- Micro-90 concentrated cleaning solution (Sigma-Aldrich cat. no. Z281506)
- Mineral oil (Sigma cat. no. M8410)
- Glasgow's MEM (Gibco cat. no. 21710025)
- MEM non-essential amino acids solution (100 X; Gibco cat. no. 11140035)
- Pen/Strep (10,000 U/ml; Gibco cat. no. 15140122)
- Sodium Pyruvate (100 mM; Gibco cat. no. 11360039)
- GlutaMAX supplement (100 X; Gibco cat. no. 35050038)
- Fetal Bovine Serum (FBS; Sigma cat. no. F7524)
- ESGRO mLIF Medium Supplement (10,000,000 U/ml; Milipore cat. no. ESG1107)
- B-mercaptoethanol (1M; Sigma cat. no. M3148) **! CAUTION:** *Acutely toxic for humans and aquatic systems. Use hand, eye, nose and mouth protection and do not pour in drain.*
- Indole-3-acetic acid sodium salt (IAA, auxin; Sigma-Aldrich cat. no. I5148)
- PBS pH 7.4 (Ambion cat. no. 10010001)
- Trypan Blue solution (Sigma-Aldrich cat. no. T8154) **! CAUTION:** *Can cause cancer. Avoid contact with eyes and skin and do not pour in drain.*
- Hoechst 34580 (Sigma cat. no. 63493)
- Propidium iodide (Sigma cat. no. P4864)
- Nuclease-free water (Invitrogen cat. no. 1097035)
- Magnesium acetate solution (1 M; Sigma-Aldrich cat. no. 63052)
- Potassium acetate solution (5 M; Sigma-Aldric cat. no. 95843)
- Tween 20 (Sigma-Aldrich cat. no. P1379)
- ERCC RNA Spike-In mix 1 (Ambion cat. no. 4456740)
- Igepal (Sigma cat. no. I8896) **! CAUTION:** *Skin, mouth and eye irritant. Use with care and wear gloves and mouth mask in case of insufficient ventilation. Chronically toxic for the aquatic systems. Do not pour in drain.*
- Recombinant ribonuclease inhibitor (Clontech cat. no. 2313A)
- dNTPs set (10mM each; Invitrogen cat. no. 10297018)

- 5x First strand buffer provided with Superscript II package (Thermo Fisher Scientific cat. no. 18064014)
- DTT 0.1 M provided with Superscript II package (Thermo Fisher Scientific cat. no. 18064014) **! CAUTION:** *Work with DTT in a ventilated hood. Wear gloves and a coat, and do not pour it into the drain.*
- RNase OUT (Invitrogen cat. no. 10777019)
- SuperScript II (Thermo Fisher Scientific cat. no. 18064014)
- 5x Second strand buffer (Thermo Fisher Scientific cat. no. 10812014)
- *E. coli* DNA ligase (Invitrogen cat. no. 18052019)
- DNA polymerase I (Thermo Fisher Scientific cat. no. 18010025)
- RibonucleaseH (Thermo Fisher Scientific cat. no. 18021071)
- 10x CutSmart buffer (NEB cat. no. B7204S)
- Proteinase K (Roche cat. no. 3115879001)
- DpnI (NEB cat. no. R0176L)
- Tris pH 7.5 (1 M; Roche cat. no. 10708976001)
- NaCl (5 M; Sigma-Aldrich cat. no. S5150)
- EDTA pH 8 (0.5 M; Invitrogen cat. no. 15575020)
- 10x T4 ligation buffer provided with T4 ligase (Roche cat. no. 1102430292001)
- T4 ligase 5 U/µl (Roche cat. no. 10799009001)
- AMPure XP beads (Beckman cat. no. A63881)
- PEG8000 (Merck cat. no 1546605)
- NaCl (2.5 M; Sigma-Aldrich cat. no. S7653)
- Tris-HCl (1M; Roche cat. no. 10812846001)
- Ethanol absolute (Scharlau cat. no. ET00052500) **! CAUTION:** *Highly flammable; keep away from heat, hot surfaces, sparks and open flames. Avoid contact with eyes and skin, and wear protective gloves.*
- MEGAscript T7 Transcription Kit (Thermo Fisher Scientific cat. no. 1334)
- Trizma acetate powder (Sigma-Aldrich cat. no. 93337)
- Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB cat. no. M0531S)
- Agilent RNA 6000 Pico Kit (Agilent cat. no. 50671513)
- Agilent High Sensitivity DNA Kit (Agilent cat. no. 50674627)
- Qubit 1x dsDNA HS Assay Kit (Invitrogen cat. no. Q33230)
- PhiX control V3 (Illumina cat. no. FC-110-3001)
- NextSeq 500/550 High Output Kit v2.5 (150 Cycles) (Illumina cat. no. 20024907)

Box 1 | Mosquito robot handling

Primer plate preparation – Timing: 20 min for one plate (30 min for 4 max plates)

1. Turn on the computer workstation and the robot and initialize.
2. Use the "Humidify" function to humidify plate deck chamber to 80 % (wt/vol).
1. Remove seal and insert the source CEL-Seq2 primer plate (500 nM) in position 1 with corner A1 facing the upper left corner of the magnetic holder.
3. Remove seals and insert the destination plate(s) containing 5 µl mineral oil in position(s) 2, 3, 4 and 5 with corner A1 facing the upper left corner of the magnetic holder.
4. Copy the source plate by pipetting 100 nl of CEL-Seq2 primer into the destination plates. Change needles after copying a column to avoid contamination across wells.

DamID adapter dispension – Timing: 20 min for one plate (30 min for 4 max plates)

1. Turn on the computer workstation and the robot and initialize.
2. Use the "Humidify" function to humidify plate deck chamber to 80 % (wt/vol).
3. Remove seal and insert the source DamID adapter plate in position 1 with corner A1 facing the upper left corner of the magnetic holder.
4. Insert the destination plate(s) in position(s) 2, 3, 4 and 5 with corner A1 facing the upper left corner of the magnetic holder. Keep on ice whenever not in the robot.
5. Copy the source plate by pipetting 50 nl of DamID adapter into the destination plates. Change needles after copying a column and between plates to avoid contamination across samples.

CEL-Seq2 primer

CRITICAL: Dissolve primers or pre-order them diluted in nuclease-free water to 500 µM and store at -80 °C indefinitely. Order primers as standard desalted.

CRITICAL: Dilute primers in nuclease-free water to a working concentration of 500 nM in a 384-well plate and store at -20 °C indefinitely. Use this as a source plate for the preparation of primer plates (Box 1).

· Example of one CEL-Seq2 primer (5' → 3'): GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNGATGNNNTCATTTTTTTTTTTTTTTTTTTTTTTTTV

CRITICAL: This primer anneals to the poly-A tail of mRNA transcripts with its 3' poly-dT tail. The 3' end contains a "V" base, which is G, C or A. This degenerate base prevents the polymerase from slipping over the poly-A sequence and locks the annealing of the primer immediately upstream of the poly-A tail. The primer contains an 8 nt unique barcode (here: GATGTCAT) that labels a single cell. It also contains a 6 nt unique molecular identifier (UMI) that labels a transcript molecule uniquely (here: NNNNNN, where "N" is G, C, A or T). The barcode is split in 2 x 4 nt and alternates in the primer sequence with the UMI which is split in 2 x 3 nt as follows: NNN-GATG-NNN-TCAT (Experimental Design: Design and concentration of DamID adapters and CEL-Seq2 primers). CEL-Seq2 primer sequences 1-384 can be found in Supplementary Table 1.

DamID adapter

CRITICAL: Dissolve the adapters or pre-order them diluted in nuclease-free water to 500 µM and store at -80 °C indefinitely. Order adapters as standard desalted.

CRITICAL: Dilute DamID adapters in annealing buffer to the desired concentration in separate 384-well plates for top and bottom oligos and store at -20 °C indefinitely. Use these plates for the adapter annealing.

CRITICAL: Anneal the DamID adapters by combining the complementary top and bottom sequences in one 384-well plate by hand-pipetting or with a robotic system at an equal molar ratio and resuspend. Immerse the plate in a container with water heated up to 100 °C, in a way that the wells are in contact with the water, and the seal stays dry. Leave in container till room temperature (20 – 22 °C) is reached. Vortex plate for 5 sec to resuspend adapters and pulse-spin at room temperature for 10 sec. Store adapters at -20 °C indefinitely. Use this plate for adapter dispensing.

- Example of one DamID adapter (5' → 3'):
  Top oligo GGTGATCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATCN-NNTGCANNNTATGGA
  Bottom oligo /5Phos/TCCATANNNTGCANNNGATCGTCGGACTGTAGAACTCTGAAC-CCCTATAGTGAGTCGTATTACCGGGAGCTT
  CRITICAL: The 5' end of the DamID adapter contains a 6 nt non-complementary sequence forming a fork to prevent adapter concatemer formation. The DamID adapter contains an 8 nt unique barcode (here: TGCATATG) that labels a single cell. It contains a 6 nt unique molecular identifier (UMI) that labels a cut GATC site uniquely (here: NNNNNN, where "N" is G, C, A or T). The barcode is split in 2 x 4 nt and alternates in the primer sequence with the UMI which is split in 2 x 3 nt as follows: NNN-GATG-NNN-TCAT (Experimental Design: Design and concentration of DamID adapters and CEL-Seq2 primers). DamID adapter sequences 1-384 can be found in Supplementary Table 2.

Library primers

- Randomhex RT primer: GCCTTGGCACCCGAGAATTCCANNNNNN, where "N" is G, C, A or T. **CRITICAL:** The N bases should be ordered as hand-mixed whenever this option is available.
- RNA PCR Index Primers should follow the Illumina TruSeq Small RNA library prep guidelines (https://www.illumina.com). Example of a RPi index primer (5' → 3'): CAAGCAGAAGAC-GGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCC*A. **CRITICAL:** The "*" indicates a phosphorothioate bond which protects the DNA from endo- and exonuclease activity therefore increasing the stability of the oligo.
- RNA PCR Primer 1 (RP1) primer, according to the Illumina Truseq Small RNA library prep guidelines (https://www.illumina.com) (5' → 3'): AATGATACGGCGACCACCGAGATCTACAC-GTTCAGAGTTCTACAGTCCG*A. **CRITICAL:** The "*" indicates a phosphorothioate bond which protects the DNA from endo- and exonuclease activity therefore increasing the stability of the oligo.

Equipment

- Hardshell 384-well PCR plates (Bio-Rad cat. no. HSP3805)
- Silverseal sealer, aluminium (Greiner Bio cat. no. 676090)
- Cell culture incubator, set at 37 °C and 5 % CO_2 (Panasonic cat. no. MCO-170AIC)

- Mosquito HTS robot (TTP Labtech)
- Vortex (we use the VWR Analog Vortex Mixer VM 3000)
- PCR workstation (WVR cat. no. 7322542)
- Tabletop centrifuge (we use the Eppendorf cat. no. 5810R)
- Burker-Turk hemocytometer (LO-Laboroptik)
- Microscope (we use the Nikon Eclipse TS100)
- Cell strainer caps (Corning cat. no. 352235)
- Falcon Round-bottom polypropylene tubes (Corning cat. no. 352063)
- FACS sorter (we sort on the BD Biosciences BD FACSJazz)
- Low-retention filter tips (we use the Greiner Bio Sapphire low retention pipette tips)
- Nanodrop II robot (BioNex)
- Microcentrifuge MiniStar blueline (Vwr cat. no. 5212320)
- 384-well plate-compatible thermocycler (we use the Eppendorf Mastercycler Pro Thermal Cycler 384 cat. no. 950030030)
- Thermocycler with a 96-well holder (we use the Bio-Rad T100 Thermal Cycler cat. no. 1861096)
- Magnetic rack (we use DynaMag-2 from Life Technologies cat. no. 12321D)
- Heat block (we use the Eppendorf Thermomixer F1.5 cat. no. 5384000012)
- Nanodrop 2000 Spectrophotometer (Thermo Scientific cat. no. ND2000)
- 2100 Bioanalyzer instrument (Agilent cat. no. G2939BA)
- Qubit 3.0 Fluorometer (Life Technologies cat. no. Q33216)

Software

- Unix/Linux operating system (used version: Ubuntu 16.04.6 LTS)
- Bash Shell (used version: bash == 4.2.46(2)-release)
- HISAT2 (https://ccb.jhu.edu/software/hisat2/index.shtml, used version: v2.1.0)[40]
- Python3 (https://www.python.org, used version: v3.6.3)
- Samtools (http://www.htslib.org/, used version: v1.6)[41]
- scDam&T-seq scripts (https://github.com/KindLab/scDamAndTools), functions are explained in Table 1.

Reagent setup

Auxin42 solution (250 mM)
Weigh 492.93 mg of IAA and dissolve in 10 ml sterile water. Filter-sterilize and aliquot. Keep at -20 °C. Protect from light. When moved to 4 °C it can stored up to a week.

mESC complete culture media without β-mercaptoethanol and ESGROmLIF
In 430 ml of Glasgow's MEM, add 50 ml FBS, 5 ml Pen/Strep, 5 ml GlutaMAX 100 X, 5 ml non-essential amino acids 100 X, 5 ml sodium pyruvate 100 mM. The final concentrations in the solution are FBS 10 % (vol/vol), Pen/Strep 100 U/ml, GlutaMAX 1 X, non-essential amino acids 1 X, sodium pyruvate 1mM. Store at 4 °C up to 1 month.

mESC complete culture media with β-mercaptoethanol and ESGROmLIF
In 50 ml of mESC complete culture media without β-mercaptoethanol and ESGROmLIF add 5 µl β-mercaptoethanol 1M and 5 µl ESGROmLIF 10,000,000 U/ml. The final concentrations in the solution are β-mercaptoethanol 0.1 mM and ESGRO mLIF 1,000 U/ml. Store at 4 °C up to 1 week.

Hoechst (1 mg/ml)
Add 5 ml of sterile MiliQ water to 5 mg of Hoechst 34580 to make a dilution of 1 mg/ml. Store at -20 °C for up to 6 months. Protect from light.

Propidium iodide (1 mg/ml)
Dissolve 1 mg of propidium iodide in 1 ml of sterile water and filter-sterilize. Store at 4 °C up to 1 year.

2 % (vol/vol) cleaning solution for Nanodrop II robot
To 49 ml nuclease-free water add 1 ml of Micro-90 concentrated cleaning solution. Store at room temperature indefinitely.

ERCC RNA Spike-In (1:50,000)
Add 9990 µl of nuclease-free water to 10 µl of ERCC RNA Spike-In mix 1 to make a dilution 1:1,000 and make aliquots of 20 µl. To make the 1:50,000 working stock add 98 µl of nuclease-free water to 2 µl of 1:1,000 diluted ERCC RNA Spike-Ins. Store at -20 °C up to the date of expiry indicated and do not freeze-thaw the ERCC RNA Spike-Ins more than 8 cycles.

Proteinase K solution 20 mg/ml
Dissolve 100 mg of Proteinase K in 5 ml nuclease-free water. Store at -20 °C up to the date of expiry indicated by the manufacturer.

Tris pH 7.5 (1 M)
Dissolve 6.05 g of Tris base in 30 ml of nuclease-free water. Adjust the pH to 7.5 with HCl and fill up to 50 ml with nuclease-free water and filter-sterilize. Keep at room temperature.

Annealing buffer 5x
Add 89 ml nuclease-free water to a sterile glass bottle and add 5 ml of Tris 1 M pH 7.5, 5ml of NaCl 5 M and 1 ml EDTA 0.5 M. Final concentrations in the solution are Tris 10 mM, NaCl 50 mM and EDTA 1 mM. Store at room temperature. Prepare new buffer when precipitates start to form.

Igepal 1 % (vol/vol)
Add 49.5 ml nuclease-free water to a 50 ml tube. With a cut tip, take 0.5 ml Igepal and slowly add it to the water. Resuspend till tip is empty. Close tube, invert a few times and leave on rotor until homogenous. Store at room temperature for up to 1 year.

dNTP mix (10 mM)
Add 60 µl of nuclease-free water to a tube. Add 10 µl of dCTP 10 mM, 10 µl of dGTP 10 mM, 10 µl of dATP 10 mM and 10 µl of dTTP 10 mM. Make aliquots of 20 µl and store at -20 °C up to a year. Avoid multiple freeze-thaw cycles.

Bead buffer
Dissolve 20 g of PEG 8000 with 48.75 ml nuclease-free water. Add 1 ml Tris-HCl 1 M, 0.2 ml EDTA 0.5 M, 50 ml NaCl 5 M and 50 µl Tween 20 pH 8.0. Final concentrations in the solution are PEG 8000 20 % (wt/vol), Tris-HCL 10 mM, EDTA 1 mM, NaCl 2.5 M and Tween 20 0.05 % (vol/vol). Store at room temperature for up to 1 year.

Diluted AMPure XP beads
To make a 1:8 bead dilution, add 700 µl of bead buffer to a tube and add 100 µl of AMPure XP beads. Resuspend and vortex till homogenous. Store at 4 °C until indicated expiration date of AMPure XP beads.

Ethanol (80 % vol/vol)
CRITICAL: Measure volumes by pipette and not by "adding up" EtOH to 10 ml.
Add 2 ml of nuclease-free water to 15 ml tube and add 8 ml of ethanol 100 % (vol/vol, absolute). Store at room temperature up to one day. Make fresh for each bead purification.

Tris acetate pH 8.1 (1 M)
Weigh 3.623 g of Trizma acetate powder and dissolve in 10 ml of nuclease-free water. Adjust the pH to 8.1 with a base such as NaOH and fill up to 20 ml with nuclease-free water and filter-sterilize. Keep at room temperature up to 1 year.

Fragmentation buffer
Add 5 ml nuclease-free water in a tube and add 2 ml Potassium acetate 5 M, 3 ml Magnesium acetate 1 M and 10 ml Tris-acetate 1 M. Final concentrations in the solution are Potassium acetate 500 mM , Magnesium acetate 150 mM and Tris-acetate 200 mM. Keep at room temperature up to 1 year.

Stop buffer
Use EDTA pH 8 0.5 M. Keep at room temperature up to 1 year.

Lysis buffer 1 x pH 6-6.5 for scDamID2
In a clean bottle, add 96.66 ml of nuclease-free water, 1 ml of 1 M Tris acetate pH 8.1, 1 ml of 1 M magnesium acetate, 1 ml of 5 M potassium acetate, 0.67 ml of Igepal and 0.67 ml of Tween 20. Final concentrations in the solution are 10 mM Tris acetate, 10 mM magnesium acetate, 50 mM potassium acetate, 0.67% (vol/vol) Igepal and 0.67% (vol/vol) Tween 20. Keep at room temperature up to 1 year.

Procedure

(Day 1) Primer plate preparation – Timing: 30 min for one plate

CRITICAL: It is crucial that the working area is sufficiently clean when working with single-cell material. DNAZap and RNAseZAP treatment is required in steps 1-3 and 21-78. RNAseZAP treatment is sufficient for steps 79-128. Ethanol 80 % (vol/vol) treatment is sufficient for steps 129-145.
CRITICAL: We recommend preparing primer plates, adapter plates and master mixes before single-cell material amplification by IVT in a PCR workstation, in order to avoid contamination and degradation of the single-cell material.

1. Pipet 5 µl of mineral oil in each well of a 384-well plate and seal the plate with an aluminium seal.
2. Thaw the CEL-Seq2 primer source plate at 4 °C and keep on ice.
   CRITICAL STEP: Verify seal is covering all wells of the source plate, vortex plate for 5 sec to resuspend primers and pulse-spin at 4 °C for 10 sec.
3. Copy CEL-Seq2 primers from source plate to the mineral oil plate (Box 1). Seal the plate.
   PAUSE POINT: The primer plates can be kept at -20 °C indefinitely.

Dam-POI induction – Timing: 45 min performed in a cell culture hood

CRITICAL: The timing of induction can differ depending on the Dam-POI construct, cell line and other parameters (Experimental design). For the data shown in the 'Anticipated Results', mESCs expressing Dam-LmnB1 were cultured for 2 days on feeder cells till 80 % colony confluency and the Dam-LmnB1 construct was expressed for 6 hr before collection.

4. Pre-warm cell culture medium and PBS at 37 °C for 30 min.
5. Remove indole-3-acetic acid[42]-containing medium from cells and wash 3 x with warm PBS.
   CRITICAL STEP: Cells that are easily detached might need to be washed with medium instead of PBS.
6. Add cell culture medium without IAA and place cells in 37 °C incubator until harvest.

(Day 2) Prepare cells for FACS sorting by Hoechst staining – Timing: 1 hr 30 min

7. Harvest cells χ hr after induction and prepare a single-cell suspension. For the mESCs presented in the 'Anticipated Results', induction was 6 hr prior to harvest.
8. Clean hemocytometer with a tissue sprayed with 80 % (vol/vol) ethanol and secure the coverslip.
9. Take 100 µl of cell suspension and put in a tube.
10. Add 400 µl of Trypan-blue to the tube and mix by pipetting up and down 3-4 times.
11. Pipet carefully 100 µl of the Trypan blue-cell suspension mix into the cavity between the hemocytometer and the coverslip. Capillary forces will draw the liquid inside.
12. Place the hemocytometer under the microscope set at a 10 x objective and focus the microscope on the grid lines of the hemocytometer.
13. Count the live cells (unstained) at the upper left square containing 16 smaller squares. Once done, move to the other 16-corner set square until all the 4 squares are counted.

14. To calculate the number of viable cells/ml of cell suspension, divide the number on the tally counter by 4 to obtain the average number per corner square. Multiply the average by 10,000 and then by 5 to correct for the 1:5 dilution of the cell suspension with the Trypan Blue.
15. Dilute the cell suspension to 1 x 10^6 cells/ml. Pipet a minimum of 600 µl of cell suspension in a 15 ml falcon tube.
16. Thaw the Hoechst solution on ice and avoid light exposure.
17. Add Hoechst to the cell suspension to stain the DNA. For the mESCs presented in the 'Anticipated Results', induced cells were stained with 30 µg/ml Hoechst.
    CRITICAL STEP: The final concentration of the Hoechst needs to be optimized for the cell type used (Experimental design: Sample collection).
18. Incubate the cells at 37 °C for 45 min in a cell culture incubator, avoiding light exposure.
19. Pass the cells through a falcon cell-strainer cap and in a polypropylene round-bottom tube to exclude cell clumps by gently pipetting the cells on the cap and leting them pass through the filter without force from the pipet tip.
20. Keep cells on ice till FACS sort. Do not keep cells on ice for longer than 3 hr.

FACS sorting – Timing: 1 hr for 3 plates

21. Thaw the primer plate from Step 3 on ice.
22. Centrifuge plate at 2,000 g for 1 min at 4 °C for primer droplets to fall at the bottom of each well.
23. Transport plate and cells on ice to the FACS sorter.
24. Add propidium iodide to a final concentration of 1 µg/ml to the cell suspension for live/dead cell gating.
25. Remove the plate seal and load both plate and tube with cell suspension onto machine.
26. Exclude debris based on side scatter-forward scatter. See Supplementary Fig. 1 for gating strategy.
27. Exclude dead cells based on propidium iodide intensity.
28. Create a histogram plot for event counts and Hoechst intensity, to visualize the DNA content. Cells in G2/M should show Hoechst intensity that is twice that of cells in G1.
    ? TROUBLESHOOTING
29. On the DNA histogram create a gate for preferred cell cycle phase.
30. Sort cells that are alived and in preferred cell cycle phase either as single cells or as small populations of 10, 20 or 100 cells per well (Experimental Design: Sample Collection).
31. Seal the plate and spin at 2,000 g for 1 min at 4 °C and store at -80 °C immediately to keep the RNA intact.
    PAUSE POINT: Sorted plates can be kept at -80 °C for several months.

Box 2 | Nanodrop II robot handling

Master mix dispension – Timing: 20 min for one plate (30 min for 4 max plates)

1. Perform a "Daily clean" program with 2 % (vol/vol) cleaning solution.
2. Prepare a "mock" 8-well PCR strip containing nuclease-free water at the same volume as the master mix.
3. Insert the mock PCR strip and a sealed mock 384-well plate in corresponding positions.
4. Dispense the desired volume of the particular step in the protocol, to check if robot aspirates the correct volume and that the seal at positions of all wells contains the desired amount of water.
5. If water check confirms that robot dispenses correctly, repeat steps 3-4 with the actual master mix and destination plate. Remove the seal of the destination plate before dispension.
6. After the Proteinase K dispension (steps 53-56) perform a "Daily clean" program with 2 % (vol/vol) cleaning solution to remove excess Proteinase K from the tubing systems to avoid contamination in next dispensing steps.

Lysis – Timing: 45 min

32. Thaw the RNA ERCC spike-in dilution and the dNTPs on ice. Keep the recombinant ribonuclease inhibitor on an ice block at all times. Prepare the lysis mix according to the table below. Keep the mix on ice at all times. Incubation at 65 °C (at Step 37) with Igepal will permeabilize/lyse the cells.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	41	38.8	
ERCC RNA Spike-in (1:50,000)	20	18.1	1:250,000
Igepal 1 % (vol/vol)	15	14.2	0.15 % (vol/vol)
Recombinant ribonuclease inhibitor (40 U/**µl)**	4	3.7	1.6 U/µl
dNTP mix (10 mM)	20	18.1	2 mM each
Total volume	100	92.9	

33. In an 8-well PCR strip, aliquot 11.1 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
    CRITICAL STEP: It is important to perform steps 34-52 as fast as possible. Keep the plate cold at all times (unless dispension is taking place) to avoid RNA degradation.
34. Remove sorted plate (Step 31) from -80 °C, remove seal and dispense the lysis mix immediately.
35. Dispense 100 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 200 nl.
36. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C
37. Heat the plate in a thermocycler heated at 65 °C for 5 min with the lid at 100 °C and then place on ice immediately for 1-2 min to cool down.
38. Centrifuge at 2,000 g for 1 min at 4 °C.

Reverse transcription – Timing: 1 hr 45 min

39. Thaw the 5 x First strand buffer and DTT on ice. Keep the RNase OUT and Superscript II on an ice block at all times.
40. Prepare the reverse transcription mix according to the table below. Keep the mix on ice at all times. Reverse transcription will generate a DNA molecule complementary to the transcript sequence (cDNA) by using the barcoded CEL-Seq2 primers added in step 3.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	10	8.6	
5x First strand buffer	70	60.5	2.3 x
DTT (0.1 M)	35	30.2	0.023 M
RNaseOUT (40 U/µl)	17.5	15.1	4.6 U/µl
Superscript II (200 U/µl)	17.5	15.1	23.33 U/µl
Total volume	150	129.5	

41. In an 8-well PCR strip, aliquot 15.5 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
42. Dispense 150 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 350 nl.
43. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C.
44. Put the plate in a thermocycler at 42 °C for 1 hr, 4 °C for 5 min and 70 °C for 10 min, with the lid at 100 °C and then place on ice for 1-2 min to cool down.
45. Centrifuge at 2,000 g for 1 min at 4 °C.

Second strand synthesis – Timing: 2 h 30 min

46. Thaw the 5 x Second strand buffer and dNTPs on ice. Keep the *E. coli* ligase, the DNA polymerase and the Ribonuclase H on an ice block at all times.
47. Prepare the second strand mix according to the table below. Keep the mix on ice at all times. Second strand synthesis will generate a second strand of DNA complementary to the cDNA molecule generated during reverse transcription. This is done by using the transcript fragments cleaved by Ribonuclease H, as primers.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	1,347.5	569.3	
5x Second strand buffer	437.5	184.8	1.1 x
dNTP mix (10 mM)	43.7	18.5	0.22 mM each
E. coli ligase (10 U/µl)	15.7	6.6	0.08 U/µl
DNA polymerase I (10 U/µl)	61.2	25.9	3.18 U/µl
Ribonuclease H (2 U/µl)	15.7	6.6	0.016 U/µl
Total volume	1,920.8	811.8	

48. In an 8-well PCR strip, aliquot 101 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
49. Dispense 1,920 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 2270 nl.
50. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C.
51. Put the plate in a thermocycler at 16 °C for 2 hr with the lid at 100 °C and let the thermocycler go to 4 °C at the end of the program.
52. Centrifuge at 2,000 g for 1 min at 4 °C.

Proteinase K treatment – Timing: 11 hr overnight reaction

53. Thaw the Proteinase K and 10 x CutSmart buffer at room temperature and put on ice.
54. Prepare the proteinase K mix according to the table below. Keep the mix on ice at all times. Proteinase K will cleave all proteins in the cell, including DNA-bound proteins and nucleases.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	84.5	41.5	
10x CutSmart buffer	277	136.3	5.5 x
Proteinase K (20 mg/ml)	138.5	68.1	5.54 mg/ml
Total volume	500	245.9	

55. In an 8-well PCR strip, aliquot 30.3 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
56. Dispense 500 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 2770 nl.
    CRITICAL STEP: It is important to perform a "Daily clean" on the Nanodrop II robot after dispensing Proteinase K, to remove traces of the proteinase and avoid subsequent reaction contaminations (Box 2).
57. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C.
58. Put the plate in a thermocycler at 50 °C for 10 hr, 80 °C for 20 min with the lid at 100 °C and let the machine go to 4 °C at the end of the program.
59. Centrifuge at 2,000 g for 1 min at 4 °C.

(Day 3) DpnI digestion – Timing: 7 hr

60. Thaw the 10x CutSmart buffer at room temperature and put on ice. Keep the DpnI enzyme on an ice block at all times.
61. Prepare the DpnI mix according to the table below. Keep the mix on ice at all times. DpnI will digest all methylated GATC sites in the genome, leaving blunt free ends.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	177	108.5	
10x CutSmart buffer	23	14.1	1 x
DpnI (20 U/**µl**)	30	18.4	0.4 U/µl
Total volume	230	141	

62. In an 8-well PCR strip, aliquot 17.4 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
63. Dispense 230 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 3000 nl.
64. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C
65. Put the plate in a thermocycler at 37 °C for 6 hr, then at 80 °C for 20 min with the lid at 100 °C and let the thermocycler go to 4 °C at the end of the program.
66. Centrifuge at 2,000 g for 1 min at 4 °C.

Adapter dispension – Timing: 45 min

67. Thaw the adapter plate at 4 °C during the DpnI digestion.
68. Prepare the Mosquito robot (Box 1).
69. Dispense 50 nl of DamID adapter per well. The cumulative reaction volume is 3050 nl.
    CRITICAL STEP: The final concentration of DamID adapter needs to be optimized for the expression of Dam-POI construct of choice. We recommend the final concentration falls within the 1.25 – 100 nM range.
70. Remove the plate and the adapter plate from the robot and seal.
71. Centrifuge at 2,000 g for 1 min at 4 °C.

Adapter ligation – Timing: 12 h 45 min overnight reaction

72. Thaw the 10x Ligase buffer on ice. Keep the T4 ligase on an ice block at all times.
73. Prepare the ligation mix according to the table below. Keep the mix on ice at all times.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
10x Ligase buffer	350	176.2	7.7 x
T4 Ligase (5 U/**µl**)	100	50.3	0.12 U/µl
Total volume	450	226.5	

74. In an 8-well PCR strip, aliquot 28 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
75. Dispense 450 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 3500 nl.
76. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C.
77. Put the plate in a thermocycler at 16 °C for 12 hr, then 65 °C for 10 min with the lid at 100 °C. Let the thermocycler go to 4 °C at the end of the program.
78. Centrifuge at 2,000 g for 1 min at 4 °C.
    PAUSE POINT: The processed plate can be kept at -20 °C up to a month.

(Day 4) Pool cells – Timing: 1 hr

79. If frozen, thaw the plate from step 78 on ice.
    CRITICAL STEP: Depending on the number of plates, pooling can take longer. We indicate 1 hr approximately for 1 plate.
    CRITICAL STEP: Depending on the number of adapters used and the distribution of barcodes in the plate, pooling of non-overlapping barcoded material can be done manually or by inversion of the plate onto a collection container such as a clean lid of a tip box and centrifugation at 200 g for 1 min.
80. Pool cell lysates in a way that the barcodes do not overlap in 5 ml tubes.
    CRITICAL STEP: The higher the number of pooled wells, the higher the pool volume. The final reaction volume in each well is 3.5 µl. As an example, when pooling 384 wells, the expected cumulative volume is 384 x 3.5 µl = 1,344 µl. In this case, we recommend splitting the volume over three clean tubes, resulting in a volume of 448 µl per tube.
81. Separate oil from aqueous phase by pulse spin and collect the aqueous solution at the bottom of the tube containing the barcoded material. Keep on ice.
82. Repeat step 81 and transfer aqueous phase to new tube to remove mineral oil completely.

Purification of barcoded material – Timing: 1 hr

CRITICAL: Depending on the number of pools, bead purifications can be a bottleneck. We therefore do not recommend cleaning more than 8 reactions simultaneously.
CRITICAL: In the following steps bead cleanups are needed to remove byproducts of the previous reactions and for size selection. For pool volumes higher than 30 µl, we recommend using AMPure XP beads that have been diluted with bead binding buffer (Reagent setup). By doing so, the volume of AMPure XP beads is reduced while the size selection is not affected. This enables proper elution of the AMPure XP beads in the small volume of 6 µl in step 91. Taking the example of 448 µl pool volume mentioned above, we recommend using a bead dilution of 1:8, meaning 1 part AMPure XP beads and 7 parts bead binding buffer (Reagent setup). For smaller pool volumes we recommend smaller bead dilutions. Keep an approximate of 30 µl of AMPure XP beads in the final mix, in order to enable the water elution of step 91.

83. Equilibrate diluted AMPure XP beads to room temperature for 30 min. Vortex till bead-buffer mix is homogenous.
84. Add 0.8 volume diluted AMPure XP beads to 1 volume of pool (Step 82). Allow material to bind to the AMPure XP beads for 10 min at room temperature. Steps 84-95 need to be carried out at room temperature.
85. Put the tube on a magnetic rack and allow AMPure XP beads to accumulate. Keep samples on magnetic rack until step 90.
86. Remove the aqueous phase carefully without disturbing AMPure XP beads.
87. Add 500 µl of fresh 80% (vol/vol) ethanol and leave for 30 sec.
88. Remove the ethanol carefully without disturbing AMPure XP beads.
89. Repeat steps 87-88. Pulse-spin tube and place in magnetic rack to remove excess ethanol.
    ? TROUBLESHOOTING

90. Let AMPure XP beads to air-dry for 5 min or until they appear “matte”.
    CRITICAL STEP: Do not let AMPure XP beads overdry. Elute in water before cracks start appearing in the bead pellet.
91. Add 6 µl of nuclease-free water to the AMPure XP beads to elute the material and resuspend until beads and water form a homogenous mix. Place tube on ice.

Amplification by in vitro transcription – Timing: 14 h 15 min

92. Thaw the Megascript T7 10x buffer at room temperature. Vortex thoroughly to dissolve precipitates and keep at room temperature.
93. Thaw the Megascript T7 NTPs on ice and keep on ice. Keep the enzyme mix on an ice block at all times.
94. Prepare the Megascript T7 mix as indicated in the table below

Reagent	Amount (µl)	Final concentration in mix
10x T7 buffer	1.5	9.3 x
ATP (75 mM)	1.5	7 mM
UTP (75 mM)	1.5	7 mM
GTP (75 mM)	1.5	7 mM
CTP (75 mM)	1.5	7 mM
T7 enzyme mix	1.5	
Total volume	9	

95. Add 9 µl of the Megascript T7 mix to the eluted material from Step 91 and resuspend. Cumulative volume is 16 µl.
96. Incubate the mix in a thermocycler at 37 °C for 14 h with the lid heated to 70 °C. Let thermocycler go to 4 °C after the program is finished.
    CRITICAL STEP: Do not let reaction stay in the thermocycler for more than a few hours after the programs is finished, to keep RNA integrity.
    PAUSE POINT: The reaction can be kept at -20 °C up to a day.

(Day 5) Purification of aRNA – Timing: 1 hr

CRITICAL: Depending on the number of samples, bead purifications can be a bottleneck. We therefore do not recommend cleaning too many samples simultaneously.

97. Thaw or put the IVT reaction from Step 96 in ice.
98. Place reaction on a magnet and transfer the reaction without the AMPure XP beads to a new tube.
99. Equilibrate fresh undiluted AMPure XP beads to room temperature for 30 min. Vortex till bead-buffer mix is homogenous.
100. Add 0.8 volume undiluted AMPure XP beads to the reaction and allow the material to bind to the AMPure XP beads for 10 min. Cummulative volume is 28.8 µl**.** Steps 100-108 need to be carried out at room temperature.
101. Put the tube on a magnetic rack and allow AMPure XP beads to accumulate. Keep samples on magnetic rack until step 106.

102. Remove the liquid carefully without disturbing AMPure XP beads.
103. Add 500 µl of fresh 80 % (vol/vol) ethanol and leave for 30 sec.
104. Remove the ethanol carefully without disturbing AMPure XP beads.
105. Repeat steps 103-104. Pulse-spin tube and place in magnetic rack to remove excess ethanol.
106. Let AMPure XP beads to air-dry for 5 min or until they appear "matte".
     CRITICAL STEP: Do not let AMPure XP beads overdry. Elute in water before cracks start appearing in the bead pellet.
107. Add 23 µl of nuclease-free water to the AMPure XP beads and resuspend until beads and water form a homogenous mix. Remove tube from magnetic rack and allow the material to elute for 5 min.
108. Place tube in magnetic rack and without disturbing the AMPure XP beads carefully transfer 22 µl solution to a clean tube and place on ice.

aRNA fragmentation – Timing: 5 min

109. Bring a heat block to 94 °C.
110. Add 0.2 volumes of fragmentation buffer to amplified material (Step 108) while on ice and resuspend. Cumulative volume is 26.4 µl.
111. Quickly transfer tube to 94 °C for 2 min.
112. Remove tube and quickly put on ice.
113. Add 0.1 volume of fragmentation STOP buffer to tube as fast as possible and resuspend. Keep on ice. Cumulative volume is 29.04 µl.
     ? TROUBLESHOOTING

Purification and quantification of fragmented aRNA – Timing: 1 hr 45 min

CRITICAL: Depending on the number of samples, bead purifications can be a bottleneck. We therefore do not recommend cleaning too many samples simultaneously.

CRITICAL: Depending on the number of pooled cell lysates, extra rounds of bead purifications can increase the library prep efficiency because of adapter depletion. For 384 pooled cells we recommend at least 2 rounds of bead purification at this stage of the protocol. For 96 pooled cells we recommend 1 round of bead purification, as stated in steps 114-124.

114. Equilibrate undiluted AMPure XP beads to room temperature for 30 min. Vortex till bead-buffer mix is homogenous.
115. Add 0.8 volume undiluted AMPure XP beads to the reaction and allow the material to bind to the AMPure XP beads for 10 min. Cumulative volume is 52.3 µl. Steps 115-124 need to be carried out at room temperature.
116. Put the tube on a magnetic rack and allow AMPure XP beads to accumulate. Keep samples on magnetic rack until step 122.
117. Remove the aqueous phase carefully without disturbing AMPure XP beads.
118. Add 500 µl of fresh 80 % (vol/vol) ethanol and leave for 30 sec.
119. Remove the ethanol carefully without disturbing AMPure XP beads.
120. Repeat steps 118-119.
121. Pulse-spin tube and place in magnetic rack to remove excess ethanol.

122. Let AMPure XP beads to air-dry for 5 min or until they appear “matte”.
CRITICAL STEP: Do not let AMPure XP beads overdry. Elute in water before cracks start appearing in the bead pellet.
123. Add 13 µl of nuclease-free water to the AMPure XP beads and resuspend until beads and water form a homogenous mix. Remove tube from magnetic rack and allow the material to elute for 5 min.
124. Place tube in magnetic rack and without disturbing the AMPure XP beads carefully transfer 12 µl solution to a clean tube and place on ice.
PAUSE POINT: The sample can be stored at -80 °C up to 6 months.
125. Measure 1 µl of aRNA with a Bioanalyzer RNA pico chip by following the kit manual. Examples of successful and less successful IVT reactions and aRNA preparation for library prepatation are shown in Fig. 4.
? TROUBLESHOOTING

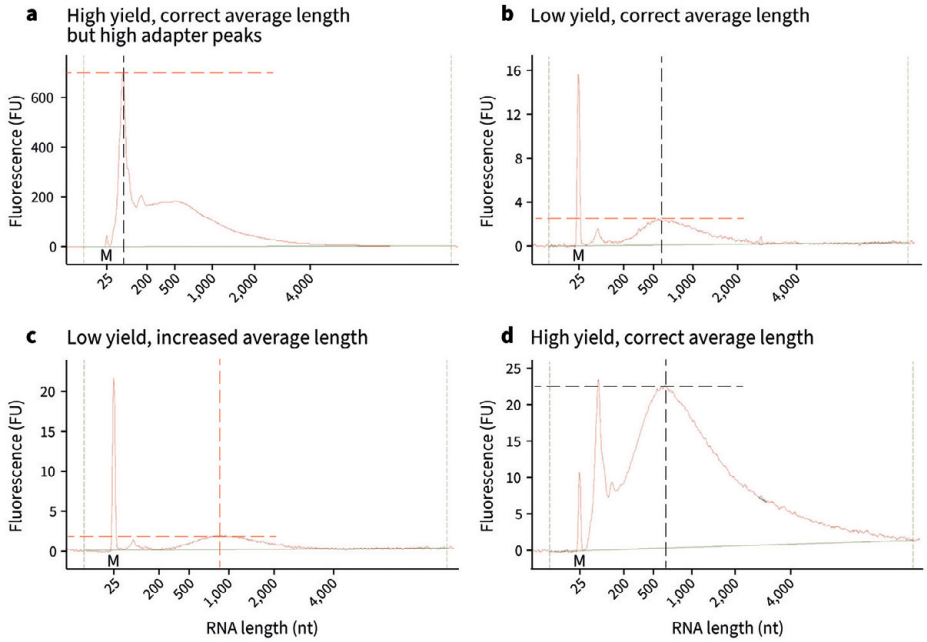


Figure 4: Examples of aRNA bioanalyzer plots
Bioanalyzer results after IVT, bead purification, aRNA fragmentation and another bead purification. **a**, The aRNA has the correct size distribution of 300–700 nt, but the adapter peak is extremely high (>600 FU), which can inhibit the library preparation. Extra rounds of aRNA bead purifications are recommended. **b**, The aRNA has the correct size distribution, but the yield is low (<4 FU), possibly due to loss during bead purification. **c**, The aRNA has an increased size distribution of 500–2,000 nt, indicating that fragmentation was not complete. In addition, the yield is low (<2.5 FU), indicating loss during bead purification. **d**, The aRNA has the correct size distribution and good yield (>20 FU) and can be used for library preparation. Peaks marked with an ‘M’ indicate the reference marker; black and red dashed lines indicate the relevant optimal and suboptimal features, respectively.

Reverse transcription – Timing: 1 hr 30 min

126. Prepare the randomhexRT mix as indicated in the table below and heat in thermocycler at 65 °C for 5 min with lid at 100 °C. Immediately put on ice. Reverse transcription will generate a DNA molecule complementary to the aRNA molecule. This is done by using poly-N primers containing the P7 Illumina adapter in their overhang.

Reagent	Amount (µl)	Final concentration in mix
aRNA (step 125)	5	
randomhexRT primer (20 µM)	1	13.33 µM
dNTP mix (10 mM)	0.5	3.33 mM
Total volume of reaction	6.5	

127. Prepare the RT mix as indicated in the table below. Keep mix on ice and enzymes in ice block at all times. Heat in thermocycler at 25 °C for 10 min, then 42 °C for 1 hr with the cycler lid at 50 °C.

Reagent	Amount (µl)	Final concentration in mix
hexRT mix with aRNA (step 126)	6.5	
5x First strand buffer	2	2.5 x
DTT (0.1 M)	1	25 mM
RNase OUT (40 U/µl)	0.5	5
Superscript II (200 U/µl)	0.5	25
Total volume of reaction	10.5	

PCR indexing – Timing: 30 min

CRITICAL: Do not overamplify the material. We recommend always a minimum of 8 cycles of PCR for the generation of enough Illumina indexed molecules. We recommend 10 PCR cycles or more for aRNA product between 1 and 10 Fluorescent Units (FU) and 8-9 cycles for aRNA product higher than 10 FU (Fig. 4). Prepare the indexing mix as indicated in the table below. Keep mix on ice. Index each sample with a unique RPi primer for multiplexing. This step completes the P5 and P7 ends of the molecules by indexing the libraries.

Reagent	Amount (µl)	Final concentration in mix
Reverse transcribed aRNA (step 127)	10.5	
Nuclease-free water	10.5	
2x NEBNext High-Fidelity PCR Master Mix	25	1.26 x
RNA PCR primer RP1 (10 µM)	2	0.5 µM
RNA PCR index primer RPi (10 µM)	2	0.5 µM
Total volume of reaction	50	

128. Run PCR program in thermocycler with the lid heated at 105 °C as indicated in the table below.

Cycle number	Denature	Anneal	Extend
1	98°C, 30 s		
2-12	98°C, 10 s	60°C, 30 s	72°C, 30 s
13			72°C, 10 min

Library purification – Timing: 1 hr 30 min

CRITICAL: Depending on the number of samples, bead purifications can be a bottleneck. We therefore do not recommend cleaning too many samples simultaneously.

129. Equilibrate undiluted AMPure XP beads to room temperature for 30 min. Vortex till bead-buffer mix is homogenous.
130. Add 0.8 volume undiluted AMPure XP beads to the reaction and allow the material to bind to the AMPure XP beads for 10 min. Cumulative volume is 90 µl. Steps 130-143 need to be carried out at room temperature.
131. Put the tube on a magnetic rack and allow AMPure XP beads to accumulate. Keep samples on magnetic rack until step 138.
132. Remove the aqueous phase carefully without disturbing AMPure XP beads.
133. Add 500 µl of fresh 80 % (vol/vol) ethanol and leave for 30 sec.
134. Remove the ethanol carefully without disturbing AMPure XP beads.
135. Repeat steps 133-134. Pulse-spin tube and place in magnetic rack to remove excess ethanol.
136. Let AMPure XP beads to air-dry for 5 min or until they appear “matte”.

CRITICAL STEP Do not let AMPure XP beads overdry. Elute in water before cracks start appearing in the bead pellet.

137. Add 26 µl of nuclease-free water to the AMPure XP beads and resuspend until beads and water form a homogenous mix. Remove tube from magnetic rack and allow the material to elute for 5 min.
138. Place tube in magnetic rack and without disturbing the AMPure XP beads carefully transfer 25 µl elution to a clean tube and place on ice.
139. Repeat steps 130-137 to re-clean the eluted material of step 138.
140. Add 16 µl of nuclease-free water to the AMPure XP beads and resuspend until beads and water form a homogenous mix. Remove tube from magnetic rack and allow the material to elute for 5 min.
141. Place tube in magnetic rack and without disturbing the AMPure XP beads carefully transfer 15 µl solution to a clean tube and place on ice.

PAUSE POINT: The finished libraries can be kept at -20 °C indefinitely.

Library quantification and sequencing – Timing: 19 hr

142. Measure the concentration of the sample with the Qubit dsDNA HS Assay kit by following the kit manual.

143. Measure 1 µl of sample with a Bioanalyzer HS DNA chip by following the kit manual. Examples of successful and less successful library prep reactions and purifications are shown in Fig. 5.
? TROUBLESHOOTING

144. Dilute samples to appropriate molarity for sequencing and pool according to the desired number of reads per sample. We dilute to 4 nM and in case of low concentration, samples can be diluted to 2 nM.

145. Submit samples to the sequencing facility. To ensure cluster formation, avoid sequencing fewer than 8 different indexed libraries in one run. We sequence samples on the NextSeq 500, using 75 bp for both Read 1 and Read 2 (paired-end) and spike-in 20 % PhiX. This percentage of PhiX control is specific for the samples generated with this protocol and ensures cluster formation by enriching the complexity of the sequencing pool.

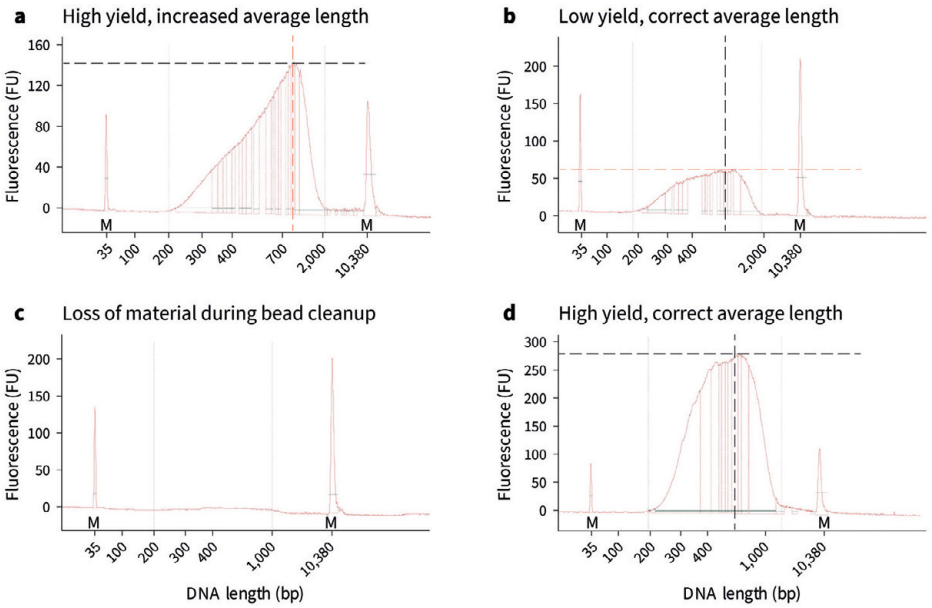


Figure 5: Examples of DNA bioanalyzer plots
Bioanalyzer results after library construction: **a**, The library shows good yield (>150 FU) but size distribution is slightly increased (peaking at 1,000 - 1,500 nt), probably due to increased fragment length of aRNA. **b**, The library shows correct length distribution (250 - 900 nt) but low yield. If library was constructed from aRNA similar to example 4 from 3a, material was probably lost during bead purification. Repeat library preparation with leftover aRNA. **c**, Lack of both the adapter and the PCR product indicate complete loss during bead purification or failed library preparation. **d**, The library shows high yield >250 FU and correct size distribution of 250 - 700 nt. FU, fluorescence units. nt, nucleotides. Peaks marked with an "M" indicate the reference markers; black and red dashed lines indicate the relevant optimal and suboptimal features, respectively.

Downloading genome reference files and generating HISAT2 index – Timing: 1 hr 30 min

CRITICAL: For the purpose of this protocol, the FASTA file, GTF file and HISAT2 index will all be placed in a folder named "references". For that reason, generate this folder in your own working directory, or replace all mentions of the "references" directory with the path of your own choice: mkdir ./references

CRITICAL: Note that this command and all future commands are executed from the terminal.

146. Download the reference sequence of the relevant species, for example from http://www.ensembl.org/info/data/ftp/index.html. Click the "DNA (FASTA)" link and download the file ending with 'dna.primary_assembly.fa.gz'. Unzip the downloaded file and place it in the "references" directory.
147. Download the GTF file for the relevant species, for example from http://www.ensembl.org/info/data/ftp/index.html. Unzip downloaded file and place it in the "references" directory.
148. If using ERCC spike-ins, their sequences should be added to the genome FASTA file and the GTF file. For clarity, we add "with_ERCC" to the FASTA and GTF filename in this walkthrough.
149. Generate a HISAT2 index:

```
HISAT2_INDEX="./references/Mus_musculus.GRCm38.dna.primary_assembly.with_ERCC"
FASTAFN="./references/Mus_musculus.GRCm38.dna.primary_assembly.with_ERCC.fa"
hisat2-build $FASTAFN $HISAT2_INDEX
```

Installing scDam&T-seq scripts – Timing: 10 min

CRITICAL: The analysis steps in this and subsequent sections demonstrate how scDam&T-seq data can be analysed using the provided software package (scDamAndTools). File names and genome references are chosen to match the test data available as part of the GitHub repository (in the "tutorial" folder). The included data represents five single cells from a mESC Dam-LaminB1 experiment. Care should be taken to modify file names when applying the analysis on other data. A more detailed explanation of all steps and functions can be found in Table 1 and on the GitHub page (https://github.com/KindLab/scDamAndTools), where we have also included the expected results from processing the test data.

150. Generate a Python3 virtual environment and activate the virtual environment:

```
python3 -m venv $HOME/.venvs/tutorial
source ~/.venvs/tutorial/bin/activate
```

PAUSE POINT: To deactivate the virtual environment:

```
deactivate
```

151. Install prerequisite modules:

```
pip install --upgrade pip wheel setuptools
pip install cython
```

152. Install scDam&T-seq package:

```
pip install git+https://github.com/KindLab/scDamAndTools.git
```

Generate GATC reference arrays – Timing: 1 hr 30 min

153. To efficiently match obtained DamID reads to specific instances of the GATC motif in the genome, we generate two reference arrays. The first array ("position array" or "posarray") contains the positions of all GATC positions in the genome. The second array ("mappability array" or "maparray") indicates whether it is possible to uniquely align a read derived from a particular (strand-specific) GATC instance. The mappability array is used to filter out ambiguously aligning GATCs and can serve as an indicator of the (mappable) GATC density along the chromosomes. During the generation of the mappability array, in silico reads are generated for each GATC instance and are subsequently mapped back to the reference genome. The length of the reads should be chosen to be the same as the length of the reads obtained in the experiment (excluding the UMI and barcode):

```
IN_SILICO_READLENGTH=62
ARRAY_PREFIX="./refarrays/Mus_musculus.GRCm38.dna.primary_assem-
bly"
create_motif_refarrays \
 -m "GATC" \
 -o $ARRAY_PREFIX \
 -r $IN_SILICO_READLENGTH \
 -x $HISAT2_INDEX \
 $FASTAFN
```

The script generates three files ending in ".positions.bed.gz" (all occurrences of the GATC motif in BED format), ".posarray.hdf5" (all occurrences of the GATC motif as a HDF5 array), and ".maparray.hdf5" (the mappability of all GATC motifs as a HDF5 array), respectively.

Demultiplex raw data – Timing: 1 hr

154. If not already done so by the sequencing facility, demultiplex the raw data based on the used Illumina indices.

155. In a text editor or Microsoft Excel, create a tab-delimited text file [43] that describes the barcodes that were used in the library. The file should have two columns, listing the adapter names and sequences respectively. The location and length of UMIs should be indicated with numbers and dashes. Using DamID and CEL-Seq2 barcodes as specified in Experimental Design (Desgin and concentration of DamID adapters and CEL-Seq2 primers), the barcode file of a library with two samples should look as follows:

```
DamID_BC_001 3-TGCT-3-GAGAGA
DamID_BC_002 3-ATTG-3-GAACGA
CELseq_BC_001 3-ACAG-3-AGGC
CELseq_BC_002 3-GTCT-3-GCCA
```

Example data and relevant barcode file are included in the scDamAndTools package in the folder "tutorial".

CRITICAL STEP: There may be multiple raw sequencing files pertaining to the same samples, for example from the different sequencing lanes. These files should be concatenated prior to the processing of DamID and CEL-Seq2 reads (step 157 and 159, respectively).

156. For each Illumina library, demultiplex the reads based on the used adapters. Make sure the output file format contains the fields "{name}" and "{readname}", where the barcode name and paired-end read name will be inserted:

```
OUTFMT="./data/demultiplexed/index01.{name}.{readname}.fastq.gz"
INFOFN="./data/demultiplexed/index01.demultiplex_info.txt"
demultiplex.py \
 -vvv \
 --mismatches 0 \
 --outfmt $OUTFMT \
 --infofile $INFOFN \
 ./metadata/index01.barcodes.tsv \
 ./data/raw/index01_R1_001.fastq.gz \
 ./data/raw/index01_R2_001.fastq.gz
```

The demultiplex script generates a separate FASTQ file for each barcode provided in the barcode information file (see step 155) that contains all reads matching this barcode. In addition, a text file ("index01.demultiplex_info.txt") is generated that details the number of reads matched to each barcode.

Process DamID demultiplexed files – Timing: 2 hr

CRITICAL: The subsequent steps (157-158) need to be performed on all DamID demultiplexed files. It is highly recommended that this process be parallelized on a high-performance computing cluster. The amount of time necessary for these steps depends entirely on the number of libraries, samples per library and available computing cores.

157. Process the DamID reads to arrays of (UMI-unique) GATC counts. The script aligns the DamID reads to the genome and subsequently matches them to positions as indicated in the position array (see step 153). Since the GATC motif is cleaved in half by DpnI, the prefix "GA" is added to all reads prior to alignment. PCR-duplicates are filtered out based on the available UMI information. For this step, only the R1 reads are used since these contain the genomic sequence aligning to the GATC motif:

```
OUTPREFIX="./data/damid/index01.DamID_BC_001";
POSARRAY="./refarrays/Mus_musculus.GRCm38.dna.primary_assembly.
GATC.posarray.hdf5";
process_damid_reads \
-o $OUTPREFIX \
-m "GA" \
-p $POSARRAY \
-x $HISAT2_INDEX \
-u \
./data/demultiplexed/index01.DamID_BC_001.R1.fastq.gz
```

The script generates an alignment file ending in ".sorted.bam", a GATC count file ending in ".counts.hdf5" and an information file ending in ".counts.stats.tsv".

158. Bin the GATC count files into genomically equal-sized bins. The resulting HDF5 file contains for each chromosome the number of observed UMI-unique counts for each bin:

```
MAPARRAY="./refarrays/Mus_musculus.GRCm38.dna.primary_assembly.
GATC.readlength_62.maparray.hdf5"
OUTFN="./data/damid/index01.DamID_BC_001.counts.binsize_100000.
hdf5"
bin_damid_counts.py \
-vvv \
--mapfile $MAPARRAY \
--posfile $POSARRAY \
--binsize 100000 \
--outfile $OUTFN \
./data/damid/index01.DamID_BC_001.counts.hdf5
```

The output of this step is a single HDF5 file ending in ".binsize_100000.hdf5" that contains the number of unique counts observed in all 100kb bins in the genome.

Process CEL-Seq2 demultiplexed files – Timing: 4 hr

CRITICAL: The subsequent step (159) needs to be performed on all CEL-Seq2 demultiplexed files. It is highly recommended that this process be parallelized on a high-performance computing cluster. The amount of time necessary for these steps depends entirely on the number of libraries, samples per library and available computing cores.

159. Process the CEL-Seq2 reads to an array of UMI-unique counts per gene. For this step, the R2 reads are used, since these contain the genomic sequence:

```
OUTPREFIX="./data/celseq/index01.CELseq_BC_001"
GTF="./references/Mus_musculus.GRCm38.98.with_ERCC.gtf"
process_celseq_reads \
```

```
-o $OUTPREFIX \
-g $GTF \
-x $HISAT2_INDEX \
./data/demultiplexed/index01.CELseq_BC_001.R2.fastq.gz
```

The script generates an alignment file ending in ".bam" and a count file ending in ".counts.hdf5". The count file contains the number of observed UMI-unique transcripts per gene, sorted by their Ensembl gene ID as provided in the GTF file.

Timing

Steps 1 - 6, preparation of primer plates, induction of Dam-POI: 1 hr 15 min
Steps 7 - 31, cell harvest, Hoecht staining, sorting: 2 hr 30 min
Steps 32 - 59, lysis, reverse transcription, second strand synthesis, proteinase K: 16 hr
Steps 60 - 78, DpnI digestion, adapter dispension, adapter ligation: 21 hr
Steps 79 - 96, pooling, bead cleanups, in vitro transcription: 16 hr 15 min
Steps 97 - 125, bead cleanups, fragmentation, bead cleanups, RNA quantification: 3 hr
Steps 126 - 145, library prep, DNA quantification, library pooling, sequencing: 23 hr
Steps 146 - 159, analysis: 10 hr

Anticipated results

Figure 6 shows statistics of two example libraries containing 96 single-cell samples of a Dam-LMNB1 mESC line. The two libraries are biological replicates that were collected, processed and sequenced at different times. We obtained ~45 and ~92 million reads for replicate 1 and 2, respectively, which we consider a high sequencing depth for these samples. Nearly all these reads (>95%) can be successfully assigned to a DamID or CEL-Seq2 barcode (Fig. 6a). Most of the demultiplexed reads successfully align (Fig. 6b), after which invalid reads are filtered out. CEL-Seq2 reads are considered to be invalid if they do not align properly to a gene or when a read's mapping score is lower than a set threshold. DamID reads are considered to be invalid when they do not align to a GATC position or when their mapping quality is too low. In a successful experiment, ~60% of the demultiplexed reads are valid, and 20–40% are unique (Fig. 6b).

Although most reads are demultiplexed to a DamID barcode (~90%) and a much smaller fraction to a CEL-Seq2 barcode (~5%), the resulting CEL-Seq2 data still provide a median of 11,000–12,000 unique transcripts (Fig. 6c) and 3,700–3,900 unique genes per cell, which is more than sufficient to perform typical single-cell transcriptome analyses. The DamID data, on the other hand, contain a median of ~125,000–175,000 unique GATC counts per sample. We typically exclude samples from our analyses that contain <10,000 unique counts, but the appropriate threshold depends on the POI. The reason why more DamID material than CEL-Seq2 material is obtained is not entirely clear, but likely has to do with the efficiency with which transcripts and gDNA anneal to the primer and adapter, respectively. We also find that the ratio between DamID and CEL-Seq2 reads varies depending on the POI and methylation level in the cell, with a higher fraction of CEL-Seq2 reads for POIs that methylate a smaller fraction of the genome. If the depth and quality are satisfactory, the resulting DamID data can be used for downstream analysis. We typically work with binned data at a resolution of 50–100 kb for single-cell samples.

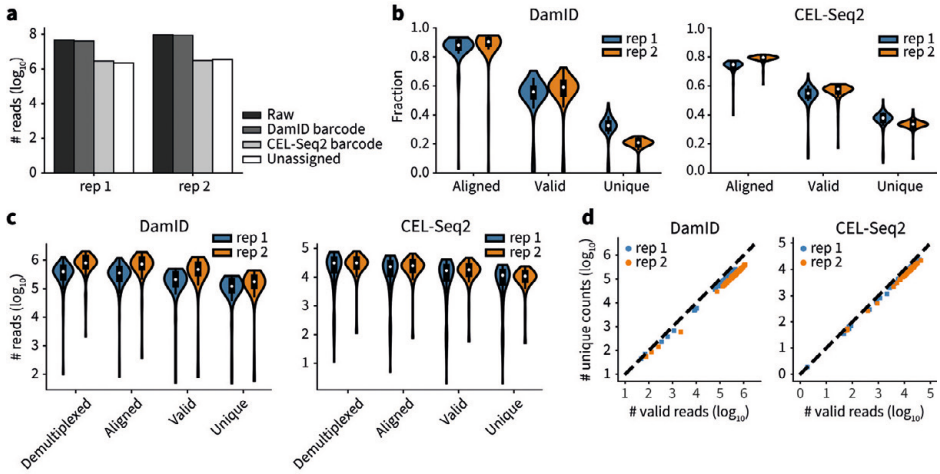


Figure 6: Technical statistics of a scDam&T-seq run

a, Barplot showing the number of raw sequencing reads and the number of DamID, CEL-Seq2 and unassigned reads for two technical replicates (rep1 and rep2) after demultiplexing. **b-c**, Overview of progressive read loss during analysis of DamID (left) and CEL-Seq2 (right) data as a fraction of the number of demultiplexed reads per sample (**b**) and in absolute numbers (**c**). The black boxplots show the median (white dot), interquartile range (black box) and range of the data within 1.5 times the interquartile range (IQR) of the median (black lines). **d**, Complexity plot of the DamID (left) and CEL-Seq2 (right) data. Each replicate in plot **b–d** shows data of one library of 96 single-cell samples (i.e., 96 data points). F1 mESCs with a hybrid genetic background of 129/Sv and Cast/EiJ were used (RRID: CVCL_XY63)39. Cells were negative for mycoplasma contamination.

In general, the final number of unique GATCs is strongly influenced by the number of samples in the library, the number of PCR cycles used during library preparation and the sequencing depth. However, loss of DamID reads can also occur when something goes wrong during the experiment or when the expression of Dam-POI is too low, which results in a high fraction of invalid reads. For that reason, we look at the fraction of valid DamID reads, which should be >50% for most samples (Fig. 6b). To establish the complexity of the samples, we compare the number of valid reads to the final number of unique counts (Fig. 6d). The libraries shown here have a DamID complexity of 58% and 35% and a CEL-Seq2 complexity of 68% and 58% for replicate 1 and 2, respectively. If the data are too sparse to execute the desired analyses (e.g., for LMNB1, a median of <10,000 unique counts per cell), libraries with a complexity of >30% may be considered for resequencing. Possible reasons for poor data quality and potential solutions are discussed in Table 2 (Troubleshooting).

Table 2: **Troubleshooting**

Step	Problem	Possible reason	Solution
28	Unexpected cell-cycle profile	Hoechst too old. Hoechst undergone too many freeze-thaw cycles.	Prepare a fresh Hoechst solution.
89, 125	Low or no aRNA product on Bioanalyzer (Fig. 3a)	Loss of material during bead purifications.	Remove all ethanol before elution.
113	Increased aRNA product distribution (Fig. 3a)	Inefficient fragmentation	Resuspend fragmentation buffer and make sure whole tube makes contact with the heat block during fragmentation at 94 °C.
143	Low or no library product on Bioanalyzer (Fig. 3b)	Loss of material during bead purifications. Failed library preparation.	Remove all ethanol before elution. Repeat library prep with leftover aRNA. Increase PCR cycles.
143	Low product (Fig. 3b)	High adapter/product ratio inhibits library prep.	Bead purify the aRNA 1-2 times extra Repeat library prep.
Post data processing	Low complexity	Too little material Too many PCR cycles Too deeply sequenced	If possible, increase the number of samples in a library and decrease the number of PCR cycles. Make sure the amount of material included in a sequencing run is proportional to the expected output.
Post data processing	A uniform DamID signal over the whole genome; signal is very similar to the mappable GATC density	Too high expression of the Dam-POI Too long induction of the Dam-POI Leaky induction system for Dam-POI expression	Select a clone with lower expression levels or modify the Dam-POI construct. Optimize the time of induction of the selected clone. Ensure there is no expression of Dam-POI prior to induction.
Post data processing	Little DamID signal, despite good signal in positive control	Too low expression of Dam-POI Too short induction	Select a clone with a higher expression level or modify the Dam-POI construct. Induce expression for a longer time.
Post data processing	Sparse CEL-Seq2 data compared to DamID data	Low transcript content of cells Transcript degradation Errors in CEL-Seq2 primer plate preparation Very efficient preparation and amplification of DamID material	Include positive controls to make sure single-cell transcription data can be obtained from the material (e.g. WT control). Renew CEL-Seq2 primer plate. Optimize CEL-Seq2 primer and DamID adapter concentrations, e.g. reduce DamID adapter concentration.

Table 2: **Continued**

Step	Problem	Possible reason	Solution
Post data processing	Low fraction of valid DamID reads	Too low expression of Dam-POI Presence of many random gDNA breaks High adapter contamination	Select clone with optimal expression levels. Include negative control (e.g. WT control). Reduce adapter concentrations or perform additional bead purifications.

Acknowledgements

We would like to thank the members of the Kind lab for their comments on the manuscript. S.S.D. acknowledges support from the Center for Scientific Computing at UCSB: an NSF MRSEC (DMR-1720256) and NSF CNS-1725797. This work was funded by a European Research Council Starting grant (ERC-StG 678423-EpiID), a Nederlandse Organisatie voor Wetenschappelijk Onderzoek[37] Open (824.15.019) and ALW/VENI grant (016.181.013). The Oncode Institute is supported by KWF Dutch Cancer Society.

Author contributions

K.R., S.S.D. and J.K. designed the study. S.S.D. developed the method with input and assistance from D.M. K.R. supervised and performed bioinformatic analyses and developed the scDam&T computational pipeline. F.J.R. performed cloning and bioinformatic analyses on mESC scDam&T data and developed the clonal selection strategy. C.M.M. optimized the method, performed experiments, generated cell lines and designed the protocol for scDamID2. S.S.d.V. generated cell lines, assisted with experiments and designed the protocol for bulk DamID2. S.J.A.L. generated cell lines. K.L.d.L. assisted with experiments. A.C. assisted with analyses. J.K. and S.S.D. conceived and supervised the study. C.M.M. and F.J.R. wrote the manuscript with input from J.K.

Data availability

A test dataset is available from GitHub (https://github.com/KindLab/scDamAndTools).

Code availability

All codes are available from GitHub (https://github.com/KindLab/scDamAndTools). The code in this manuscript has been peer-reviewed.

References

1. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502 (2007).
2. Vogel, M.J., Peric-Hupkes, D. & van Steensel, B. Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat Protoc* **2**, 1467-1478 (2007).
3. Crawford, G.E. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**, 123-131 (2006).
4. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
5. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
6. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134-147 (2015).
7. Flyamer, I.M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114 (2017).
8. Stevens, T.J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59-64 (2017).
9. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).
10. Cusanovich, D.A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
11. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142-146 (2015).
12. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **23**, 2126-2135 (2013).
13. Smallwood, S.A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817-820 (2014).
14. Farlik, M. et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* **10**, 1386-1397 (2015).
15. Mooijman, D., Dey, S.S., Boisset, J.C., Crosetto, N. & van Oudenaarden, A. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat Biotechnol* **34**, 852-856 (2016).
16. Wu, X., Inoue, A., Suzuki, T. & Zhang, Y. Simultaneous mapping of active DNA demethylation and sister chromatid exchange in single cells. *Genes Dev* **31**, 511-523 (2017).
17. Zhu, C. et al. Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **20**, 720-731 e725 (2017).
18. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33**, 1165-1172 (2015).
19. Harada, A. et al. A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat Cell Biol* **21**, 287-296 (2019).
20. Hainer, S.J., Boskovic, A., McCannell, K.N., Rando, O.J. & Fazzio, T.G. Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* **177**, 1319-1329 e1311 (2019).
21. Ku, W.L. et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nat Methods* **16**, 323-325 (2019).
22. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* **13**, 229-232 (2016).
23. Hou, Y. et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* **26**, 304-319 (2016).
24. Clark, S.J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**, 781 (2018).

25. Rooijers, K. et al. Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. *Nat Biotechnol* **37**, 766-772 (2019).

26. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666-673 (2012).

27. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 77 (2016).

28. Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T. & Kanemaki, M. An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nat Methods* **6**, 917-922 (2009).

29. Boers, R. et al. Genome-wide DNA methylation profiling using the methylation-dependent restriction enzyme LpnPI. *Genome Res* **28**, 88-99 (2018).

30. Sen, M. et al. Strand-specific single-cell methylomics reveals distinct modes of DNA demethylation dynamics during early mammalian development. *bioRxiv*, 804526 (2019).

31. Borsos, M. et al. Genome-lamina interactions are established de novo in the early mouse embryo. *Nature* **569**, 729-733 (2019).

32. Liu, C.L., Schreiber, S.L. & Bernstein, B.E. Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* **4**, 19 (2003).

33. Grosselin, K. et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* **51**, 1060-1066 (2019).

34. Kaya-Okur, H.S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).

35. Schmid, M., Durussel, T. & Laemmli, U.K. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell* **16**, 147-157 (2004).

36. Skene, P.J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6** (2017).

37. Sirunyan, A.M. et al. Search for rare decays of Z and Higgs bosons to J / psi and a photon in proton-proton collisions at s = 13 TeV. *Eur Phys J C Part Fields* **79**, 94 (2019).

38. Tosti, L. et al. Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo. *Genome Res* **28**, 592-605 (2018).

39. Monkhorst, K., Jonkers, I., Rentmeester, E., Grosveld, F. & Gribnau, J. X inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator. *Cell* **132**, 410-421 (2008).

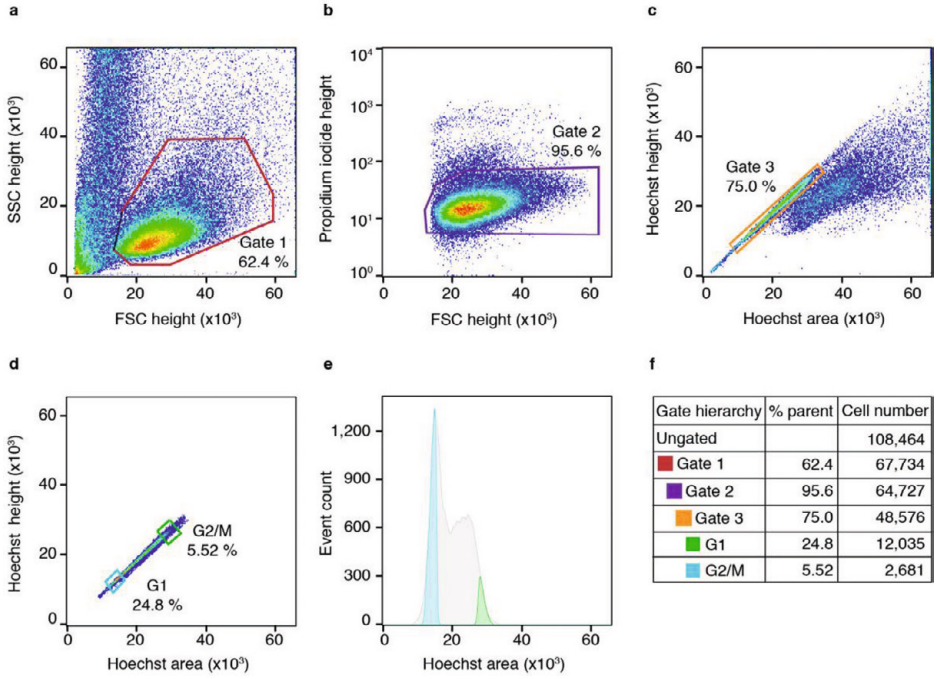
40. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).

41. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

42. Ditzel, M. et al. Biologic meshes are not superior to synthetic meshes in ventral hernia repair: an experimental study with long-term follow-up evaluation. *Surg Endosc* **27**, 3654-3662 (2013).

43. Aad, G. et al. Observation of associated near-side and away-side long-range correlations in sqrt[s(NN)]=5.02 TeV proton-lead collisions with the ATLAS detector. *Phys Rev Lett* **110**, 182302 (2013).

Supplementary Figures



f

Gate hierarchy	% parent	Cell number
Ungated		108,464
Gate 1	62.4	67,734
Gate 2	95.6	64,727
Gate 3	75.0	48,576
G1	24.8	12,035
G2/M	5.52	2,681

g

Population	Mean (Hoechst area)
G1	14,463
G2/M	28,670

Supplementary Figure 1: Gating strategy for FACS
a, Dot plot of ungated mESCs showing gating strategy to exclude debris in FSC (forward scatter) versus SSC (side scatter). The percentage of events in Gate 1 is indicated. **b**, Dot plot of mESCs passing Gate 1, showing the gating strategy to exclude dead cells in propidium iodide versus FSC. The percentage of events in Gate 2 is indicated. **c**, Dot plot of mESCs passing Gate 2, showing the gating strategy to exclude duplet cells in Hoechst versus Hoechst area. The percentage of events in Gate 3 is indicated. **d**, Dot plot of mESCs passing Gate 3 were gated for DNA content in G1 and G2/M phase of the cell cycle. The gate for the G2/M population was defined by doubling the intensity value of the G1 peak maximum. **e**, DNA content histogram events in Gate 3, showing counted events versus Hoechst area. Only cells passing gate G2/M were sorted. **f**, Table indicating gate hierarchy, percentage of events in each gate relative to parent population and total numbers of events within each gate. **g**, Table indicating the mean value for the G1 and G2/M populations. All measurements were done on the BD FACSJazz and analyzed with FlowJo software, version 10.1r5.

Supplementary Tables

Supplementary Table 1: **CEL-Seq2 primers**
Available in online version: https://doi.org/10.1038/s41596-020-0314-8

Supplementary Table 2: **scDamID double-stranded adapters**
Available in online version: https://doi.org/10.1038/s41596-020-0314-8

Supplementary Methods: scDamID2

Lysis plate preparation – Timing: 45 min for one plate

CRITICAL: It is crucial that the working area is sufficiently clean when working with single cell material. DNAZap and RNAseZAP treatment is required in steps 1-5 and 8-25. RNAseZAP treatment is sufficient for steps 26-33. Ethanol 80 % (vol/vol) treatment is sufficient for steps 34-35.

CRITICAL: We recommend preparing lysis plates, adapter plates and master mixes before single-cell material amplification by IVT in a PCR workstation.

1. Pipet 5 µl of mineral oil in each well of a 384-well plate and seal the plate with an aluminium seal.
2. Prepare the lysis mix according to the table below. Keep the mix on ice at all times.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Lysis buffer 1.03 x	290	162.5	1 x
Proteinase K (20 mg/ml)	10	5.6	0.67 mg/ml
Total volume	300		

3. In an 8-well PCR strip, aliquot 20.7 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
4. Dispense 300 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 300 nl.
5. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C.
6. **PAUSE POINT:** The lysis plates can be kept at 4 °C until sort on the same day or maximum overnight.

Dam-POI induction – Timing: 45 min performed in a cell culture hood

7. Induce cells as stated in **scDam&T Procedure** steps 4-6.

Prepare cells for FACS sorting by Hoechst staining – Timing: 1 hr 30 min

8. Prepare cells for FACS sorting as stated in **scDam&T Procedure** steps 7-31.
   PAUSE POINT: Proceed with lysis after FACS sorting or store the sorted plates at -20 °C for several months.

Lysis – Timing: 8 hr 20 min

9. Thaw the plates on ice if stored at -20 °C.
10. Put the plate in a thermocycler at 50 °C for 8 hr, then 80 °C for 20 min with the thermocycler lid at 100 °C and let the machine go to 4 °C at the end of the program.
    PAUSE POINT: The lysed plates can be kept at -20 °C for several months.

DpnI digestion – Timing: 8 hr 35 min

11. Thaw the plates on ice if stored at -20 °C.
12. Thaw the 10x CutSmart buffer at room temperature and keep on ice. Keep the DpnI enzyme on an ice block at all times.
13. Prepare the DpnI mix according to the table below. Keep the mix on ice at all times.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	590	272.8	
10x CutSmart buffer	100	46.2	1.6 x
DpnI (20 U/µl)	10	4.6	0.33 U/µl
Total volume	700		

14. In an 8-well PCR strip, aliquot 39.9 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
15. Dispense 700 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 1000 nl.
16. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C
17. Put the plate in a thermocycler at 37 °C for 8 hr, then at 80 °C for 20 min with the lid at 100 °C and then place on ice for 1-2 min to cool down.
18. Centrifuge at 2,000 g for 1 min at 4 °C

Adapter dispension – Timing: 45 min for 1 plate

19. Dispense adapters as stated in **scDam&T Procedure** steps 67-71 but dispense 400 nl of adapter instead of 50 nl. The cumulative reaction volume is 1400 nl.

Adapter ligation – Timing: 12 h 45 min overnight reaction

20. Thaw the 10x Ligase buffer on ice. Keep the T4 ligase on an ice block at all times.
21. Prepare the ligation mix according to the table below. Keep the mix on ice at all times.

Reagent	Amount for 1 well (nl)	Amount for 1 plate (µl)	Final concentration in mix
Nuclease-free water	375	178.0	
10x Ligase buffer	200	94.9	3.33 x
T4 Ligase (5 U/**µl)**	25	11.8	0.20 U/µl
Total volume	600		

22. In an 8-well PCR strip, aliquot 35.1 µl of mix per well of the strip. Spin for 3-5 sec on tabletop spinner. Keep on ice at all times.
23. Dispense 600 nl per well with the Nanodrop II robot (Box 2). The cumulative reaction volume is 2000 nl.
24. Seal the plate and centrifuge at 2,000 g for 1 min at 4 °C
25. Put the plate in a thermocycler at 16 °C for 12 hr, then 65 °C for 10 min with the lid at 100 °C. Let the thermocycler go to 4 °C at the end of the program.

26. Centrifuge at 2,000 g for 1 min at 4 °C
    PAUSE POINT: The processed plate can be kept at -20 °C up to a month.

Pool cells – Timing: 1 hr for one plate

27. Pool cells as stated in **Procedure** steps 79-82.

Purification of barcoded material – Timing: 1 hr

28. Purify the barcoded material as stated in **scDam&T Procedure** steps 83-91 but use a lower dilution of AMPure XP beads (Reagent setup) and 1.0 volume diluted AMPure XP beads to purify material instead of 0.8 volume.

Amplification by in vitro transcription – Timing: 14 hr 15 min

29. Amplify the barcoded material as stated in **scDam&T Procedure** steps 92-96.

Purification of aRNA – Timing: 1 hr

30. Purify the aRNA as stated in **scDam&T Procedure** steps 97-108.

aRNA fragmentation – Timing: 5 min

31. Fragment the aRNA as stated in **scDam&T Procedure** steps 109-113.

Purification and quantification of fragmented aRNA – Timing: 1 hr 45 min

32. Purify the aRNA as stated in **scDam&T Procedure** steps 114-125.

Reverse transcription – Timing: 1 hr 30 min

33. Reverse transcribe the aRNA for library preparation as stated in **scDam&T Procedure** steps 126-127.

PCR indexing – Timing: 30 min

34. Index PCR the reverse transcribed material as stated in **scDam&T Procedure** step 128.

Library purification – Timing: 1 hr 30 min

35. Purify the libraries as stated in **scDam&T Procedure** steps 129-141.

Library quantification and sequencing – Timing: 19 hr

36. Prepare the libraries for sequencing submission as stated in **scDam&T Procedure** steps 142-145.
    PAUSE POINT: The finished libraries can be kept at -20 °C indefinitely.

Supplementary Methods: DamID2 in bulk

Dam-POI induction – Timing: 45 min performed in a cell culture hood

1. Induce cells as stated in **scDam&T Procedure** steps 4-6.

Genomic DNA isolation – Timing: 3 hr

2. Harvest cells χ hr after induction and extract gDNA following a standard gDNA extraction protocol. We use the Wizard Genomic DNA Purification Kit by Promega.

DpnI digestion – Timing: 12 hr 30 min

3. Prepare the following mix.

Reagent	Amount per reaction (µl)	Final concentration in reaction
gDNA (50 ng/µl)	5	25 ng/µl
10x CutSmart buffer	1	1 x
DpnI (20 U/µl)	0.25	0.5 (U/µl)
Nuclease-free water	3.75	
Total volume	10	

4. Digest the gDNA at 37 °C for 12 hr, then 80 °C for 20 min with the lid at 100 °C. Let the thermocycler go to 4 °C at the end of the program.

Adapter ligation – Timing: 16 hr 30 min overnight reaction

5. Prepare the following mix.
   CRITICAL: In case of more than one sample, use a uniquely barcoded DamID adapter per sample.

Reagent	Amount per reaction (µl)	Final concentration
Digested gDNA (step 4)	10	
DamID adapter (50 µM)	0.5	2 µM
10x Ligase buffer	1.25	1 x
Ligase (5 U/µl)	0.25	0.1 U/µl
Nuclease-free water	0.5	
Total volume	12.5	

6. Ligate the adapters by incubating reaction at 16 °C for 16 hr, then 65 °C for 10 min with the lid at 100 °C. Let the thermocycler go to 4 °C at the end of the program.

Pool samples – Timing: 10 min

CRITICAL: The pooling weight of each sample depends on the yield estimated from a methyl-PCR, and can vary between 0.5 and 10 µl per sample.

CRITICAL: Work on a clean bench free from RNAses. RNAseZAP treatment is sufficient for steps 7-25. Ethanol 80 % (vol/vol) treatment is sufficient for steps 26-28.

7. Pool samples to a maximum total of 2 µg with non-overlapping barcodes in a clean tube.

Purification of barcoded material – Timing: 1 hr

8. Purify the barcoded material as stated in **scDam&T Procedure** steps 83-91 but use undiluted instead of diluted AMPure XP beads, 1.0 volume AMPure XP beads to purify material instead of 0.8 volume and elute in 50 µl nuclease-free water instead of 7 µl.
9. Repeat step 8, but use 0.8 volume AMPure XP beads to purify material instead of 1.0 volume and elute in 25 µl nuclease-free water instead of 50 µl.

Amplification by in vitro transcription – Timing: 2 hr 15 min

10. Thaw the Megascript T7 10x buffer at room temperature. Vortex thoroughly to dissolve precipitates and keep at room temperature.
11. Thaw the Megascript T7 NTPs on ice and keep on ice. Keep the enzyme mix on an ice block at all times.
12. Prepare the Megascript T7 mix as indicated in the table below

Reagent	Amount (µl)	Final concentration in reaction
Cleaned material (step 9)	4	
Nuclease-free water	4	
10x T7 buffer	2	1 x
ATP (75 mM)	2	7.5 mM
UTP (75 mM)	2	7.5 mM
GTP (75 mM)	2	7.5 mM
CTP (75 mM)	2	7.5 mM
T7 enzyme mix	2	
Total volume	20	

13. Incubate the mix in a thermocycler at 37 °C for 2 hr with the lid heated to 70 °C. Let the thermocycler go to 4 °C at the end of the program.

Purification of aRNA – Timing: 1 hr

14. Add 40 µl nuclease-free water to the aRNA (step 13).
15. Purify the aRNA as stated in **scDam&T Procedure** steps 97-108 but elute in 50 µl instead of 23 µl.
16. Quantify the aRNA by measurement on a Nanodrop spectrophotometer machine and dilute aRNA to 50 ng/µl.

aRNA fragmentation – Timing: 5 min

17. Take 20 µl of diluted aRNA (step 16) and fragment as stated in **scDam&T Procedure** steps 109-113.

Purification and quantification of fragmented aRNA – Timing: 1 hr 45 min

18. Purify the aRNA as stated in scDam&T Procedure steps 114-125 but elute in 15 µl instead of 13 µl.

19. Quantify the aRNA by measurement on a Nanodrop spectrophotometer machine and dilute to a concentration between 2 and 15 ng/µl for subsequent quantification on a Bioanalyzer RNA pico chip. Run chip according to the instructions of the kit.

Reverse transcription – Timing: 1 hr 30 min

20. Prepare the randomhexRT mix as indicated in the table below.

Reagent	Amount (µl)	Final concentration in reaction
aRNA (step 19)	4	
randomhexRT primer (20 µM)	1	3.33 µM
dNTP mix (10 mM)	1	1.66 mM
Total volume	6	

21. Heat mix in thermocycler at 65 °C for 5 min with lid at 100 °C. Immediately put on ice.
22. Prepare the RT mix as indicated in the table below. Keep mix on ice and enzymes on ice block at all times.

Reagent	Amount (µl)	Final concentration in reaction
hexRT mix with aRNA (step 21)	6	
5x First strand buffer	2	1 x
DTT (0.1 M)	1	10 mM
RNase OUT (40 U/µl)	0.5	2 U/µl
Superscript II (200 U/µl)	0.5	10 U/µl
Total volume	10	

23. Heat mix in thermocycler at 25 °C for 10 min, then 42 °C for 1 hr with the cycler lid at 50 °C. Let the thermocycler go to 4 °C at the end of the program.

PCR indexing – Timing: 30 min

CRITICAL: PCR amplification must be kept to a minimum to avoid over-amplification. We recommend 6-8 PCR cycles.

24. Prepare the indexing mix as indicated in the table below. Keep mix on ice. Index each sample with a unique RPi primer for multiplexing.

Reagent	Amount (µl)	Final concentration in reaction
Reverse transcribed material (step 23)	10	
Nuclease-free water	11	
2x NEBNext High-Fidelity PCR Master Mix	25	1x
RNA PCR primer RP1 (10 µM)	2	0.4 µM
RNA PCR index primer RPi (10 µM)	2	0.4 µM
Total volume	50	

25. Put mix in thermocycler and run PCR program with the lid heated at 105 °C as indicated in the table below.

Cycle number	Denature	Anneal	Extend
1	98°C, 30 s		
2-11	98°C, 10 s	60°C, 30 s	72°C, 30 s
12			72°C, 10 min

Library purification – Timing: 1 hr 30 min

26. Purify the libraries as stated in **scDam&T Procedure** steps 129-141.
27. Quantify the aRNA by measurement on a Nanodrop spectrophotometer machine and dilute aRNA to 2 ng/µl for quantification on a Bioanalyzer HS DNA chip. Run chip according to the kit manual.

Library quantification and sequencing – Timing: 19 hr

28. Prepare the libraries for sequencing submission as stated in **scDam&T Procedure** steps 142-145.
    PAUSE POINT: The finished libraries can be kept at -20 °C indefinitely.

Supplementary Manual

Molecule after each protocol step – DamID

After ligation of ds DamID adapters (example adapter with barcode 1) 5'→ 3'

```
Fork  T7 promoter        TSS Illumina P5 sequence     UMI barcode UMI barcode
GGTGATCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATCNNNTGCANNNTATGGA-DNA sequence
TTCGAGGGCCATTATGCTGAGTGATATCCCCAAGTCTCAAGATGTCAGGCTGCTAGNNNACGTNNNATACCT-DNA sequence
```

After IVT 5'→ 3'

```
TSS Illumina P5 sequence     UMI barcode UMI barcode
GGGGUUCAGAGUUCUACAGUCCGACGAUCNNNUGCANNNUAUGGA-DNA sequence
```

After hexRT-mediated RT (library prep) 5' → 3'

By using the random hexRT primer, containing a sequence used as in the RA3 sequence (Illumina Truseq Small RNA) and serving as the P7

```
TSS Illumina P5 sequence     UMI barcode UMI barcode           hexRT containing P7
GGGGUUCAGAGUUCUACAGUCCGACGAUCNNNUGCANNNUAUGGA-DNA sequence-
CCCCAAGTCTCAAGATGTCAGGCTGCTAGNNNACGTNNNATACCT-DNA sequence-NNNNNNTGGAATTCTCGGGTGCCAAGGC
```

After PCR (library prep; example RPi index primer 1) 5' → 3'

By using the universal RP1 primer, containing an overlapping sequence with the Illumina P5 sequence in
adapter (Illumina Truseq Small RNA)

By using the indexed RPi primer, containing an overlapping sequence with the P7 sequence of the hexRT, an index, and a sequence to anneal to the flow cell

```
universal RP1 primer       Illumina P5 sequence UMI barcode UMI barcode
ATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGATCNNNUGCANNNUAUGGA-DNA sequence…
TACTATGCCGCTGGTGGCTCTAGATGTGCAAGTCTCAAGATGTCAGGCTAGNNNACGTNNNATACCT-DNA sequence…

          hexRT containing P7 universal RP1 overlapping with P7 sequence (index underlined)
…(continued)DNA sequence-NNNNNNACCTTAAGAGCCCACGGTTCCTTGAGGTCAGTGTAGTGCTAGAGCATACGGCAGAAGACGAAC
…(continued)DNA sequence-NNNNNNTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
```

Molecule after each protocol step – CEL-Seq (as in CEL-Seq2 protocol)

After annealing of CEL-Seq primer to the mRNA molecule (example primer with barcode 1) 5'→ 3'

```
                polyT                barcode UMI barcode UMI Illumina P5 sequence TSS T7 promoter
mRNA sequence-AAAAAAAAAAAAAAAAAAAAAAAA
              VTTTTTTTTTTTTTTTTTTTTTTTTTACTNNNGTAGNNNCTAGCAGCCTGACATCTTGAGGGATATCACTCAGCATAATGGCCG
```

After RT with the CEL-Seq primer 5'→ 3'

```
                polyT                barcode UMI barcode UMI Illumina P5 sequence TSS T7 promoter
mRNA sequence-AAAAAAAAAAAAAAAAAAAAAAAA
mRNA sequence-TTTTTTTTTTTTTTTTTTTTTTTTTACTNNNGTAGNNNCTAGCAGCCTGACATCTTGAGGGATATCACTCAGCATAATGGCCG
```

After second strand synthesis 5' → 3'

```
                polyT                barcode UMI barcode UMI Illumina P5 sequence TSS T7 promoter
mRNA sequence-AAAAAAAAAAAAAAAAAAAAAAAAATGANNNCATCNNNGATCGTCGGACTGTAGAACTCCCTATAGTGAGTCGTATTACCGGC
mRNA sequence-TTTTTTTTTTTTTTTTTTTTTTTTTACTNNNGTAGNNNCTAGCAGCCTGACATCTTGAGGGATATCACTCAGCATAATGGCCG
```

After IVT 5'→ 3'

```
TSS Illumina P5 sequence UMI barcode UMI barcode
GGGAGUUCUACAGUCCGACGAUCNNNGAUGNNNUCAU-UUUUUUUUUUUUUUUUUUUUUUUU-mRNA sequence
```

After hexRT-mediated RT (library prep) 5’ → 3’

By using the random hexRT primer, containing a sequence used as in the RA3 sequence (Illumina Truseq Small RNA) and serving as the P7

```
TSS Illumina P5 sequence UMI barcode UMI barcode                              hexRT containing P7
GGGAGUUCUACAGUCCGACGAUCNNNGAUGNNNUCAU-UUUUUUUUUUUUUUUUUUUUUUUUUU
CCCTCAAGATGTCAGGCTGCTAGNNNCTACNNNAGTA-AAAAAAAAAAAAAAAAAAAAAAAAA-mRNA sequence-NNNNNNTGGAATTCTCGGGTGCCAAGGC
```

After PCR (library prep; example RPi index primer 1) 5’ → 3’

By using the universal RP1 primer, containing an overlapping sequence with the Illumina P5 sequence in t adapter (Illumina Truseq Small RNA)

By using the indexed RPi primer, containing an overlapping sequence with the P7 sequence of the hexRT, an index, and a sequence to anneal to the flow cell

```
RP1 primer                    Illumina P5 sequence UMI barcode UMI barcode
ATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATCNNNGATGNNNTCAT-TTTTTTTTTTTTTTTTTTTTTTTT-mRNA sequence…
TACTATGCCGCTGGTGGCTCTAGATGTGCAAGTCTCAAGATGTCAGGCTGCTAGNNNCTACNNNAGTA-AAAAAAAAAAAAAAAAAAAAAAAA-mRNA sequence…

          hexRT containing P7 universal RP1 overlapping with P7 sequence (index underlined)
…(continued)mRNA sequence-NNNNNNACCTTAAGAGCCCACGGTTCCTTGAGGTCAGTGTAGTGCTAGAGCATACGGCAGAAGACGAAC
…(continued)mRNA sequence-NNNNNNTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
```

CHAPTER 4

Single-cell profiling of transcriptome and histone modifications with EpiDamID

Franka J. Rang[1,2,6], Kim L. de Luca[1,2,6], Sandra S. de Vries[1,2], Christian Valdes-Quezada[1,2], Ellen Boele[1,2], Phong D. Nguyen[1], Isabel Guerreiro[1,2], Yuko Sato[5], Hiroshi Kimura[5], Jeroen Bakkers[1,3], Jop Kind[1,2,4,7]*

1: Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and University Medical Center Utrecht, Utrecht, The Netherlands

2: Oncode Institute, The Netherlands

3: Department of Pediatric Cardiology, Division of Pediatrics, University Medical Center Utrecht, Utrecht, The Netherlands

4: Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, The Netherlands

5: Cell Biology Center, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, 226-8503, Japan

6: These authors contributed equally

7: Lead Contact

*Correspondence: J.K. (j.kind@hubrecht.eu)

Molecular Cell, 2022

Abstract

Recent advances in single-cell sequencing technologies have enabled simultaneous measurement of multiple cellular modalities, but the combined detection of histone post-translational modifications and transcription at single-cell resolution has remained limited. Here, we introduce EpiDamID, an experimental approach to target a diverse set of chromatin types by leveraging the binding specificities of single-chain variable fragment antibodies, engineered chromatin reader domains, and endogenous chromatin-binding proteins. Using these, we render the DamID technology compatible with the genome-wide identification of histone post-translational modifications. Importantly, this includes the possibility to jointly measure chromatin marks and transcriptome at the single-cell level. We use EpiDamID to profile single-cell Polycomb occupancy in mouse embryoid bodies and provide evidence for hierarchical gene regulatory networks. In addition, we map H3K9me3 in early zebrafish embryogenesis, and detect striking heterochromatic regions specific to notochord. Overall, EpiDamID is a new addition to a vast toolbox to study the role of chromatin states during dynamic cellular processes.

Introduction

Histone post-translational modifications (PTMs) contribute to chromatin structure and gene regulation. The addition of PTMs to histone tails can modulate the accessibility of the underlying DNA and form a binding platform for myriad downstream effector proteins. As such, histone PTMs play key roles in a multitude of biological processes, including lineage specification (e.g. refs[1–3]), cell cycle regulation (e.g. refs[4,5]), and response to DNA damage (e.g. refs[6,7]).

Over the past decade, antibody-based DNA-sequencing methods have provided valuable insights into the function of histone PTMs in a variety of biological contexts. Most studies employ ChIP-seq (chromatin immunoprecipitation after formaldehyde fixation[8]), or strategies based on *in situ* enzyme tethering such as chromatin immunocleavage (ChIC)[9], and its derivative Cleavage Under Targets and Release Using Nuclease[10]. However, the requirement of high numbers of input cells consequently provides a population-average view, which disregards the complexity of most biological systems. As a result, several low-input methods have been developed that can assay histone PTMs in individual cells, including but not limited to Drop-ChIP[11], ChIL-seq[12], ACT-seq[13], single-cell ChIP-seq[14], single-cell ChIC-seq[15], single-cell adaptation of CUT&RUN[16], CUT&Tag[17], CoBATCH[18], single-cell itChIP[19], and sortChIC[20]. While these techniques offer an understanding of the epigenetic heterogeneity between cells, they do not provide a direct link to other measurable outputs. Recently, however, three methods have been developed that jointly profile histone modifications and gene expression: Paired-Tag (parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation by sequencing)[21], CoTECH (combined assay of transcriptome and enriched chromatin binding)[22], and SET-seq (same cell epigenome and transcriptome sequencing)[23]. These techniques thus enable linking of gene regulatory mechanisms to transcriptional output and cellular state. Of note, all three methods rely on antibody binding for detection of histone modifications and Tn5-mediated tagmentation for sequencing library preparation. As can be expected from its implementation in ATAC-seq (assay for transposable-accessible chromatin using sequencing)[24], the Tn5 transposase has a high affinity for exposed DNA in open chromatin. While approaches exist to mitigate this bias[25], a recent systematic analysis of Tn5-based studies has provided preliminary indications that accessibility artefacts persist[26].

We recently developed scDam&T-seq, a method that measures DNA-protein contacts and transcription in single cells by combining single-cell DamID and CEL-Seq2[27]. DamID-based techniques attain specificity by tagging a protein of interest (POI) with the *E. coli* Dam methyltransferase, which methylates adenines in a GATC motif in the proximity of the POI[28–30]. The approach is especially suited for single-cell studies, because DNA-protein contacts are recorded directly on the DNA in the living cell, and downstream sample handling is limited. However, Dam cannot be tethered directly to post-translationally modified proteins by genetic engineering, which has precluded the use of DamID for studying histone PTMs.

Here, we present EpiDamID, an extension of existing DamID protocols, based on the fusion of Dam to chromatin-binding modules for the detection of various types of histone PTMs. We validate the specificity of EpiDamID in population (Fig. 1) and single-cell samples (Fig. 2). Subsequently, we leverage its single-cell resolution to study the Polycomb mark H3K27me3 and its relationship to transcription in mouse embryoid bodies (EBs) (Fig. 3), and identify distinct Polycomb-regulated and Polycomb-independent hierarchical TF networks (Fig. 4). Finally, we implement a protocol to assay cell type-specific patterns of the heterochromatic mark H3K9me3 in the zebrafish embryo and discover broad domains of heterochromatin specific to the notochord (Fig. 5). Together, these results show that EpiDamID provides a versatile tool that can be implemented in diverse biological settings to obtain single-cell histone PTM profiles.

Design

The conventional DamID approach involves genetically engineering a protein of interest (POI) to the bacterial methyltransferase Dam (Fig. 1a). In this study, we adapted the DamID method to detect histone PTMs by fusing Dam to one of the following: 1) full-length chromatin proteins, 2) tuples of well-characterized reader domains[31–33], or 3) single-chain variable fragments (scFv) also known as mintbodies[34–36] (Fig. 1a, Methods). Similar strategies have been successfully applied in microscopy, proteomics and ChIP experiments[34–38]. Our approach is henceforth referred to as EpiDamID, and the construct fused to Dam as the targeting domain. Since this approach can be applied to any existing DamID method, EpiDamID makes all these protocols available to the study of chromatin modifications. This includes the possibility to perform (live) imaging of Dam-methylated DNA[39–41], tissue-specific study of model organisms without cell isolation via Targeted DamID (TaDa)[42], DamID-directed proteomics[43], (multi-modal) single-cell[27,39,40,44] and single-molecule[45] sequencing studies, and the processing of samples with little material[40,46].

Results

Targeting domains specific to histone modifications mark distinct chromatin types with EpiDamID

We categorized the various targeting domains into the following chromatin types: accessible, active, heterochromatin, and Polycomb. We generated various expression constructs for each of the different targeting domains, testing promoters (HSP, PGK), orientations (Dam-POI, POI-Dam) and two versions of the Dam protein (DamWT, Dam126) (Table S1). The choice of promoter influences the expression level of the Dam-POI, whereas the orientation may affect target binding. In the Dam126 mutant, the N126A substitution diminishes off-target methylation[47,48]. We introduced the Dam constructs by viral transduction in hTERT-immortalized RPE-1 cells and performed DamID2 followed by high-throughput sequencing[49]. To validate our data with an orthogonal method, we generated ChIP-seq samples for various histone modifications.

The DamID samples were filtered on sequencing depth and information content (IC), a metric for determining signal-to-noise levels (Fig. S1a-b) (Methods). IC additionally showed that tuples of reader domains fused to Dam typically perform better than single domains ($p < 0.05$ for three out of four domains, Fig. S1b), in agreement with a recent study employing similar domains for proteomics purposes[38] (Fig. S1a-b). Therefore, only data from the triple reader domains were included in further analyses.

Visualization of all filtered samples by uniform manifold approximation and projection (UMAP) shows that EpiDamID mapping identifies distinct chromatin types and that samples consistently group with their corresponding ChIP-seq datasets (Fig. 1b). Genome-wide DamID signal also correlates well with ChIP-seq signal (mean Pearson's correlation coefficients from 0.40-0.64 for active marks, 0.58-0.61 for heterochromatin marks, and 0.56-0.60 for Polycomb marks) (Fig. 1c and S1c). Importantly, DamID samples do not group based on construct type, promoter, Dam type, sequencing depth, or IC (Fig. S1d-e), indicating that those properties do not influence target specificity. All targets display the expected patterns of enrichment along the linear genome (Fig. 1d, left), as well as genome-wide on-target signal (Fig. 1d, right). To further explore the specificity of constructs that target active chromatin, we compared signal of Dam-H3K9ac and Dam-TAF3 at H3K9ac ChIP-seq peaks with high and low H3K4me3 ChIP-seq levels. Dam-H3K9ac shows enrichment in both categories, while Dam-TAF3 is enriched specifically in the high-H3K4me3 category (Fig S1F). This confirms that, while the untethered Dam protein preferentially marks accessible chromatin, targeting it to active regions of the genome yields specific methylation patterns.

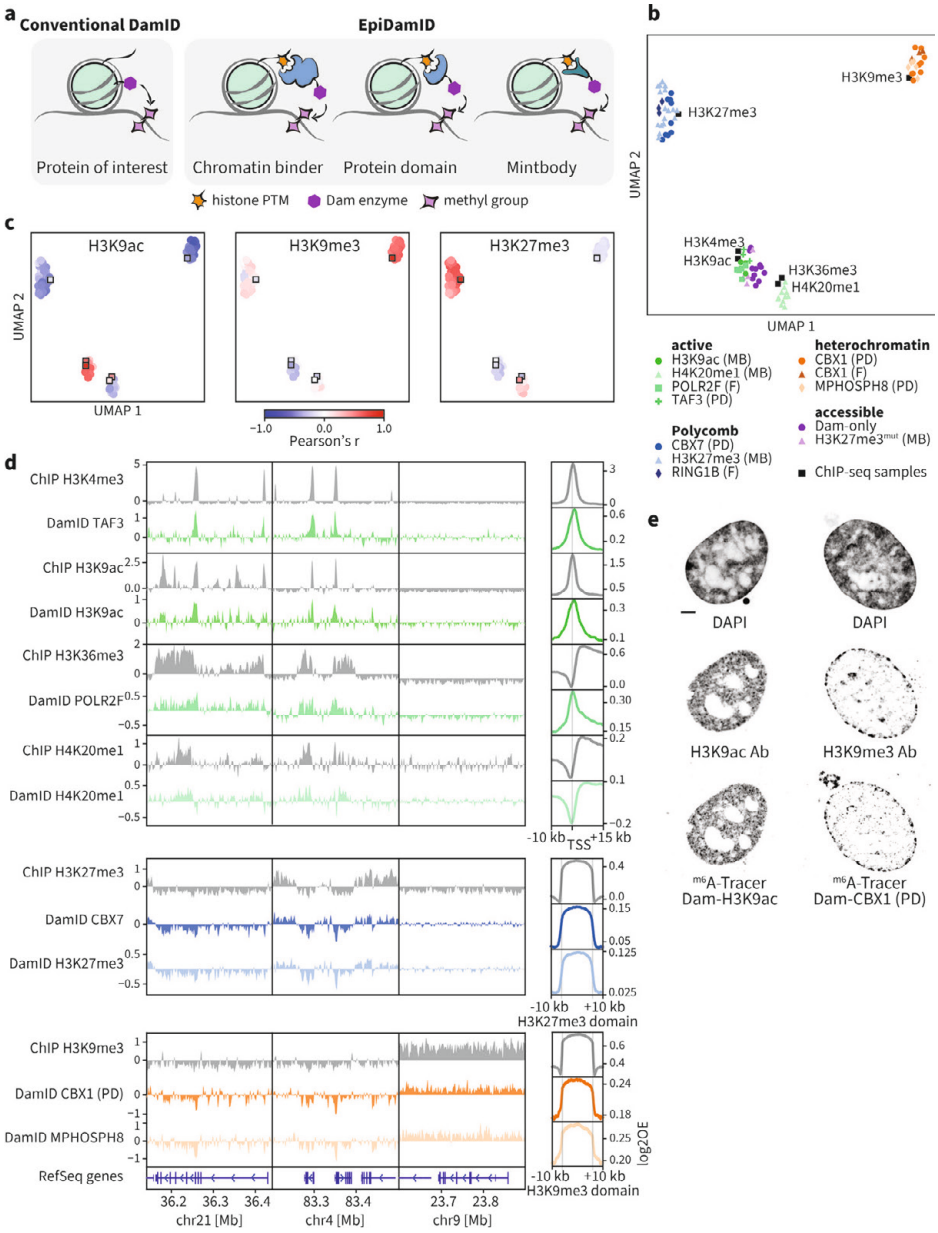


Figure 1: Targeting domains specific to histone modifications mark distinct chromatin types with EpiDamID

a, Schematic overview of EpiDamID concept compared to conventional DamID. **b**, UMAP of DamID samples colored by targeting construct, and ChIP-seq samples of corresponding histone modifications. MB: mintbody; PD: protein domain; F: full protein. **c**, UMAPs as in **b**, colored by correlation with selected ChIP-seq samples (H3K9ac, H3K9me3, and H3K27me3). Correlation values reflect the Pearson's correlation coefficient of Dam-normalized samples with the indicated ChIP-seq sample. Control constructs (Dam, H3K27me3^{mut}) are excluded from the UMAP. DamID samples are circles; ChIP-seq samples are squares. **d**,

Left: three genome browser views of ChIP-seq (gray) and DamID (colored) enrichment. Data represent the combined signal of all samples of each targeting domain. Right: average DamID and ChIP-seq enrichment plots over genomic regions of interest. Signal is normalized for untethered Dam or input, respectively. Regions are the TSS (-10/+15 kb) of the top 25% H3K9ac-enriched genes for the active marks (top), and ChIP-seq domains (+/- 10 kb) for H3K27me3 (middle), and H3K9me3 (bottom). **e**, Confocal images of nuclear chromatin showing DAPI (top), immunofluorescent staining against an endogenous histone modification (middle), and its corresponding EpiDamID construct visualized with ^{m6}A-Tracer (bottom). Left: H3K9ac, right: H3K9me3. Scale bar: 3 µm.

Next, we quantified the spreading of Dam signal from its binding location to determine the resolution for all chromatin types. We found that DamID signal decays to 50% (from 100% at peak center or domain border) across a distance that extends ~1 kb past the ChIP-seq 50% decay point (Fig. S1g), implying a resolution of ~1-2 kb, similar to earlier studies with transcription factors[50,51]. It was previously reported that the Dam126 mutant improves signal quality compared to DamWT[47]. Indeed, this mutant markedly improved sensitivity and reduced background methylation (mean IC increase of 0.07-0.21 per construct) (Fig. S1h-i).

We further validated the correct nuclear localization of Dam-marked chromatin with microscopy, by immunofluorescent staining of endogenous histone PTMs and DamID visualization using ^{m6}A-Tracer protein[41,52] (Fig. 1e).

Together, these results show that EpiDamID specifically targets histone PTMs and enables identification of their genomic distributions by next-generation sequencing.

Detection of histone PTMs in single mouse embryonic stem cells with EpiDamID

We next established EpiDamID for single-cell sequencing. To this end, we generated clonal, inducible mESC lines for the following targeting domains fused to Dam: H4K20me1 mintbody, H3K27me3 mintbody, and the H3K27me3-specific CBX7 protein domain (3x tuple). While H4K20me1 is enriched over the gene body of active genes[53], the heterochromatic mark H3K27me3 is enriched over the promoter of developmentally regulated genes[54,55]. As controls, we included an H3K27me3^{mut} mintbody construct whose antigen-binding ability is abrogated by a point mutation in the third complementarity determining region of the heavy chain (Y105F), and a published mESC line expressing untethered Dam[27]. We performed scDam&T-seq to generate 442-1,402 single-cell samples per construct, retaining 283-855 samples after filtering on the number of unique GATCs and IC (10,417-45,067 median unique counts per construct and median IC of 2.0-2.9) (Fig. S2a-c, Table S2). For subsequent analyses, we also included a published dataset of Dam fused to RING1B [27] as an example of a full-length chromatin reader targeting Polycomb chromatin. All constructs contained DamWT, as the Dam126 methylation levels were found insufficient to produce high-quality single-cell signal (data not shown).

Dimensionality reduction of the single-cell datasets revealed that the samples primarily separated on chromatin type (Fig. 2a). To further confirm the specificity of the constructs, we used mESC H3K27me3 (ENCSR059MBO) and H3K9ac (ENCSR000CGP) ChIP-seq datasets from the ENCODE portal[56] and generated our own for H4K20me1. For all single cells, we computed

the enrichment of counts within H3K27me3, H3K9ac and H4K20me1 ChIP-seq domains. These results show a strong enrichment of EpiDamID counts within domains for the corresponding histone PTMs (Fig. 2b-d, Fig. S2d), indicating that the methylation patterns are specific for their respective chromatin targets, even at the single-cell level. The combined single-cell data also showed the expected enrichment over H3K27me3 ChIP-seq domains (Fig. 2e) and active gene bodies (Fig. 2f) for the Polycomb-targeting constructs and H4K20me1, respectively. Contrary to the H3K27me3 construct, H3K27me3^{mut} showed little enrichment over H3K27me3 ChIP-seq domains (Fig. S2e). The specificity of the signal is also evident at individual loci in both the in silico populations and single cells (Fig. 2g-h and Fig. S2f).

These results demonstrate that mintbodies and protein domains can be used to map histone PTMs in single cells with EpiDamID.

Joint profiling of Polycomb chromatin and gene expression in mouse embryoid bodies

To exploit the benefits of simultaneously measuring histone PTMs and transcriptome, we profiled Polycomb chromatin in mouse EBs. We targeted the two main Polycomb repressive complexes (PRC) with EpiDamID using the full-length protein RING1B and H3K27me3-mintbody fused to Dam. RING1B is a core PRC1 protein that mediates H2AK119 ubiquitylation[57,58], and H3K27me3 is the histone PTM deposited by PRC2[59–62]. Both PRC1 and PRC2 have key roles in gene regulation during stem cell differentiation and early embryonic development (see refs[63,64] for recent reviews on this topic).

To assay a diversity of cell types at various stages of differentiation, we harvested EBs for scDam&T-seq at day 7, 10 and 14 post aggregation, next to ESCs grown in 2i/LIF (Fig. 3a). We used Hoechst incorporation in combination with fluorescence-activated cell sorting (FACS) to deposit live, single cells into 384-well plates and record their corresponding cell cycle phase (Methods). In addition to RING1B and H3K27me3-mintbody, we included the untethered Dam protein for all time points as a control for chromatin accessibility. Collectively, we obtained 2,943 cells after filtering (Fig. S3a-b), in a similar range as CoTECH (~7,000 cells), higher than SET-seq (~500 cells) and lower than Paired-Tag (~65,000 nuclei). The number of unique genomic and transcriptomic counts per cell was similar or higher compared to the other methods (Fig. S3a-b). Based on the transcriptional readout, we identified eight distinct clusters across time points (Fig. 3b). We integrated the EB transcriptome data with the publicly available mouse embryo atlas[65] to confirm the correspondence of cell types with early mouse development and guide cluster annotations (Fig. S3c-d). This indicated the presence of pluripotent and more differentiated cellular states, including epiblast, endoderm, and mesoderm lineages. Notably, the DamID readout alone was sufficient to consistently separate cells on chromatin type (Fig. 3c) and to distinguish between the pluripotent and more lineage-committed cells (Fig. 3d-e). Thus, the EpiDamID profiles display cell type-specific patterns of chromatin accessibility and Polycomb association. Prompted by this observation, we trained a linear discriminant analysis (LDA) classifier to assign an additional 1,543 cells with poor transcriptional data to cell type clusters, based on their DamID signal (Fig. S3e, Table S2).

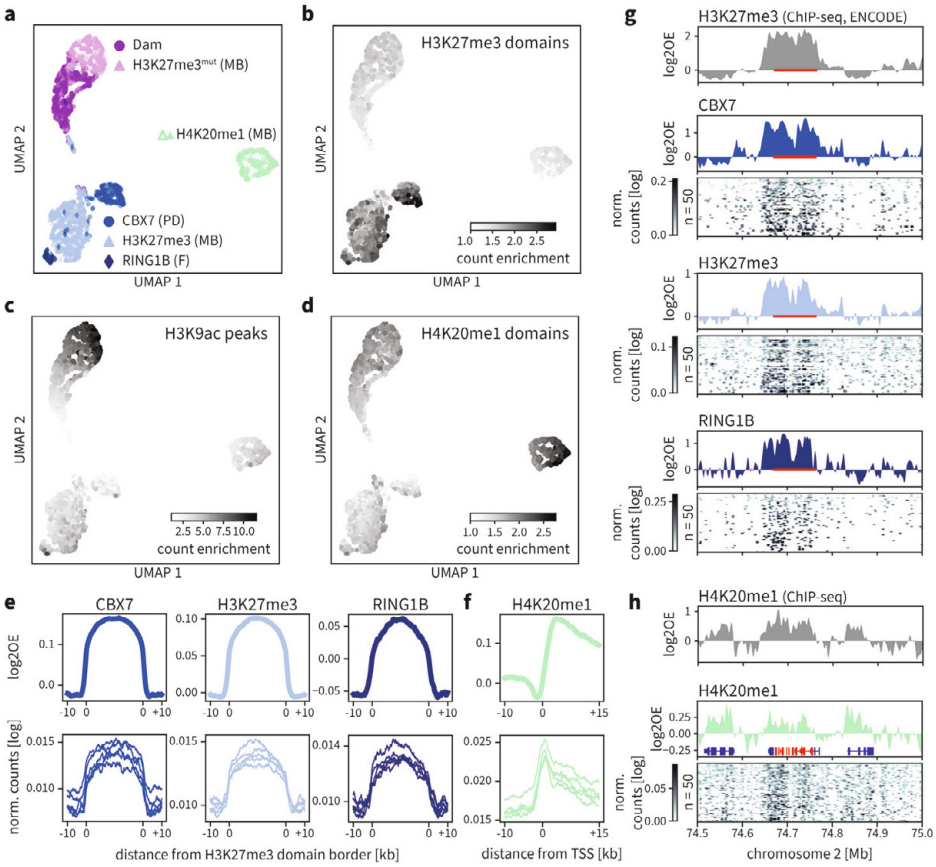


Figure 2: Detection of histone PTMs in single mouse embryonic stem cells with EpiDamID
a, UMAP based on the single-cell DamID readout of all single-cell samples. MB: mintbody; PD: protein domain; F: full protein. **b-d**, DamID UMAP as in **a**, colored by the enrichment of counts within H3K27me3 ChIP-seq domains (**b**), H3K9ac ChIP-seq peaks (**c**), and H4K20me1 ChIP-seq domains (**d**). **e**, Average signal over H3K27me3 ChIP-seq domains of CBX7 and H3K27me3 targeting domains and full-length RINGB1B protein. **f**, Average H4K20me1 signal over the TSS of the top 25% active genes (based on H3K9ac ChIP-seq signal). **e-f**, Top: in silico populations normalized for Dam; Bottom: five of the best single-cell samples (bottom) normalized only by read depth. **g-h**, Signal of various marks over the HoxD cluster and neighboring regions. ChIP-seq data is normalized for input control. DamID tracks show the Dam-normalized in silico populations of the various Dam-fusion proteins, DamID heatmaps show the depth-normalized single-cell data of the fifty richest cells. The HoxD cluster is indicated in red in **g** (bar) and **h** (RefSeq); additional RefSeq genes are shown **h.**

Next, we defined the set of genes that is Polycomb-regulated in the EB system. First, we determined the H3K27me3 and RING1B signal at the promoter region of all genes and compared these two readouts across the clusters. This confirmed good correspondence between H3K27me3 and RING1B profiles (Pearson's r = 0.60-0.82, p = 0 between profiles of the same cluster) (Fig. S3f-g), albeit with a slightly higher signal amplitude for RING1B (Fig. S3g). This difference between RING1B and H3K27me3 may be biological (e.g., differential binding

sites or kinetics) and/or technical (e.g., the use of a full-length protein versus a mintbody to target Dam). Nonetheless, because of the overall similarity, we decided to classify high-confidence Polycomb targets as having both H3K27me3 and RING1B enrichment in at least one of the EB clusters (excluding cluster 7 due to the relatively low number of cells) or in the previous ESC data set. We identified 9,159 Polycomb-regulated targets across the dataset, in good concordance with previous work in mouse development (4,059 overlapping genes out of a total of 5,986; $p = 9.5 \times 10^{-135}$, Chi-square test)[66] (Fig. S3h).

Next, we intersected the cluster-specific transcriptome and DamID data to relate gene expression patterns to Polycomb associations. Based on the role of Polycomb in gene silencing, differential binding of PRC1/2 to genes is expected to be associated with changes in expression levels. As exemplified in Fig. 3f-g, the cell type-specific expression of *Tal1*, a master regulator in hematopoiesis, is indeed inversely related to Polycomb enrichment. This negative association is apparent for all PRC targets that are upregulated in the hematopoietic cluster (Fig. S3i-j). In addition, unsupervised clustering of H3K27me3 and RING1B promoter occupancy shows variation in signal between target genes as well as between cell types, indicating dynamic regulation of these targets in EBs (Fig. 3h). In line with this, Polycomb targets with variable PRC occupancy are typically more highly expressed in those clusters where Polycomb is absent (Wilcoxon's signed-rank test, $p = 2.6 \times 10^{-185}$, Fig. 3i). Since the negative relationship between Polycomb occupancy and transcription is not perfect, we were interested to see whether an additional layer of epigenetic regulation could further explain the observed transcriptional changes. To this end, we integrated our data with a publicly available scNMT-seq dataset[67], also generated in EBs (Fig. S3k). This resulted in sufficient scNMT-seq samples in four clusters to compare CpG methylation profiles with Polycomb occupancy. The integrated profiles indeed revealed a complementary relationship between the two marks, where genes with either CpG methylation or Polycomb at their promoter tend to be expressed at lower levels (Fig. S3l). This was also apparent for CpG methylation and expression of genes with variable Polycomb enrichment between the clusters (Fig. S3m). The observed trends are in line with the known repressive effects of both marks and their largely mutually exclusive localizations[68–70].

Collectively, these data illustrate the strength of EpiDamID to jointly capture transcription and chromatin dynamics during differentiation, as well as the potential to integrate the results with datasets derived from different techniques.

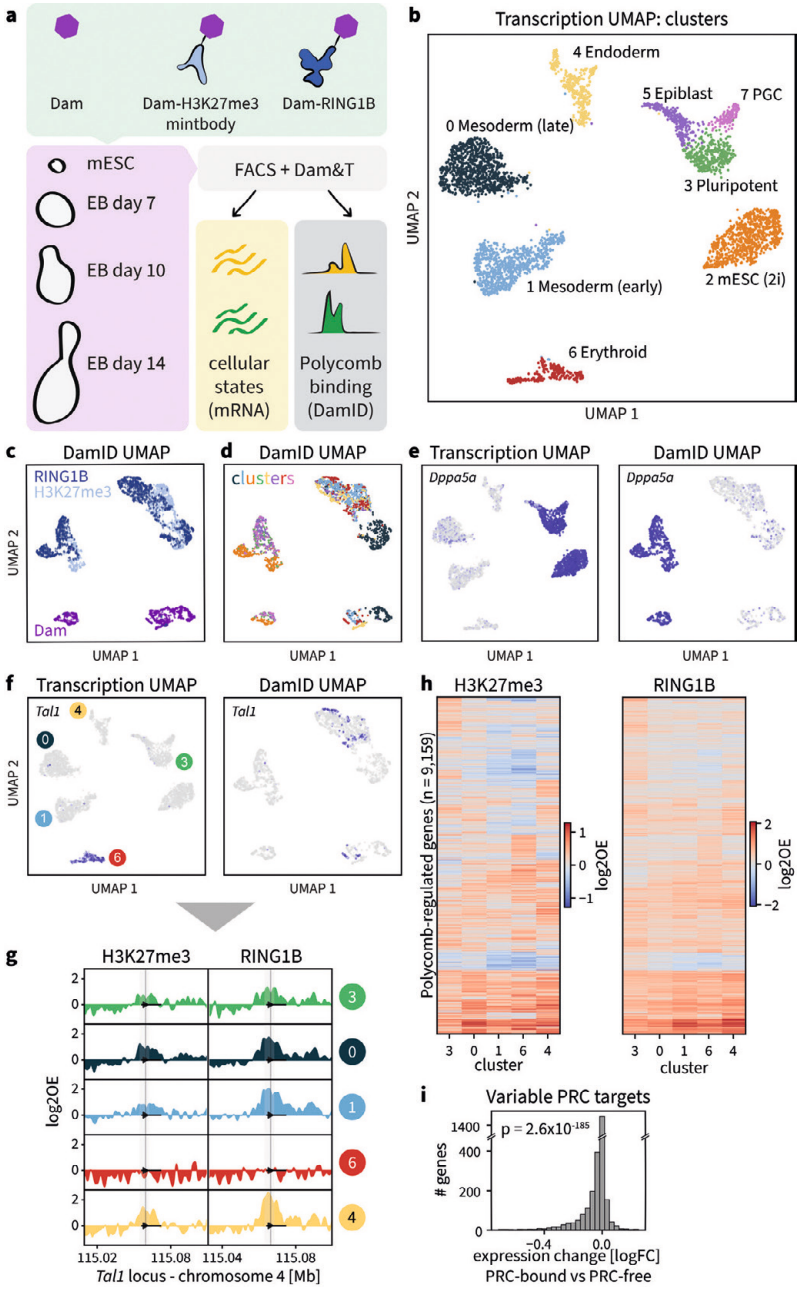


Figure 3: Joint profiling of Polycomb chromatin and gene expression in mouse embryoid bodies
a, Schematic showing the experimental design. **b**, UMAP of samples based on transcriptional readout, colored by cluster. **c-d**, UMAP of samples based on DamID readout, colored by construct (**c**) and cluster (**d**). **e**, Transcriptomic UMAP (left) and DamID UMAP (right), colored by expression of pluripotency marker Dppa5a. **f**, Transcriptomic UMAP (left) and DamID UMAP (right), colored by expression of hematopoietic regulator Tal1. **g**, Genomic tracks of H3K27me3 and RING1B DamID signal per cluster at the Tal1 locus.

h, Heatmaps showing the H3K27me3 (left) and RING1B (right) DamID signal of all identified PRC targets for transcriptional clusters 3, 0, 1, 6, and 4. PRC targets are ordered based on hierarchical clustering. **i**, Fold-change in expression of Polycomb targets between clusters where the gene is PRC-associated and clusters where the gene is PRC-free. The significance was tested with a two-sided Wilcoxon's signed rank test ($p = 2.6 \times 10^{-185}$).

Polycomb-regulated transcription factors form separate regulatory networks

We next focused on the Polycomb targets based on their function, and found that TF genes are over-represented within the Polycomb target genes (Fig. S4a), in line with previous observations[54]. Nearly half of all TF genes in the genome (761/1689) is bound by Polycomb in at least one cluster. In addition, genes encoding TFs generally accumulate higher levels of H3K27me3 and RING1B compared to other protein-coding genes (Fig. S4b). Consistent with an important role in lineage specification, Polycomb-controlled TFs are expressed in a cell type-specific pattern, as opposed to the more constitutive expression across cell types for Polycomb-independent TFs (Fig. S4c-d). Accordingly, the Polycomb-controlled TFs are enriched for Gene Ontology (GO) terms associated with animal development (Fig. S4e).

The high Polycomb occupancy at developmentally regulated TF genes prompted further investigation into the role of Polycomb in TF network hierarchies. We used SCENIC to systematically identify target genes that are associated with the expression of TFs[71,72]. SCENIC employs co-expression patterns and binding motifs to link TFs to their targets, together henceforth termed "regulons" (per SCENIC nomenclature). We identified 285 "activating" regulons after filtering (Fig. 4a and Methods). While regulons and their activity were found independently of RNA-based cluster annotations, regulon activity trends clearly matched the annotated clusters (Fig. 4a).

We first determined how overall regulon activity identified by SCENIC correlates to Polycomb binding. As illustrated for the homeobox TF gene *Msx1*, we found that regulon activity is generally inversely related to Polycomb association of both the TF gene (red dot) and its Polycomb-controlled targets (boxplots, 65% of all MSX1 targets) (Fig. 4b-c). We wondered whether there is a general preference for Polycomb-controlled TFs to target genes that themselves are regulated by Polycomb. Indeed: while Polycomb-controlled TFs have a similar number of target genes compared to other TFs (Fig. S4f), the expression of the targets is much more frequently controlled by Polycomb than expected by chance (Mann-Whitney-U test $p = 2.8 \times 10^{-20}$, Fig. 4d). This effect is even stronger when considering the subset of targets that is exclusively regulated by Polycomb TFs (Chi-square test $p = 0$, Fig. S4g). Similarly, upstream TFs controlling the regulon TFs (Fig. 4e) also tend to be Polycomb-controlled (Mann-Whitney-U test, $p = 6.6 \times 10^{-19}$, Fig. 4f). Moreover, the fractions of Polycomb-controlled upstream regulators and downstream targets are correlated (Pearson's $r = 0.61$, $p = 2.9 \times 10^{-29}$, Fig. 4g), indicating consistency in the level of Polycomb regulation across at least three layers of the TF network. This trend is especially strong for the lineage-specific genes (Pearson's $r = 0.48$, $p = 9.2 \times 10^{-8}$), but also holds for other, unspecific, genes (Pearson's $r = 0.41$, $p = 4.0 \times 10^{-4}$) (Fig. S4h-i). These results suggest that Polycomb-associated hierarchies exist, forming relatively separate

networks isolated from other gene regulatory mechanisms, and that this phenomenon extends beyond lineage-specific genes alone.

Together, the above findings demonstrate that single-cell EpiDamID can be successfully applied in complex developmental systems to gather detailed information on cell type-specific Polycomb regulation and its interaction with transcriptional networks.

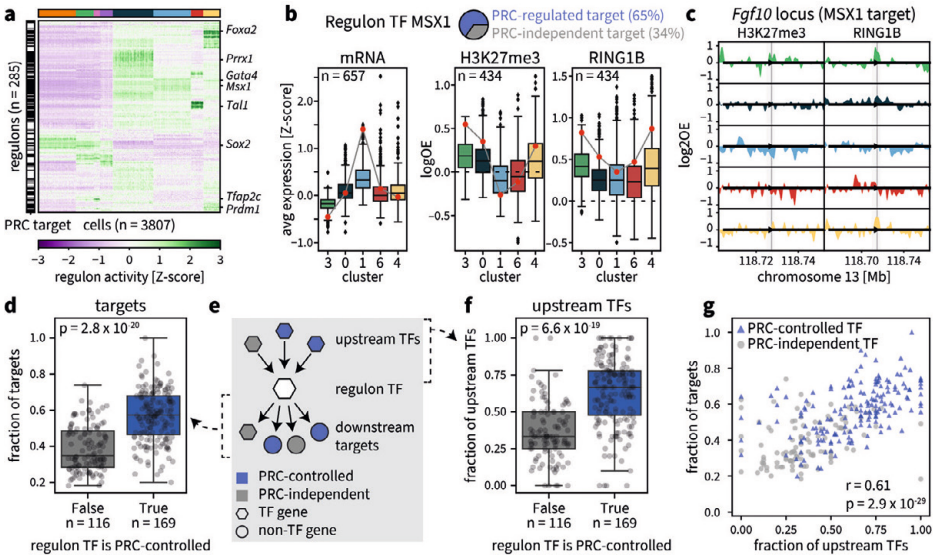


Figure 4: Polycomb-regulated transcription factors form separate regulatory networks
a, Heatmap showing SCENIC regulon activity per single cell. Cells (columns) are ordered by transcriptional cluster; regulon (rows) are ordered by hierarchical clustering. The black and white bar on the left indicates whether the regulon TF is a PRC target (black) or not (white). **b**, Example of the relationship between expression and Polycomb regulation for the MSX1 regulon. Pie chart indicates the percentages of Polycomb-controlled (blue) or Polycomb-independent (grey) target genes. Left: boxplots showing target gene expression per cluster for all target genes. Middle and right: boxplots showing the H3K27me3 and RING1B DamID signal at the TSS per cluster for the Polycomb-controlled target genes. The expression and DamID signal of Msx1 is indicated with a red circle. **c**, Genomic tracks of H3K27me3 and RING1B DamID signal per cluster at the Fgf10 locus, one of the target genes of MSX1. Arrowhead indicates the location of the TSS; shaded area indicates -5kb/+3kb around the TSS. **d**, Boxplots showing the fraction of Polycomb-controlled target genes, split by whether the TF itself is Polycomb-controlled. The significance was tested with a two-sided Mann-Whitney U test ($p = 2.8 \times 10^{-20}$). **e**, Schematic of the regulatory network, indicating the relationship between a regulon TF (white hexagon), its upstream regulators (colored hexagons), and its downstream targets (colored hexagons/circles). **f**, Boxplots showing the fraction of Polycomb-controlled upstream regulators, split by whether the regulon TF is Polycomb-controlled. The significance was tested with a two-sided Mann-Whitney U test ($p = 6.6 \times 10^{-19}$). **g**, Scatter plot showing the relationship between the fraction of Polycomb-controlled targets and regulators of a regulon TF. Regulon TFs that are PRC controlled are indicated in blue; regulon TFs that are PRC independent are indicated in grey. Correlation was computed using Pearson's correlation ($p = 2.9 \times 10^{-29}$).

Implementation of EpiDamID during zebrafish embryogenesis

Next, we applied EpiDamID in an in vivo system to study the heterochromatic mark H3K9me3 during zebrafish development. To bypass the need for genetic engineering, we employed microinjection of mRNA into the zygote (Fig. 5a), a strategy successfully applied in the mouse embryo[40]. H3K9me3 is reprogrammed during the early stages of development in several species[73–77] and the deposition of this mark coincides with decreased developmental potential[78]. It was previously shown that H3K9me3 is largely absent before the maternal-to-zygotic transition (MZT)[74], but it remains unclear whether the H3K9me3 distribution undergoes further remodeling after this stage, and whether its establishment differs across cell types during development.

We injected mRNA encoding the H3K9me3-specific construct *Dam-Mphosph8* and untethered *Dam* into the yolk at the one-cell stage and collected embryos at the 15-somite stage (Fig 5a), which comprises a wide diversity of cell types corresponding to all germ layers. We generated 2,127 single-cell samples passing both DamID and CEL-Seq2 thresholds (Fig. S5a, Supplementary Table 2). Comparing the DamID data of an in silico whole-embryo sample to published H3K9me3 ChIP-seq data of 6-hpf embryos[74] showed good concordance (Pearson's $r = 0.72$, $p = 0$; Fig. S5b).

Broad domains of notochord-specific H3K9me3 enrichment revealed by scDam&T-seq

Analysis of the single-cell transcriptome data resulted in 22 clusters of diverse cell types (Fig. 5b), which we annotated according to expression of known marker genes (Fig. S5c). After dimensionality reduction based on the DamID signal, we observed a clear visual separation of cells in accordance with their Dam construct, and to a lesser extent with their cell type (Fig. 5c-d). Cluster-specific DamID profiles allowed us to employ the LDA classifier to assign a further 705 cells with poor transcriptional readout to a cluster (Fig. S5d, Table S2). Notably, the MPHOSPH8 samples of hatching gland (cluster 1, *he1.1* expression) and notochord (cluster 2, *col9a2* expression) segregated strongly from the other cell types (Fig. 5d), implying differences in their single-cell H3K9me3 profiles. In particular, we observed the appearance of large domains of H3K9me3 enrichment in the notochord, and seemingly lower levels of H3K9me3 in the hatching gland (Fig. 5e and Fig. S5e).

Next, to more systematically identify and characterize regions of differential H3K9me3 enrichment between cell clusters, we performed ChromHMM [79,80]. The approach uses the H3K9me3 signal per cluster to annotate genomic segments as belonging to different H3K9me3 states. We included the 12 cell clusters containing >30 cells per construct and identified five H3K9me3 states across the genome. These represented: A) three states of constitutive H3K9me3 with different enrichment levels [A1-A3], B) notochord-specific H3K9me3 enrichment, and C) constitutive depletion of H3K9me3 (Fig. 5f-g). While all 12 clusters had the highest H3K9me3 enrichment in state A1, cells belonging to the hatching gland (cluster 1) tended to have lower signal in these regions compared to other cell types (Fig. S5f). Notochord cells (cluster 2), conversely, displayed somewhat higher enrichment in

state A1 and dramatically higher enrichment in state B compared to the others. State A (A1-3) chromatin forms broad domains (Fig. S5g) that together comprise 27% of the genome (Fig. S5h) and, as expected for H3K9me3-associated chromatin regions, are characterized by sparser gene density and lower gene activity compared to the H3K9me3-depleted state C (Fig. 5h). Moreover, state A1 is strongly enriched for zinc-finger transcription factors (Fig. S5i), which are known to be demarcated by H3K9me3 in other species [81,82]. The notochord-specific state B has similar characteristics to states A1-A3 (Fig. 5h, S5g-i), yet exhibits broader consecutive regions of H3K9me3 enrichment (Fig. 5g and S5g) and an even lower active gene density (Fig. 5h). However, we did not find a notable increase in H3K9me3 at genes downregulated in notochord (Fig. S5j), implying that these domains do not play a role in gene expression regulation.

One of the known functions of H3K9me3 chromatin is the repression of transposable elements[83–85]. Indeed, it was previously observed in zebrafish that nearly all H3K9me3 domains in early embryos are associated with repeats[74]. We determined whether distinct repeat classes were over-represented in each H3K9me3 ChromHMM state (Fig. S6a) and found a strong enrichment of several repeat classes in state A1, including LTR and tRNA. Further discrimination within the classes showed a high frequency of pericentromeric satellite repeats SAT-1 and BRSATI in state A1 (Fig. 5i), in line with the known occupancy of H3K9me3 at pericentromeric regions. Inspection of the DamID patterns showed a clear increase of signal centered on specific repeat regions in state A1, and to lesser extents in other states (Fig. S6b). In addition, we found that state B harbors specific enrichment of certain repeats (Fig. 5i and Fig. S6c), although further study is required to determine whether H3K9me3 is involved in cell type-specific repression of repetitive genomic regions in the notochord.

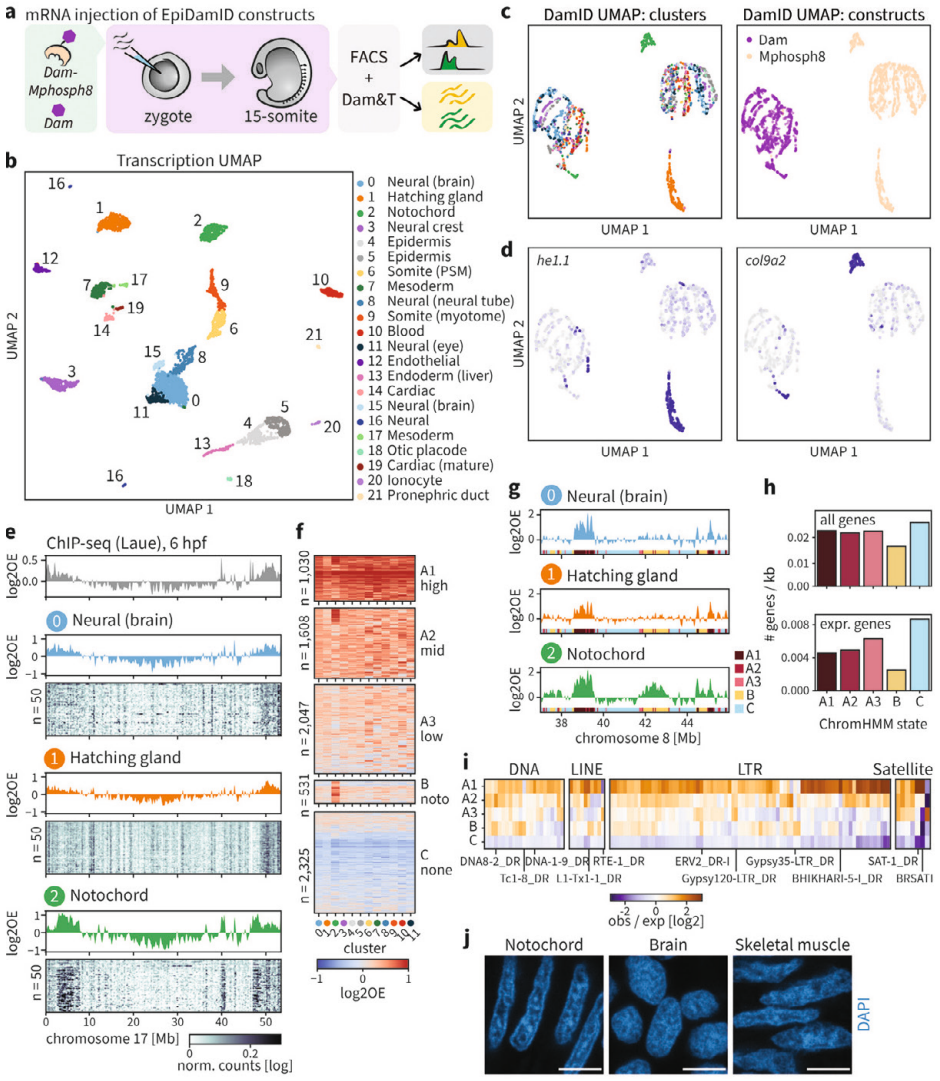


Figure 5: Notochord-specific H3K9me3 enrichment in the zebrafish embryo

a, Schematic representation of the experimental design and workflow. **b**, UMAP based on the transcriptional readout of all single-cell samples passing CEL-Seq2 thresholds (n = 3902). **c**, UMAP based on the genomic readout of all single-cell samples passing DamID thresholds (n = 2833). Samples are colored by transcriptional cluster (left) and Dam-targeting domain (right). **d**, Expression of the hatching gland marker he1.1 (left) and the notochord marker col9a2 (right) projected onto the DamID UMAP. **e**, Genomic H3K9me3 signal over chromosome 17. Top track: H3K9me3 ChIP-seq signal of 6-hpf embryo. Remaining tracks: combined single-cell Dam-MPHOSPH8 data for clusters 0-2. Heatmaps show the depth-normalized Dam-MPHOSPH8 data of the 50 richest cells. **f**, Heatmap showing the cluster-specific average H3K9me3 enrichment over all domains called per ChromHMM state. Per state, domains were clustered using hierarchical clustering. **g**, Genomic H3K9me3 signal over a part of chromosome 8 for clusters 0-2. The colored regions at the bottom of each track indicate the ChromHMM state. **h**, Gene density of all genes (top) and expressed genes (bottom) per state. **i**, Enrichment of repeats among the ChromHMM states. Example repeats are indicated. **j**, Representative images of DAPI staining in cryosections of zebrafish embryos at 15-somite stage. Scale bars represent 4 μm.

Altered expression of chromatin proteins and pronounced nuclear compartmentalization in notochord

Finally, we evaluated cluster-specific expression of known chromatin proteins in relation to the differential H3K9me3 patterns. Expression levels of histone methyltransferases, demethylases and other chromatin factors did not show an upregulation of known H3K9 methyltransferases (*setdb2*, *setdb1a/b*, *suv39h11a/b*, *ehmt2*) nor demethylases (*kdm4aa/ab/b/c*, *phf8*) in notochord (Fig. S6d). However, the H3K9- and H3K36-specific demethylase *kdm4c* was exclusively upregulated in hatching gland, which could explain the low H3K9me3 levels in this cluster. Notably, the notochord cluster showed significant upregulation of *lmna*, the gene encoding nuclear lamina protein Lamin A/C that associates with heterochromatin[86] and plays an important structural role in the nucleus[86,87]. This could be relevant in relation to the structural role of the notochord and the resulting mechanical forces the cells are subjected to[88]. To more directly investigate chromatin state and nuclear organization in these embryos, we performed confocal imaging of H3K9me3 and DAPI stainings in notochord, brain, and skeletal muscle. H3K9me3-marked chromatin displayed a typical nuclear distribution in all tissues, including heterochromatin foci as previously reported[74] (Fig. S6e). DAPI staining showed more structure in the notochord compared to the other tissues (Fig. 5j), visible as a clear rim along the nuclear periphery and denser foci within the nuclear interior. This indicates a stronger separation between euchromatin and heterochromatin, although it remains to be elucidated whether these features are related to the notochord-specific H3K9me3 domains in the genome.

The implementation of EpiDamID in zebrafish embryos shows that this strategy provides a flexible and accessible approach to generate high-resolution single-cell information on the epigenetic states that underlie biological processes during organismal development.

Discussion

Advantages of DamID for single-cell multi-modal omics during embryo development

The DamID workflow involves few enzymatic steps and is thus especially suitable for integration with other single-cell protocols to achieve multi-modal measurements[49]. Minimal sample handling prior to molecular processing results in a high recovery rate of collected cells[40]; for example, scDam&T-seq with EpiDamID constructs could be used to individually assay all cells of a single preimplantation mouse embryo and examine epigenetic and transcriptomic differences that may point towards cell fate commitment, while tracking intra-embryonic variability. Further, DamID genomic marks are stable upon deposition, offering the possibility to track ancestral EpiDamID signatures through mitosis to study inheritance and spatial distribution of epigenetic states in daughter cells[41,89].

Comparison to other single-cell transcriptome and chromatin profiling techniques

In the past year, three other techniques have been published that are capable of simultaneously measuring chromatin modifications and transcription: Paired-Tag[21], CoTECH[22], and SET-seq[23]. One major conceptual difference between above methods and DamID-based techniques is the

manner of capturing DNA in proximity of the chromatin mark of interest. Strategies leveraging CUT&Tag obtain a readout of chromatin by targeting protein A fused to transposase Tn5 (pA-Tn5) to antibody-bound regions, and integrating barcoded adapters into the surrounding DNA. DamID deposits signal in living cells over time; consequently, it represents a historic record of chromatin state over a period of multiple hours up to a full cell cycle, while antibody-based techniques provide a snapshot view. In DamID, regions that are only transiently bound by the mark of interest will thus be represented more strongly in the signal relative to CUT&Tag-based methods. Another key difference is the extent to which chromatin accessibility affects the data. DamID techniques are known to have an accessibility signature due to extended exposure to free-floating Dam protein (discussed in more detail under *Limitations*), which is controlled for by performing experiments with untethered Dam. While CUT&Tag- and CUT&RUN-based methods have reported less of such an accessibility bias and do not customarily include explicit control experiments, early results[26] suggest that such a bias may indeed be present. The question of data interpretation and normalization in light of this bias should be carefully considered among all existing single-cell genomics techniques. With regard to the transcriptional readout, the four techniques also employ different approaches: Paired-Tag exclusively amplifies the nuclear fraction of mRNA, SET-seq separates and measures total RNA in the cytoplasm, while CoTECH and scDam&T-seq both amplify the total mRNA. Finally, the Paired-Tag and CoTECH protocols have been adapted for combinatorial indexing and consequently have a higher throughput compared to scDam&T-seq and SET-seq.

Limitations

EpiDamID requires the expression of a construct encoding for the Dam-fusion protein in the system of interest. This may involve a substantial time investment depending on the system of choice and conditions generally need to be optimized for each Dam-fusion protein to optimize signal quality. DamID techniques are also limited in their resolution by the distribution of GATC motifs in the genome (median inter-GATC distance: 263 bp in mouse, 265 bp in human). In addition, we and others[47,50,51] have found that the methylation spreads ~1 kb from the site of binding (Fig. S1g), thus yielding an empirical resolution of 1-2 kb. This is sufficient to study the localization of many chromatin factors, but may be restrictive when exact binding sites are required. Finally, due to the in vivo expression and consequent roaming of the Dam-POI in the nucleus, spurious methylation gradually accumulates in unspecific, mostly accessible, chromatin regions. The degree of accumulated background signal differs substantially between different Dam-POIs, yet interferes most with proteins that reside within active chromatin. This can be overcome through computational normalization to the untethered Dam protein. In the case of single-cell experiments, this requires the grouping of similar cells into in silico populations. While this strategy yields good results, it does not provide a way to eliminate the accessibility component in individual cells, and the signal in single cells should therefore be interpreted as convolution of on-target and accessibility signal. Computational imputation of accessibility signal based on transcriptional similarity between targeted samples and Dam control samples could provide a solution to this problem, similar to current single-cell

transcriptional imputation methods (see ref[90] for an overview). We explored one experimental strategy to reduce off-target effects by implementing Dam mutants with decreased affinity for DNA, which yielded promising results in population data but insufficient ^{m6}A-events for single-cell profiling. Further adaptation of the Dam protein to engineer an enzyme with high enzymatic activity and reduced DNA-binding affinity may further improve the quality of EpiDamID profiles in single cells. Alternatively, molecular processing could be extended to facilitate an orthogonal accessibility readout from the same sample.

Acknowledgements

We would like to thank the members of the Kind laboratory for their helpful comments and suggestions. In particular, we thank Koos Rooijers for providing input and support on the computational work. This work was funded by an ERC Starting grant (ERC-StG 678423-EpiID) to JK. The Oncode Institute is partially funded by the KWF Dutch Cancer Society. IG is supported by an EMBO Long-Term Fellowship ALTF1214-2016, Swiss National Science Fund grant P400PB_186758 and NWO-ENW Veni grant VI.Veni.202.073. PDN is supported by an EMBO Long-Term Fellowship ALTF1129-2015, HFSPO Fellowship (LT001404/2017-L) and an NWO-ZonMW Veni grant (016.186.017-3). The laboratory of JB is supported by the Netherlands Cardiovascular Research Initiative: An initiative with support of the Dutch Heart Foundation and Hartekind, CVON2019-002 OUTREACH. The laboratory of HK is supported by MEXT/JSPS KAKENHI (JP18H05527, JP20K06484, and JP21H04764), and Japan Science and Technology Agency (JPMJCR16G1 and JPMJCR20S6). We additionally thank the Hubrecht Sorting Facility, the Hubrecht Imaging Center, and the Utrecht Sequencing Facility (USEQ) subsidized by the University Medical Center Utrecht.

Author contributions

Conceptualization: FJR, KLdL, SSdV, JK. Data curation & Validation: FJR, KLdL, SSdV. Formal analysis & Software: FJR. Funding acquisition & Project administration: JK. Investigation & Methodology: KLdL and SSdV designed and performed all experiments unless noted otherwise. CVQ and EB designed and generated knock-in mouse ESC lines. PDN performed all zebrafish experiments, with assistance from IG and SSdV. Resources: YS, HK. Supervision: JB, JK. Visualization: FJR, KLdL. Writing – original draft: FJR, KLdL, JK. Writing – review & editing: all authors.

Methods

Cell lines

All cell lines were grown in a humidified chamber at 37°C in 5% CO_2, and were routinely tested for mycoplasma. Human TERT-immortalized RPE-1 cells were cultured in DMEM/F12 (Gibco) containing 10% FBS (Sigma F7524 lot BCBW6329) and 1% Pen/Strep (Gibco). This cell line does not contain a Y chromosome. Human HEK293T cells were cultured in DMEM (Gibco) containing

10% FBS and 1% Pen/Strep (Gibco). This cell line does not contain a Y chromosome. Mouse F1 hybrid Cast/EiJ x 129SvJae embryonic stem cells (mESCs; a kind gift from the Joost Gribnau laboratory) were cultured on irradiated primary mouse embryonic fibroblasts (MEFs), in mESC culture media CM+/+ defined as follows: G-MEM (Gibco) supplemented with 10% FBS (Sigma F7524 lot BCBW6329), 1% Pen/Strep (Gibco), 1x GlutaMAX (Gibco), 1x non-essential amino acids (Gibco), 1x sodium pyruvate (Gibco), 0.1 mM β-mercaptoethanol (Sigma) and 1000 U/mL ESGROmLIF (EMD Millipore ESG1107). Cells were split every 3 days and medium was changed every other day. Expression of the Dam-POI constructs was suppressed by addition of 0.5 mM indole-3-acetic acid (IAA; Sigma, I5148). This cell line does not contain a Y chromosome.

Zebrafish

All experiments were conducted under the guidelines of the animal welfare committee of the Royal Netherlands Academy of Arts and Sciences (KNAW). Adult Tüpfel long fin (wild type) zebrafish (*Danio rerio*) were maintained and embryos raised and staged as previously described[91,92].

ChIP-seq

ChIP-seq was performed as described previously[93], with the following adaptations. Cells were harvested by trypsinization, and chemically crosslinked with fresh formaldehyde solution (1% in PBS) for 8 minutes while rotating at room temperature. Crosslinking was quenched with glycine on ice and sample was centrifuged at 500 g for 10 min at 4 °C. Pellet was then resuspended in lysis buffer for 5 min on ice and sonicated as follows: 16 cycles of 30 s on / 30 s off at max power (Bioruptor Diagenode), and centrifuged at 14,000 rpm at 4 °C for 10 min. The chromatin in supernatant was treated with RNase A for 30 min at 37 °C, and Proteinase K for 4 hours at 65 °C to reverse crosslinks, then cleared using DNA purification columns and eluted in nuclease-free water. Chromatin was incubated with antibodies (see below), after which Protein G beads (ThermoFisher #88847) were added for antibody binding. After successive washing, samples were cleared using DNA purification columns, eluted in nuclease-free water, and measured using a Qubit fluorometer. Libraries were prepared according to the Illumina TruSeq DNA LT kit and sequenced on the Illumina HiSeq 2500 following manufacturer's protocols. Up to 50 ng of immunoprecipitated chromatin was used as input for library preparation. Antibodies used were: anti-H3K4me3 Abcam ab8580, anti-H3K9ac Abcam ab4441, anti-H3K9me3 Abcam ab8898, anti-H3K27me3 Merck Millipore 07-449, anti-H3K36me3 Active Motif 61902, anti-H4K20me1 Abcam ab9051.

DamID construct design and lentivirus production

The constructs for mintbodies, chromatin binding domains, and full-length protein constructs were fused to Dam in both possible orientations under the control of the auxin-inducible degron (AID) system[94,95] with either the hPGK or HSP promoter, and cloned into the pCCL.sin.cPPT.ΔLNGFR.Wpre lentiviral construct[96] by standard cloning procedures.

The linkers used for the triple fusion domains are, in order of appearance:
Dam; V5 linker [GKPIPNPLLGLDST]; 1st domain (e.g., chromo); GSAGSAAGSGEF; 2nd domain; linker [KESGSVSSEQLAQFRSLD]; 3rd domain. All other POIs are linked to Dam via a V5 linker, which has been commonly used in DamID constructs[82,97,98]. The Gly- and Ser-rich flexible linker, GSAGSAAGSGEF, was designed to express GFP-fusion proteins for rapid protein-folding assay[99]. The KESGSVSSEQLAQFRSLD flexible linker was previously used for the construction of a bioactive scFv[100]. For context: the Gly and Ser residues in the linker were designed to provide flexibility, whereas Glu and Lys were added to improve the solubility[101].

Bulk DamID2
hTERT-RPE1 cells were grown as described above. At 30% confluence in 6-well plates, cells were transduced with 1500 µL total volume unconcentrated lentivirus, amounts ranging between 20-1500 µL unconcentrated lentivirus (or 0.1-40 µL concentrated) in the presence of 10 µg/mL polybrene. Cells were collected for genomic DNA isolation (Wizard, Promega) 48 h after transduction. Dam methylation levels were checked by ^{m6}A-PCR as previously described[28,102] and sequenced following the DamID2 protocol[49].

Immunofluorescent staining and confocal imaging of RPE-1 cells
Viral transduction was performed as described above for bulk DamID2, with the exception that RPE-1 cells were grown on glass coverslips. Two days after transduction, cells were washed with PBS and chemically crosslinked with fresh formaldehyde solution (2% in PBS) for 10 minutes at RT, then permeabilized (with 0.5% IGEPAL® CA-630 in PBS) for 20 minutes and blocked (with 1% bovine serum albumin (BSA) in PBS) for 30 minutes. All antibody incubations were performed in final 1% BSA in PBS followed by three PBS washes at RT. Incubation with primary antibody against the endogenous histone modification as well as purified ^{m6}A-Tracer protein[52] (recognizing methylated DNA) was performed at 4 °C for 16 hours (overnight), followed by anti-GFP (against ^{m6}A-Tracer protein) incubation at RT for 1 hour, and secondary antibody incubations at RT for 1 hour. The final PBS wash was simultaneously an incubation with DAPI at 0.5 µg/mL for 2 min, followed by a wash in MilliQ and sample mounting on glass slides using VECTASHIELD Antifade mounting medium (Vector Laboratories). Primary antibodies: anti-H3K9ac abcam ab4441 (rabbit) at 1:1000, anti-H3K9me3 abcam ab8898 (rabbit) at 1:300, anti-GFP Aves GFP-1020 (chicken) at 1:1000. Secondary antibodies: AlexaFluor anti-chicken 488 at 1:500 and anti-rabbit 647 at 1:500. Purified ^{m6}A-Tracer protein (used at 1:1000) was a kind gift from the Bas van Steensel laboratory. Imaging was performed on a Leica TCS SP8 laser scanning confocal microscope with a 63X (NA 1.40) oil-immersion objective. Images were processed in Imaris 9.3 (Bitplane) by baseline subtraction. Additional background correction was done with a 1-µM Gaussian filter for the images of Dam-CBX1 ^{m6}A-Tracer and H3K9me3 stainings.

Generation of mouse embryonic stem cell lines
The various stable clonal F1 hybrid mESC lines for the initial single cell experiments were created by lentiviral co-transduction of pCCL-EF1a-Tir1-IRES-puro and pCCL-hPGK-AID-Dam-POI constructs with a 4:1 ratio in a EF1a-Tir1-IRES-neo mother line[27], after which the

cells were selected for 10 days on 0.1% gelatine coated 10-cm dishes in 60% Buffalo Rat Liver (BRL)-conditioned medium containing 0.8 µg/mL puromycin (Sigma P9620), 250 µg/mL G418 (ThermoFisher 11811031) and 0.5 mM IAA. Individual puromycin resistant colonies were handpicked and tested for the presence of the constructs by PCR using Dam-specific primers fw-ttcaacaaaagccaggatcc and rev-gacagcggtgcataaggcgg.

The clonal F1 hybrid knock-in cell lines were CRISPR targeted in a mother line carrying Tir1-Puro in the TIGRE locus[103]. For all CRISPR targeting, cells were cultured on gelatin-coated 6-wells in 60% BRL conditioned medium to 70-90% confluency and transfected with Lipofectamin3000 (Invitrogen L3000008) according to the supplier protocol with 2 µg donor vector and 1 µg Cas9/guide vector. At 24 h after transfection the cells were split to a gelatin-coated 10-cm dish and antibiotic selection of transfected cells is started 48 h after transfection. Cells were selected with 60% BRL conditioned medium containing 0.8 µg/mL puromycin for the Tir1 knock-in and 2.5 µg/mL blasticidin (Invivogen) for the AID-Dam knock-in lines. After 5-10 days of selection, individual colonies were manually picked and screened by PCR for the correct genotype.

All CRISPR knock-in lines were made in a Tir1-TIGRE mother line that was generated by co-transfection of Cas9-gRNA plasmid pX330-EN1201(Addgene plasmid #92144) and donor plasmid pEN396-pCAGGS-Tir1-V5-2A-PuroR TIGRE (Addgene plasmid #92142)[104]. The Tir1-puro clones were screened for the presence of Tir1 by PCR from the CAGG promoter to Tir1 with the primers fw-cctctgctaaccatgttcatg and rev-tccttcacagctgatcagcacc, followed by screening for correct integration in the TIGRE locus by PCR from the polyA to the TIGRE locus with primers fw-gggaagagaatagcaggcatgct and rev-accagccacttcaaagtggtacc. The Tir1 expression is further confirmed by Western blot using a V5 antibody (Invitrogen R960-25).

A knock-in of AID-Dam in the N-terminus of the RING1B locus was made by co-transfection of a donor vector carrying the blasticidin-p2A-HA-mAID-Dam cassette flanked by 2 500-bp homology arms of the endogenous RING1B locus (pHom-BSD-p2A-HA-mAID-Dam) and p225a-RING1B spCas9-gRNA vector (sgRNA: 5'gctttttattcctagaaatgtctc3') as described above. Picked clones were screened for correct integration by PCR with primers from Dam to the RING1B locus outside the targeting construct; fw-gaacaacaagcgcatctggc and rev-tcctcccctaacctgcttttgg. Presence of the RING1B wildtype allele was checked by PCR with primers fw-tcctcccctaacctgcttttgg and rev-gccttgcctgcttggtttg. The H3K27me3 mintbody coupled to ER-mAID-Dam was knocked into the Rosa26 locus by co-transfection of pHom-ER-mAID-V5-Dam-scFv_H3K27me3-P2A-BSD-Hom donor vector and p225a-Rosa26 spCas9-RNA vector (sgRNA: gtccagtctttctagaagatgggc) as described above. Picked clones were screened for correct integration by PCR from a sequence adjacent to the Rosa homology arm to the Rosa26 locus with primers fw- gaactccatatatgggctatg and rev-cttggtgcgtttgcgggga. The untethered mAID-Dam was knocked into the Rosa26 locus by co-transfection with the pHom-ER-mAID-V5-Dam-P2A-BSD-Hom donor vector and p225a-Rosa26 spCas9-RNA vector (sgRNA: gtccagtctttctagaagatgggc) as described above. Picked clones were screened for correct integration by PCR with the same primers as for the Dam-H3K27me3 mintbody knock-in line.

All clones with correct integrations were furthermore screened for their level of induction upon IAA removal by ^{m6}A-PCR evaluated by gel electrophoresis[28,102], followed by DamID2 sequencing in bulk[49], to select the clone with a correct karyotype and the best signal-to-noise ratio of enrichment over expected regions or chromatin domains. Finally, the best 3-4 clones were selected for testing of IAA removal timing in single cells by DamID2.

Mouse embryonic stem cell culture and induction of Dam-fusion proteins

When plated for targeting or genomics experiments, cells were passaged at least 2 times in feeder-free conditions, on plates coated with 0.1% gelatin, grown in 60% BRL-conditioned medium, defined as follows and containing 1 mM IAA: 40% CM+/+ medium and 60% of CM+/+ medium conditioned on BRL cells. For timed induction of the constructs the IAA was washed out at different clone-specific times before single-cell sorting.

Embryoid body differentiation and induction of Dam-fusion proteins

For EB differentiation, the stable knock-in F1ES lines were cultured for 2 weeks on plates coated with 0.1% gelatin, grown in 2i+LIF ES cell culture medium defined as follows: 48% DMEM/F12 (Gibco) and 48% Neurobasal medium (Gibco), supplemented with 1x N2 (Gibco), 1x B27 supplement + vitamin A (Gibco), 1x non-essential amino acids, 1% FBS, 1% Pen/Strep, 0.1mM β-mercaptoethanol, 1 µM PD0325901 (Axon Medchem, PZ0162-5MG), 3 µM CHIR99021 (Tocris, SML1046-5MG), 1000 U/mL ESGRO mLIF. EB differentiation was performed according to ATCC protocol. On day 1 of differentiation, 2x10^6 cells were grown in suspension on a non-coated bacterial 10-cm dish with 15 mL CM +/- (with β-mercaptoethanol, without LIF) and 0.5 mM IAA. On day 2, half the cell suspension was divided over five non-coated bacterial 10-cm dishes each containing 15mL CM+/- medium and 0.5 mM IAA. Plates were refreshed every other day. EBs were harvested at day 7, 10, and 14. Two days before single-cell sorting, the EBs were grown in CM+/- medium containing 1 mM IAA, and induced as follows: 6 h without IAA (RING1B); 20 h without IAA and 7 h with 1 µM 4OHT (Sigma SML1666) (Dam-H3K27me3-mintbody); 7 h without IAA and 4 h with 1 µM 4OHT (untethered Dam). The EBs were evaluated by brightfield microscopy and hand-picked for further handling (see below).

FACS for single-cell experiments

FACS was performed on BD FACSJazz or BD FACSInflux Cell Sorter systems with BD Sortware. mESCs and EBs were harvested by trypsinization, centrifuged at 300 g, resuspended in medium containing 20 µg/mL Hoechst 34580 (Sigma 63493) per 1x10^6 cells and incubated for 45 minutes at 37°C. Prior to sorting, cells were passed through a 40-µm cell strainer. Propidium iodide (1 µg/mL) was used as a live/dead discriminant. Single cells were gated on forward and side scatters and Hoechst cell cycle profiles. Index information was recorded for all sorts. One cell per well was sorted into 384-well hard-shell plates (Biorad, HSP3801) containing 5 µL of filtered mineral oil (Sigma #69794) and 50 nL of 0.5 µM barcoded CEL-Seq2 primer[27,49]. In the EB experiment, the knock-in mESC lines were cultured alongside on 2i+LIF medium and included as a reference at each timepoint.

Single-cell Dam&T-seq

The scDam&T-seq protocol was performed as previously described in detail[49], with the adaptation that all volumes were halved to reduce costs. Liquid reagent dispension steps were performed on a Nanodrop II robot (Innovadyne Technologies / BioNex). Addition of barcoded adapters was done with a mosquito LV (SPT Labtech). In short, after FACS, 50 nL per well of lysis mix (0.07% IGEPAL, 1 mM dNTPs, 1:50,000 ERCC RNA spike-in mix (Ambion, 4456740)) was added, followed by incubation at 65 °C for 5 min. 100 nL of reverse transcription mix (1× First Strand Buffer and 10 mM DTT (Invitrogen, 18064-014), 2 U RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, 10777019), 10 U SuperscriptII (Invitrogen, 18064-014)) was added, followed by incubation at 42 °C for 2 h, 4 °C for 5 min and 70 °C for 10 min. Next, 885 nL of second strand synthesis mix (1× second strand buffer (Invitrogen, 10812014), 192 µM dNTPs, 0.006 U *E. coli* DNA ligase (Invitrogen, 18052019), 0.013 U RNase H (Invitrogen, 18021071), 0.26 U *E. coli* DNA polymerase (Invitrogen)) was added, followed by incubation at 16 °C for 2 h. 250 nL of protease mix was added (1× NEB CutSmart buffer, 1.0 mg/mL Proteinase K (Roche, 000000003115836001)), followed by incubation at 50 °C for 10 h and 80 °C for 20 min. Next, 115 nL of DpnI mix (1× NEB CutSmart buffer, 0.1 U NEB DpnI) was added, followed by incubation at 37 °C for 6 h and 80 °C for 20 min. Finally, 50 nL of 0.5uM DamID2 adapters were dispensed (final concentrations 25 nM), followed by 400 nL of ligation mix (1× T4 Ligase buffer (Roche, 10799009001), 0.13U T4 Ligase (Roche, 10799009001)) and incubation at 16°C for 16 hrand 65°C for 10 min. Contents of all wells were pooled and the aqueous phase was recovered by centrifugation and transfer to clean tubes. Samples were purified by incubation for 10 min with 0.8 volumes magnetic beads (CleanNA, CPCR-0050) diluted 1:7 with bead binding buffer (20% PEG8000, 2.5 M NaCl), washed twice with 80% ethanol and resuspended in 8 µL of nuclease-free water before in vitro transcription at 37 °C for 14 h using the MEGAScript T7 kit (Invitrogen, AM1334). Library preparation was done as described in the CEL-Seq2 protocol with minor adjustments[105]. Amplified RNA (aRNA) was purified with 0.8 volumes beads as described above, and resuspended in 20 µL of nuclease-free water, and fragmented at 94 °C for 90 sec with the addition of 0.25 volumes fragmentation buffer. Fragmentation was stopped by addition of 0.1 volumes of 0.5 M EDTA pH 8 and quenched on ice. Fragmented aRNA was purified with beads as described above, and resuspended in 12 µL of nuclease-free water. Thereafter, library preparation was done as previously described[105] using up to 7 µL or approximately 150 ng of aRNA, and 8-10 PCR cycles depending on input material. Libraries were sequenced on the Illumina NextSeq500 (75-bp reads) or NextSeq2000 (100-bp reads) platform.

Collection of zebrafish samples and FACS

Tüpfel long fin (wild type) pairs were set up and the following morning, approximately 1 nL of 1 ng/µL *Dam-Mphosph8* mRNA or 0.5 ng/µL *Dam-Gfp* mRNA was injected into the yolk at the 1 cell stage. Embryos were slowed down overnight at 23°C and the following morning all embryos were manually dechorionated. At 15-somite stage, embryos were transferred to 2-mL Eppendorf tubes and digested with 0.1% Collagenase type II from Cl. Histolyticum (Gibco) in Hanks Balanced Salt Solution without Mg^{2+}/Ca^{2+} (Thermofisher) for 20-30 mins at 32°C with constant shaking. Once embryos were noticeably digested, cell solution was spun at 2000 g

for 5 min at room temperature and the supernatant was removed. Cell pellet was resuspended with TrypLE Express (Thermofisher) and digested for 10 min at 32°C with constant shaking. Cell solution was inactivated with 10% Fetal Bovine Serum (Thermofisher) in Hanks Balanced Salt Solution without Mg^{2+}/Ca^{2+} and filtered through a 70-µm cell strainer (Greiner Bio-One). Cells were pelleted at 2000g 5min room temperature and washed twice with 10% Fetal Bovine Serum (Thermofisher) in Hanks Balanced Salt Solution without Mg^{2+}/Ca^{2+}. Hoechst 34580 at a final concentration of 16.8 µg/mL was added to the cell solution and incubated for 30 mins at 28°C in the dark. Solution was then filtered through a 40-µmcell strainer (Greiner Bio-One), and propidium iodide was added at a final concentration of 5 µL/mL. FACS was performed on BD FACSInflux as described above, retaining only cells in G2/M phase based on Hoechst DNA content. Plates were processed for scDam&T-seq as described above.

Immunofluorescent staining and confocal imaging of zebrafish embryos

Embryos at 15-somite stage were fixed in 4% PFA (Sigma) for 2 h at RT, followed by washes in PBS. Embryos were then washed three times in 4% sucrose/PBS and allowed to equilibrate in 30% sucrose/PBS at 4°C for 3-5 h. Embryos were suspended in Tissue Freezing Medium (Leica) orientated in the sagittal plane and frozen with dry ice. Blocks were sectioned at 8 µm and slides were rehydrated in PBS, treated with -20°C pre-cooled acetone for 7 min at -20°C, washed three times with PBS and digested with Proteinase K (Promega) at a final concentration of 10 µg/mL for 3 min, washed 1x PBS and incubated in blocking buffer (10% Fetal Bovine Serum, 1% DMSO, 0.1% Tween20 in PBS) for 30 min. Primary antibody was diluted in blocking buffer and slides incubated overnight at 4°C. Slides were washed the following day and incubated with the appropriate AlexaFluor secondary antibodies (1:500), DAPI (0.5 µg/mL) and Phalloidin-TRITC (1:200) diluted in blocking buffer for 1 h at RT. Slides were washed, covered with glass coverslips with ProLong Gold Antifade Mountant (Thermofisher) and imaged at 63X with a LSM900 confocal with AiryScan2 (Zeiss). Images were viewed and processed in Imaris 9.3 (Bitplane) and Adobe Creative Cloud (Adobe). Primary antibody: anti-H3K9me3 abcam ab8898 at 1:500[106].

Processing DamID and scDam&T-seq data

Data generated by the DamID and scDam&T-seq protocols was largely processed with the workflow and scripts described in[49] (see also www.github.com/KindLab/scDamAndTools). The procedure is described in short below.

Demultiplexing

All reads are demultiplexed based on the barcode present at the start of R1 using a reference list of barcodes. In the case of scDam&T-seq data, the reference barcodes contain both DamID-specific and CEL-Seq2-specific barcodes and zero mismatches between the observed barcode and reference are allowed. In the case of the population DamID data, the reference barcodes only contain DamID-specific barcodes and one mismatch is allowed. The UMI information, also present at the start of R1, is appended to the read name.

DamID data processing

DamID reads are aligned using bowtie2 (v. 2.3.3.1)[107] with the following parameters: "--seed 42 --very-sensitive -N 1". For human samples, the hg19 reference genome is used; for mouse samples, the mm10 reference genome; and for zebrafish samples the GRCz11 reference genome. The resulting alignments are then converted to UMI-unique GATC counts by matching each alignment to known strand-specific GATC positions in the reference genome. Any reads that do not align to a known GATC position or have a mapping quality smaller than 10 are removed. In the case of bulk DamID samples, up to 64 unique UMIs are allowed per GATC position, while up to 4 unique UMIs are allowed for single-cell samples to account for the maximum number of alleles in G2. Finally, counts are binned at the desired resolution.

CEL-Seq2 data processing

CEL-Seq2 reads are aligned using tophat2 (v. 2.1.1)[108] with the following parameters: "--segment-length 22 --read-mismatches 4 --read-edit-dist 4 --min-anchor 6 --min-intron-length 25 --max-intron-length 25000 --no-novel-juncs --no-novel-indels --no-coverage-search --b2-very-sensitive --b2-N 1 --b2-gbar 200". For mouse samples, the mm10 reference genome and the GRCm38 (v. 89) transcript models are used. For zebrafish samples, the GRCz11 reference genome and the adjusted transcript models published by the Lawson lab[109] are used. Alignments are subsequently converted to transcript counts per gene with custom scripts that assign reads to genes similar to HTSeq's[110] htseq-count with mode "intersection_strict".

Processing of ChIP-seq data

External ChIP-seq datasets were downloaded from the NCBI GEO repository and the ENCODE database[56]. The external ChIP-seq data used in this manuscript consists of: H3K9ac ChIP-seq in mESC (ENCSR000CGP), H3K27me3 ChIP-seq in mESC (ENCSR059MBO), and H3K9me3 ChIP-seq in 6-hpf zebrafish embryos[74] (GSE113086). Internal and external ChIP-seq data were processed in an identical manner. First reads were aligned using bowtie2 (v. 2.3.3.1) with the following parameters: "--seed 42 --very-sensitive -N 1". Indexes for the alignments were then generated using "samtools index" and genome coverage tracks were computed using the "bamCoverage" utility from DeepTools (v. 3.3.2)[111] with the following parameters: "--ignoreDuplicates --minMappingQuality 10". For marks that exist in broad domains in the genome, domains were called using MUSIC[112] according to the suggested workflow (https://github.com/gersteinlab/MUSIC). For marks that form narrow peaks in the genome, peaks were called using MACS2 (v. 2.1.1.20160309)[113] using the "macs2 callpeak" utility with the following parameters: "-q 0.05".

Computing the Information Content (IC) of DamID samples

The Information Content (IC) of a DamID sample is a measure of how much structure is in the detected methylation signal. It is essentially an adaptation of the RNA-seq normalization strategy called PoissonSeq[114]. Its goal is to compare the obtained signal to a background signal (the density of mappable GATCs), identify regions where the signal is similar to background, and finally compare the amount of total signal (i.e. total GATC counts) to the total signal in

background regions. The IC is the ratio of total signal over background signal and can be used to filter out samples that contain little structure in their data. The code used to compute the IC is available online (https://github.com/KindLab/EpiDamID2022) and the procedure is explained below.

As an input, we use the sample counts binned at 100-kb intervals, smoothened with a 250-kb gaussian kernel. The large bin size and smoothing are necessary when working with single-cell samples that have very sparse and peaky data and would otherwise be difficult to match to the background signal. As a control, we use the number of mappable GATCs in the same 100-kb bins, similarly smoothened. We subsequently remove all genomic bins that do not have any observed counts in the sample. Our starting data is then X, a matrix with size (n,k), where n is the number of genomic bins and k is the number of samples. Since we are comparing one experimental sample with the control, k is always 2. X_{ij} denotes the number of counts observed in the ith bin of the jth sample. We first compute the expected number of counts for each X_{ij} based on the marginal probabilities of observing counts in each bin and in each sample:

$$d = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}$$

$$p = \sum_{j=1}^{k} Xj/d = (p_1 \dots p_n)^T$$

$$q = \sum_{i=1}^{n} Xi/d = (q_1, q_2)$$

$$E = d(p \cdot q)$$

Where d is the total sum of X_{ij}; p_i is the marginal probability of observing counts in bin i; q_j is the marginal probability of observing counts in sample j; and E is the matrix of size (n,k) where entry E_{ij} is the expected number of counts in bin i for sample j, computed as $p_i q_j d$.

We subsequently compute the goodness of fit of our predictions compared to the actual counts per bin:

$$g = \sum_{j=1}^{k} \frac{X_j - E_j}{E_j}$$

Where g_i is the measure of how well the predictions of E_i match the observed counts in X_i in bin i. The better the prediction, the closer g_i is to zero, indicating that the signal of the experimental sample closely resembles the background in bin i. Next, an iterative process is performed where in each step a subset of the original bins is chosen that exclude bins with extreme

values of *g*. Specifically, all bins with a goodness of fit in the top and bottom 5th percentiles are excluded to progressively move towards a stable set of bins where the sample resembles the background. After each iteration, the chosen bins are compared to the previous set of bins and when this has stabilized, or when the maximum number of iterations is reached, the procedure stops. In practice, convergence is usually reached after only a couple of iterations. The IC is then computed for the experimental sample as the ratio of its summed total counts to the sum of counts observed in the final subset of bins.

Population DamID data filtering and analyses

The population DamID samples were filtered based on a depth threshold of 300,000 UMI-unique GATC counts and an IC of at least 1.1. Per Dam-construct, the best samples based on the IC were maintained. Samples were normalized for the total number of counts using reads per kilobase per million (RPKM). Normalization for Dam controls was performed by adding a pseudo count of 1, taking the per bin fold-change with Dam, and performing a log2-transformation, resulting in log2 observed-over-expected (log2OE) values. The UMAP presented in Fig. 1b was computed by performing principal component analysis (PCA) on the RPKM-normalized samples (20-kb bins) and using the top components for UMAP computation in python with custom scripts. For the correlations presented in Figure 1c and S1c, the RPKM-normalized DamID values were normalized for the density of mappable GATCs and log-transformed. The Spearman's rank correlation was then computed with the input-normalized ChIP-seq values of the various marks.

Resolution analysis on RPE-1 samples

To evaluate the resolution of EpiDamID signal compared to ChIP-seq, we wanted to determine the spread of the signal around regions of known enrichment. To this end, we used ChIP-seq peaks for H3K9ac and H3K4me3, and domains for H3K27me3 and H3K9me3. We computed the average ChIP-seq signal and DamID signal around these regions, using a resolution (i.e. bin size) of 200 bp. The resulting signal was mildly smoothed to get a better representation of the trends. For each sample, we then determined the distance over which the signal measured at the reference point decayed to 50% relative to the background. As a reference point, we chose the center of H3K9ac and H3K4me3 peaks, or the boundary of H3K27me3 and H3K9me3 domains. The spread of the DamID signal can then be determined as the increase in this distance relative to the corresponding ChIP-seq sample.

Single-cell DamID data filtering and analyses

Filtering and normalizing scDamID data

Single-cell DamID samples were filtered based on a depth and an IC threshold. For the mouse samples, these thresholds were 3,000 unique GATCs and an IC within the range of 1.5 to 7 (the upper threshold removes samples with very sparse profiles); for zebrafish, these thresholds were 1,000 unique GATCs and an IC within the range of 1.2 to 7. For the zebrafish samples, chromosome 4 was excluded when determining depth and IC (and in all downstream analyses)

since the reference assembly of this chromosome is poor and alignments unreliable. The quality of scDam&T-seq samples is determined separately for the DamID readout and the CEL-Seq2 readout. To preserve as much of the data as possible, we used all samples passing DamID thresholds for analyses that relied exclusively on the DamID readout. Wherever single-cell data was used, samples were normalized for their total number of GATCs, scaled by a factor 10,000, and log-transformed with a pseudo-count of 1, equivalent to the normalizations customarily performed for single-cell RNA-seq samples. To generate in silico populations based on single-cell samples, the binned UMI-unique counts of all single-cells were combined and normalization was performed equivalent to population DamID samples.

scDamID UMAPs

The UMAPs presented in Fig. 2a, Fig. 3c and Fig. 5c were computed by performing PCA on the depth-normalized single-cell samples and using the top components for UMAP computation. Since in EBs inactivation of chromosome X can coincides with a strong enrichment of H3K27me3/RING1B on that chromosome, we depth-normalized these samples using the total number of GATCs on somatic chromosomes. For the zebrafish samples, chromosome 4 was completely excluded from the analysis. For the mouse UMAPs, the single-cell data were binned at a resolution of 10-kb intervals, while for the zebrafish UMAPs, the resolution was 100 kb. Notably, when the first principal components showed a strong correlation to sample depth, it was excluded.

Single-cell count enrichment

Fig. 2b-d show the enrichment of counts in ChIP-seq domains for all single-cell mESC samples; Fig. S5f shows the enrichment of counts for all MPHOSPH8 zebrafish samples. The count enrichment is equivalent to the more well-known Fraction Reads in Peaks (FRiP) metric, but has been normalized for the expected fraction of counts within the domains based on the total number of mappable GATCs covered by these domains. In other words, if the domains cover 50% of the mappable GATCs in the genome and we observe that 70% of a sample's counts fall within these domains, the count enrichment is 0.7 / 0.5 = 1.4.

Single-cell CEL-Seq2 data filtering and analyses

Filtering CEL-Seq2 data

Single-cell data sets were evaluated with respect to the number of unique transcripts, percentage mitochondrial reads, percentage ERCC-derived transcripts and the percentage of reads coming from unannoted gene models (starting with "AC" or "Gm") and appropriate thresholds were chosen. For the EB data, the used thresholds were ^{3}1,000 UMI-unique transcripts, <7.5% mitochondrial transcripts, <1% ERCC-derived transcripts, and <5% transcripts derived from unannotated gene models. In addition, a small group of cells (29/6,554 » 0.4%) from different time points, which formed a cluster that could not be annotated and was characterized by high expression of ribosomal genes, was removed from further analyses. For the zebrafish data, the used thresholds were ^{3}1,000 UMI-unique transcripts and <5% ERCC-

derived transcripts. Only genes observed in at least 5 samples across the entire dataset were maintained in further analyses. The quality of scDam&T-seq samples is determined separately for the DamID readout and the CEL-Seq2 readout. To preserve as much of the data as possible, we used all samples passing CEL-Seq2 thresholds (independent of DamID quality) for transcriptome-based analyses.

Analysis of CEL-Seq2 data with Seurat and Harmony

Single-cell transcription data was processed using Seurat (v3)[115]. First, samples were processed using the "NormalizeData", "FindVariableFeatures", "ScaleData", and "RunPCA" commands with default parameters. Subsequently, batch effects relating to processing batch and plate were removed using Harmony[116] using the "RunHarmony" command, using a theta=2 for the batch variable and theta=1 for the plate variable. Clustering and dimensionality reduction were subsequently performed with the "FindNeighbors", "FindClusters" and "RunUMAP" commands. Differentially expressed genes per cluster were found using the "FindAllMarkers" command.

Integration with external single-cell datasets

The EB data was integrated with part of the single-cell mouse embryo atlas published by[65] and with the transcription data from the scNMT-seq EB dataset published by[67]. In the case of the mouse embryo atlas, the data was loaded directly into R via the provided R package "MouseGastrulationData". One data set per time point was included (datasets 18, 14, 19, 16, 17, corresponding to embryonic stages E6.5, E7.0, E7.5, E8.0, E8.5, respectively). In the case of the scNMT-seq dataset, the transcript count tables were downloaded from the repository provided in the publication. Only cells derived fom wild type embryos were included. The external data and our own data was integrated using the SCTransform[117] and the anchor-based intregration[115] functionalities from Seurat. First, all data was normalized per batch using the "SCTransform" command. Data sets were then integrated using the "SelectIntegrationFeatures", "PrepSCTIntegration", "FindIntegrationAnchors", and "IntegrateData", as per Seurat documentation. To assign scNMT-seq samples to the previously determined EB clusters, we used Seurat's "TransferData" command.

SCENIC

We used SCENIC[72] on the command line according to the documentation provided for the python-based scalable version of the tool (pySCENIC)[71]. Specifically, we ran "pyscenic grn" with the parameters "--method grnboost2"; "pyscenic ctx" with the parameters "--all_modules"; and "pyscenic aucell" with the default parameters. We used the transcription factor annotation and the transcription factor motifs (10 kb +/- of the TSS) provided with SCENIC. This yielded 414 activating regulons. We subsequently filtered regulons based on the expression of the regulon as a whole (at least 50% of cells having an AUCell score > 0 within at least one Seurat cluster) and based on the expression of the regulon transcription factor (detected in at least 5% of cells in at least one cluster) to retain only high confidence regulons. This resulted in

285 remaining activating regulons. However, repeating all analyses with the unfiltered set of regulons yielded the same trends and relationships.

Linear Discriminant Analysis (LDA) classifier to assign samples to transcriptional clusters based on DamID signal

In both the EB results and the zebrafish results, we noticed that there was a substantial number of cells that passed DamID thresholds, but that had a poor CEL-Seq2 readout. Since most of our analyses rely on the separation of cells in transcriptional clusters (i.e. cell types) and cells with a poor CEL-Seq2 readout cannot be included in the clustering, these cells cannot be used in downstream DamID-based analyses. However, we noticed that the separation of different cell types was recapitulated to a considerable extent in low-dimensionality representations of the DamID readout (see the DamID-based UMAPs in Fig. 2a and Fig. 3d). Since cell-type information is captured in the DamID readout, we reasoned that a classifier could be trained based on cells with both good DamID and CEL-Seq2 readouts to assign cells with a poor CEL-Seq2 readout to transcriptional clusters based on their DamID readout.

To this end, we implemented a Linear Discriminant Analysis (LDA) classifier as described below. In addition, the code is available online (https://github.com/KindLab/EpiDamID2022).

Data input and preprocessing

As in input for the classifier, we used the binned DamID data of all samples passing DamID thresholds and the transcriptional cluster labels of these samples (samples with a poor CEL-Seq2 readout had the label "unknown"). The DamID data was depth-normalized (as described above) and genomic bins that contained fewer than 1 mappable GATC motif per kb were excluded, resulting in a matrix of size N x M, where N is the number of samples and M is the number of remaining genomic bins. For the EB data, a bin size of 10 kb was used, while a bin size of 100 kb was used for the zebrafish data. Subsequently, the pair-wise correlation was computed between all samples, resulting in a correlation matrix of size N x N. This transformation had two reasons: First, it served as a dimensionality reduction, since $N << M$. Second, it resulted in a data type that effectively describes the similarity of a sample with all other samples, including samples without a cluster label. Consequently, during the training phase, the classifier can indirectly use the information of these unlabeled samples to learn about the overall data structure. We found that using the correlation matrix (N x N) as an input for the classifier yielded much better results than using the original matrix (N x M).

To train the LDA classifier, we used two thirds (~66%) of all samples with cluster labels (i.e. with a good CEL-Seq2 readout). Since the number of cells per cluster varied extensively, we randomly selected two thirds of the samples per cluster and thereby ensured that all clusters were represented in both training and testing. The training data thus consisted of the correlation matrix of size N_{train} x N and a list of sample labels of size N_{train}, where N_{train} is the number of samples used for training. Consequently, we retained one third (~33%) of labelled samples to test the performance of the LDA classifier, consisting of the correlation matrix of

size N_{test} x N and a list of sample labels of size N_{test}, where N_{test} is the number of samples used for testing. In summary, this split the samples into three groups: one group for training, one group for testing, and the group of unlabeled samples.

Training the classifier
For the implementation of the LDA classifier, we used the "LinearDiscriminantAnalysis" function provided in the Python (v. 3.8.10) scikit-learn toolkit (v. 0.24.2). The number of components was set to the number of transcriptional clusters minus one and the LDA classifier was trained using the training samples.

Testing the performance
To test the performance, the trained LDA classifier was used to predict the labels of the training set of samples. Predictions with a probability larger than 0.5 were maintained, while predictions with a lower probability were discarded (and the corresponding cells were thus not labelled). The predicted labels were subsequently compared to the known labels (Fig. S3e, S5c). In general, we found a very good performance for clusters with many cells, while the performance tended to be lower for clusters with few cells. This is as expected, since the number of samples for these clusters was also very low during training.

Predicting cluster labels for unlabeled samples
After establishing that the performance was satisfactory, the LDA was retrained, this time using all labelled samples. The actual performance on the unlabeled data is likely higher than the performance on the test data, since the number of samples used for the final training is notably higher. Finally, the cluster labels were predicted for the unlabeled samples. Once again, only predictions with a probability higher than 0.5 were maintained. Fig. S3d shows the number of EB samples that were attributed to each cluster using the LDA classifier, as well as the number of samples that could not be attributed ("unassigned").

Defining PRC targets
First, we identified for each gene the region of 5 kb upstream and 3 kb downstream of the TSS. Only protein-coding genes and genes for non-coding RNA were considered. When the TSS domains of two genes overlapped, they were merged if the overlap was >4 kb, otherwise the two domains were split in the middle of the overlap. This resulted in 30,356 domains covering a total of 35,814 genes. Subsequently, for all single-cells, the number of observed GATC counts within each domain was determined. In silico populations per transcriptional cluster were generated by combining the counts of all cells belonging to each cluster per DamID construct. The in silico population counts were subsequently RPKM-normalized, using the total number of GATC counts on the somatic chromosomes of the combined single-cell samples as the depth (i.e. also counts outside the domains). Normalization for Dam controls was performed for the H3K27me3 and RING1B data per transcriptional cluster by adding a pseudo count of 1, taking the fold-change with Dam, and performing a log2-transformation, resulting in log2 observed-over-expected (log2OE) values. The correlation of the resulting H3K27me3 and RING1B values

per cluster is shown in Fig. S3f. We subsequently determined PRC targets as those genes that showed H3K27me3 and RING1B log2OE values >0.35 in at least one cluster. PRC targets were defined based on the in silico population of the H3K27me3 and RING1B data of the mESCs (Fig. 2) and the EB clusters, excluding cluster 7. Cluster 7 was excluded, because it consisted of relatively few cells and the combined data was consequently sparse.

Comparing EpiDamID and scNMT-seq data at transcription start sites

We downloaded the tables of single-cell CpG methylation values at regions +/-2 kb of gene TSS from the repository provided in the scNMT-seq publication[67]. We subsequently averaged the CpG methylation scores across cells per cluster to gain an average CpG methylation for all genes per cluster. This could be done for four out of eight transcriptional clusters to which sufficient scNMT-seq samples were attributed (cluster 3: 31 cells; cluster 5: 21 cells; cluster 1: 37 cells; cluster 4: 43 cells). We subsequently could integrate the CpG methylation scores with our own H3K27me3 and RING1B DamID data for all genes, for which the enrichment scores were computed as described in the previous section. The subsequent analyses were performed on genes that were represented in both datasets.

ChromHMM of zebrafish in silico populations

In order to determine regions that were characterized by H3K9me3-enrichment in specific (sets of) cell types in the zebrafish embryo, we made use of ChromHMM (v. 1.22)[79,80]. As input, we used the in silico H3K9me3 signal (log2OE) of all clusters that had at least 30 cells passing DamID thresholds for both Dam and MPHOSPH8 (clusters 0-11). The genome-wide signal at a resolution of 50 kb was used and the values were binarized based on a threshold of log2OE > 0.35. Bins that had fewer than 1 mappable GATC per kb were given a value of 2, indicating that the data was missing. As in all other analysis, chromosome 4 was excluded. The binarized values of clusters 0-11 were provided as input for the ChromHMM and the results were computed using the "LearnModel" function using the following parameters: -b 50000 -s 1 -pseudo. The number of ChromHMM states was varied from 2 to 10 and for each result the differences between the states (based on the emission probabilities) were inspected. We found that a ChromHMM model with 5 states was optimal, since this yielded the most diverse states and increasing the number of states just added redundant states with similar emission probabilities.

Repeat enrichment in ChromHMM states

The RepeatMasker repeat annotations for GRCz11 were downloaded from the UCSC Genome Browser website (https://genome.ucsc.edu/). The enrichment of repeats within each ChromHMM state was computed either for repeat classes as a whole (Fig. S6a) or for individual types of repeats (Fig. 5i and S6c). To compute the enrichment of a repeat class/type in a ChromHMM state, the fraction of repeats belonging to that class/type that fell within the state was computed and normalized for the fraction of the genome covered by that state. In other words, if we observe that 70% of a certain repeat falls within state B and state B covers 7% of the genome, then the repeat enrichment is 0.7 / 0.07 = 10.

GO term and PANTHER protein class enrichment analysis

GO term and PANTHER[118] protein class enrichment analyses were performed via de Gene Ontology Consortium website (http://geneontology.org/). For Fig. S4e, the list of PRC-regulated TFs was used as a query and the list of all TFs as a reference to determine enriched Biological Process GO terms. Only the top 10 most significant terms are shown. For Fig. S5g, the list of genes in ChromHMM state A1 or B was used as a query and the list of genes in all ChromHMM states as a reference to determine enriched PANTHER protein classes. All hits are shown.

Quantification and statistical analysis

The number of n samples included in analyses is provided within each figure and/or accompanying figure legend. Statistical p values are associated with the significance test as described in the figure legends. The boxes of boxplots indicate the quartiles of the dataset, the middle shows the median, and the error bars of indicate 1.5 times the inter-quartile range.

Data and Code Availability

All sequencing data generated in this manuscript are deposited on the NCBI Gene Expression Omnibus (GEO) portal and are publicly available as of the data of publication under accession number GSE184036 (see Key Resource Table for further details). Imaging data are publicly available on Mendeley Data (DOI: 10.17632/sp7hsw68c4.1). Key scripts are available at https://github.com/KindLab/EpiDamID2022. Any additional information required to reanalyze the data reported in this paper is available from the Lead Contact upon request.

References

1. Juan, A. H. *et al.* Roles of H3K27me2 and H3K27me3 Examined during Fate Specification of Embryonic Stem Cells. *Cell Rep* **17**, 1369–1382 (2016).
2. Nicetto, D. *et al.* H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. *Science (1979)* **363**, 294–297 (2019).
3. Pengelly, A. R., Copur, Ö., Jäckle, H., Herzig, A. & Müller, J. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor polycomb. *Science (1979)* **339**, 698–699 (2013).
4. Hirota, T., Lipp, J. J., Toh, B.-H. & Peters, J.-M. Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. *Nature 2005 438:7071* **438**, 1176–1180 (2005).
5. Liu, W. *et al.* PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* **466**, 508–512 (2010).
6. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA Double-stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139 *. *Journal of Biological Chemistry* **273**, 5858–5868 (1998).
7. Sanders, S. L. *et al.* Methylation of Histone H4 Lysine 20 Controls Recruitment of Crb2 to Sites of DNA Damage. *Cell* **119**, 603–614 (2004).
8. Solomon, M. J. & Varshavsky, A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences* **82**, 6470–6474 (1985).
9. Schmid, M., Durussel, T. & Laemmli, U. K. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol Cell* **16**, 147–57 (2004).
10. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, 1–35 (2017).
11. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology 2015 33:11* **33**, 1165–1172 (2015).
12. Harada, A. *et al.* A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat Cell Biol* **21**, 287–296 (2019).
13. Carter, B. *et al.* Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat Commun* **10**, (2019).
14. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* **51**, 1060–1066 (2019).
15. Ku, W. L. *et al.* Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nat Methods* **16**, 323–325 (2019).
16. Hainer, S. J., Bošković, A., McCannell, K. N., Rando, O. J. & Fazzio, T. G. Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* **177**, 1319-1329.e11 (2019).
17. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1–10 (2019).
18. Wang, Q. *et al.* CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Mol Cell* **76**, 206-216.e7 (2019).
19. Ai, S. *et al.* Profiling chromatin states using single-cell itChIP-seq. *Nat Cell Biol* **21**, 1164–1172 (2019).
20. Zeller, P. *et al.* Hierarchical chromatin regulation during blood formation uncovered by single-cell sortChIC. *bioRxiv* 2021.04.26.440606 (2021) doi:10.1101/2021.04.26.440606.
21. Zhu, C. *et al.* Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat Methods* **18**, 283–292 (2021).
22. Xiong, H., Luo, Y., Wang, Q., Yu, X. & He, A. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat Methods* **18**, 652–660 (2021).
23. Sun, Z. *et al.* Joint single-cell multiomic analysis in Wnt3a induced asymmetric stem cell division. *Nat Commun* **12**, (2021).
24. 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 2013 10:12* **10**, 1213–1218 (2013).

25. Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nature Protocols 2020 15:10* **15**, 3264–3283 (2020).

26. Zhang, W. &, Wang, M. & Zhang, Y. Tn5 transposase-based epigenomic profiling methods are prone to open chromatin bias. *bioRxiv* 2021.07.09.451758 (2021) doi:10.1101/2021.07.09.451758.

27. Rooijers, K. *et al.* Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nat Biotechnol* (2019) doi:10.1038/s41587-019-0150-y.

28. Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. Detection of in vivo protein - DNA interactions using DamID in mammalian cells. *Nat Protoc* **2**, 1467–1478 (2007).

29. Filion, G. J. *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143**, 212–224 (2010).

30. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nature Biotechnology 2000 18:4* **18**, 424–428 (2000).

31. Kungulovski, G., Mauser, R., Reinhardt, R. & Jeltsch, A. Application of recombinant TAF3 PHD domain instead of anti-H3K4me3 antibody. *Epigenetics Chromatin* **9**, (2016).

32. Kungulovski, G. *et al.* Application of histone modification-specific interaction domains as an alternative to antibodies. *Genome Res* **24**, 1842–1853 (2014).

33. Vermeulen, M. *et al.* Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell* **131**, 58–69 (2007).

34. Sato, Y. *et al.* Genetically encoded system to track histone modification in vivo. *Sci Rep* **3**, (2013).

35. Sato, Y. *et al.* A Genetically Encoded Probe for Live-Cell Imaging of H4K20 Monomethylation. *J Mol Biol* **428**, 3885–3902 (2016).

36. Tjalsma, S. J. D. *et al.* H4K20me1 and H3K27me3 are concurrently loaded onto the inactive X chromosome but dispensable for inducing gene silencing. *EMBO Rep* **22**, 1–17 (2021).

37. Sato, Y., Nakao, M. & Kimura, H. Live-cell imaging probes to track chromatin modification dynamics. *Microscopy* 1–8 (2021) doi:10.1093/JMICRO/DFAB030.

38. Villaseñor, R. *et al.* ChromID identifies the protein interactome at chromatin marks. *Nat Biotechnol* (2020) doi:10.1038/s41587-020-0434-2.

39. Altemose, N. *et al.* µDamID: A Microfluidic Approach for Joint Imaging and Sequencing of Protein-DNA Interactions in Single Cells. *Cell Syst* **11**, 354-366.e9 (2020).

40. Borsos, M. *et al.* Genome–lamina interactions are established de novo in the early mouse embryo. *Nature* **569**, 729–733 (2019).

41. Kind, J. *et al.* Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. *Cell* **153**, 178–192 (2013).

42. Southall, T. D. *et al.* Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: Assaying RNA pol II occupancy in neural stem cells. *Dev Cell* **26**, 101–112 (2013).

43. Wong, X. *et al.* Mapping the micro-proteome of the nuclear lamina and lamina-associated domains. *Life Sci Alliance* **4**, (2021).

44. Kind, J. *et al.* Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* **163**, 134–147 (2015).

45. Cheetham, S. W. *et al.* Single-molecule simultaneous profiling of DNA methylation and DNA-protein interactions with Nanopore-DamID. *bioRxiv* 2021.08.09.455753 (2021) doi:10.1101/2021.08.09.455753.

46. Pal, M., Kind, J. & Torres-Padilla, M. E. DamID to map genome-protein interactions in preimplantation mouse embryos. *Methods in Molecular Biology* **2214**, 265–282 (2021).

47. Szczesnik, T., Ho, J. W. K. & Sherwood, R. Dam mutants provide improved sensitivity and spatial resolution for profiling transcription factor binding. *Epigenetics Chromatin* **12**, 1–11 (2019).

48. Park, M., Patel, N., Keung, A. J. & Khalil, A. S. Construction of a Synthetic, Chromatin-Based Epigenetic System in Human Cells. *SSRN* (2018) doi:10.2139/ssrn.3155804.

49. Markodimitraki, C. M. *et al.* Simultaneous quantification of protein–DNA interactions and transcriptomes in single cells with sc-Dam&T-seq. *Nat Protoc* **15**, 1922–1953 (2020).

50. Tosti, L. *et al.* Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo. *Genome Res* **28**, 592–605 (2018).

51. Cheetham, S. W. *et al.* Targeted damid reveals differential binding of mammalian pluripotency factors. *Development (Cambridge)* **145**, (2018).

52. Schaik, T. van, Vos, M., Peric-Hupkes, D., Celie, P. H. & Steensel, B. van. Cell cycle dynamics of lamina-associated DNA. *EMBO Rep* **21**, e50636 (2020).

53. Shoaib, M. *et al.* Histone H4 lysine 20 mono-methylation directly facilitates chromatin openness and promotes transcription of housekeeping genes. *Nat Commun* **12**, 1–16 (2021).

54. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature 2006 441:7091* **441**, 349–353 (2006).

55. Riising, E. M. *et al.* Gene Silencing Triggers Polycomb Repressive Complex 2 Recruitment to CpG Islands Genome Wide. *Mol Cell* **55**, 347–360 (2014).

56. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794–D801 (2018).

57. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature 2004 431:7010* **431**, 873–878 (2004).

58. de Napoles, M. *et al.* Polycomb Group Proteins Ring1A/B Link Ubiquitylation of Histone H2A to Heritable Gene Silencing and X Inactivation. *Dev Cell* **7**, 663–676 (2004).

59. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in polycomb-group silencing. *Science (1979)* **298**, 1039–1043 (2002).

60. Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev* **16**, 2893–2905 (2002).

61. Czermin, B. *et al.* Drosophila Enhancer of Zeste/ESC Complexes Have a Histone H3 Methyltransferase Activity that Marks Chromosomal Polycomb Sites. *Cell* **111**, 185–196 (2002).

62. Müller, J. *et al.* Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex. *Cell* **111**, 197–208 (2002).

63. Piunti, A. & Shilatifard, A. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nature Reviews Molecular Cell Biology 2021 22:5* **22**, 326–345 (2021).

64. Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat Rev Mol Cell Biol* **0123456789**, (2021).

65. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

66. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).

67. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

68. Li, Y. *et al.* Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol* **19**, 1–16 (2018).

69. Hagarman, J. A., Motley, M. P., Kristjansdottir, K. & Soloway, P. D. Coordinate Regulation of DNA Methylation and H3K27me3 in Mouse Embryonic Stem Cells. *PLoS One* **8**, e53880 (2013).

70. Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res* **22**, 1128–1138 (2012).

71. van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* **15**, 2247–2276 (2020).

72. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).

73. Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620–631 (2018).

74. Laue, K., Rajshekar, S., Courtney, A. J., Lewis, Z. A. & Goll, M. G. The maternal to zygotic transition regulates genome-wide heterochromatin establishment in the zebrafish embryo. *Nat Commun* **10**, (2019).

75. Mutlu, B. *et al.* Regulated nuclear accumulation of a histone methyltransferase times the onset of heterochromatin formation in C. elegans embryos. *Sci Adv* **4**, 6224–6246 (2018).

76. Rudolph, T. *et al.* Heterochromatin Formation in Drosophila Is Initiated through Active Removal of H3K4 Methylation by the LSD1 Homolog SU(VAR)3-3. *Mol Cell* **26**, 103–115 (2007).

77. Santos, F., Peters, A. H., Otte, A. P., Reik, W. & Dean, W. Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. *Dev Biol* **280**, 225–236 (2005).

78. Ahmed, K. *et al.* Global Chromatin Architecture Reflects Pluripotency and Lineage Commitment in the Early Mouse Embryo. *PLoS One* **5**, e10531 (2010).

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

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

81. Hahn, M. A., Wu, X., Li, A. X., Hahn, T. & Pfeifer, G. P. Relationship between Gene Body DNA Methylation and Intragenic H3K9me3 and H3K36me3 Chromatin Marks. *PLoS One* **6**, e18844 (2011).

82. Vogel, M. J. *et al.* Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res* **16**, 1493–1504 (2006).

83. Mosch, K., Franz, H., Soeroes, S., Singh, P. B. & Fischle, W. HP1 Recruits Activity-Dependent Neuroprotective Protein to H3K9me3 Marked Pericentromeric Heterochromatin for Silencing of Major Satellite Repeats. *PLoS One* **6**, e15894 (2011).

84. Liu, S. *et al.* Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. *Genes Dev* **28**, 2041–2055 (2014).

85. Bulut-Karslioglu, A. *et al.* Suv39h-Dependent H3K9me3 Marks Intact Retrotransposons and Silences LINE Elements in Mouse Embryonic Stem Cells. *Mol Cell* **55**, 277–290 (2014).

86. Gruenbaum, Y. & Foisner, R. Lamins: Nuclear Intermediate Filament Proteins with Fundamental Functions in Nuclear Mechanics and Genome Regulation, **84**, 131–164 (2015).

87. Donnaloja, F., Carnevali, F., Jacchetti, E. & Raimondi, M. T. Lamin A/C Mechanotransduction in Laminopathies. *Cells 2020, Vol. 9, Page 1306* **9**, 1306 (2020).

88. Corallo, D., Trapani, V. & Bonaldo, P. The notochord: structure and functions. *Cellular and Molecular Life Sciences 2015 72:16* **72**, 2989–3008 (2015).

89. Park, M., Patel, N., Keung, A. J. & Khalil, A. S. Engineering Epigenetic Regulation Using Synthetic Read-Write Modules. *Cell* **176**, 227-238. e20 (2019).

90. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol* **21**, 1–30 (2020).

91. Aleström, P. *et al.* Zebrafish: Housing and husbandry recommendations, **54**, 213–224 (2019).

92. Westerfield, M. *The Zebrafish Book : A Guide for the Laboratory Use of Zebrafish, 4th Edition*. (University of Oregon Press, Eugene, 2000).

93. Collas, P. A Chromatin Immunoprecipitation Protocol for Small Cell Numbers. *Methods in Molecular Biology* **791**, 179–193 (2011).

94. Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T. & Kanemaki, M. An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nature Methods 2009 6:12* **6**, 917–922 (2009).

95. Kubota, T., Nishimura, K., Kanemaki, M. T. & Donaldson, A. D. The Elg1 Replication Factor C-like Complex Functions in PCNA Unloading during DNA Replication. *Mol Cell* **50**, 273–280 (2013).

96. Amendola, M., Venneri, M. A., Biffi, A., Vigna, E. & Naldini, L. Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. *Nature Biotechnology 2004 23:1* **23**, 108–116 (2005).

97. Meuleman, W. *et al.* Constitutive nuclear lamina–genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* **23**, 270–280 (2013).

98. Peric-Hupkes, D. *et al.* Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Mol Cell* **38**, 603–613 (2010).

99. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnology 1999 17:7* **17**, 691–695 (1999).

100. Bird, R. E. *et al.* Single-Chain Antigen-Binding Proteins. *Science (1979)* **242**, 423–426 (1988).

101. Chen, X., Zaro, J. L. & Shen, W. C. Fusion protein linkers: Property, design and functionality. *Adv Drug Deliv Rev* **65**, 1357–1369 (2013).

102. de Luca, K. L. & Kind, J. Single-cell damid to capture contacts between dna and the nuclear lamina in individual mammalian cells. in *Methods in Molecular Biology* vol. 2157 159–172 (Humana, New York, NY, 2021).

103. Zeng, H. *et al.* An Inducible and Reversible Mouse Genetic Rescue System. *PLoS Genet* **4**, e1000069 (2008).

104. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).

105. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 1–7 (2016).

106. Chandra, T. *et al.* Independence of Repressive Histone Marks and Chromatin Compaction during Senescent Heterochromatic Layer Formation. *Mol Cell* **47**, 203–214 (2012).

107. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4* **9**, 357–359 (2012).

108. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology 2013 14:4* **14**, 1–13 (2013).

109. Lawson, N. D. *et al.* An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *Elife* **9**, 1–76 (2020).

110. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

111. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–W165 (2016).

112. Harmanci, A., Rozowsky, J. & Gerstein, M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol* **15**, 474 (2014).

113. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, 1–9 (2008).

114. Li, J., Witten, D. M., Johnstone, I. M. & Tibshirani, R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**, 523–538 (2012).

115. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

116. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).

117. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, (2019).

118. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols 2013 8:8* **8**, 1551–1566 (2013).

Supplementary Figures

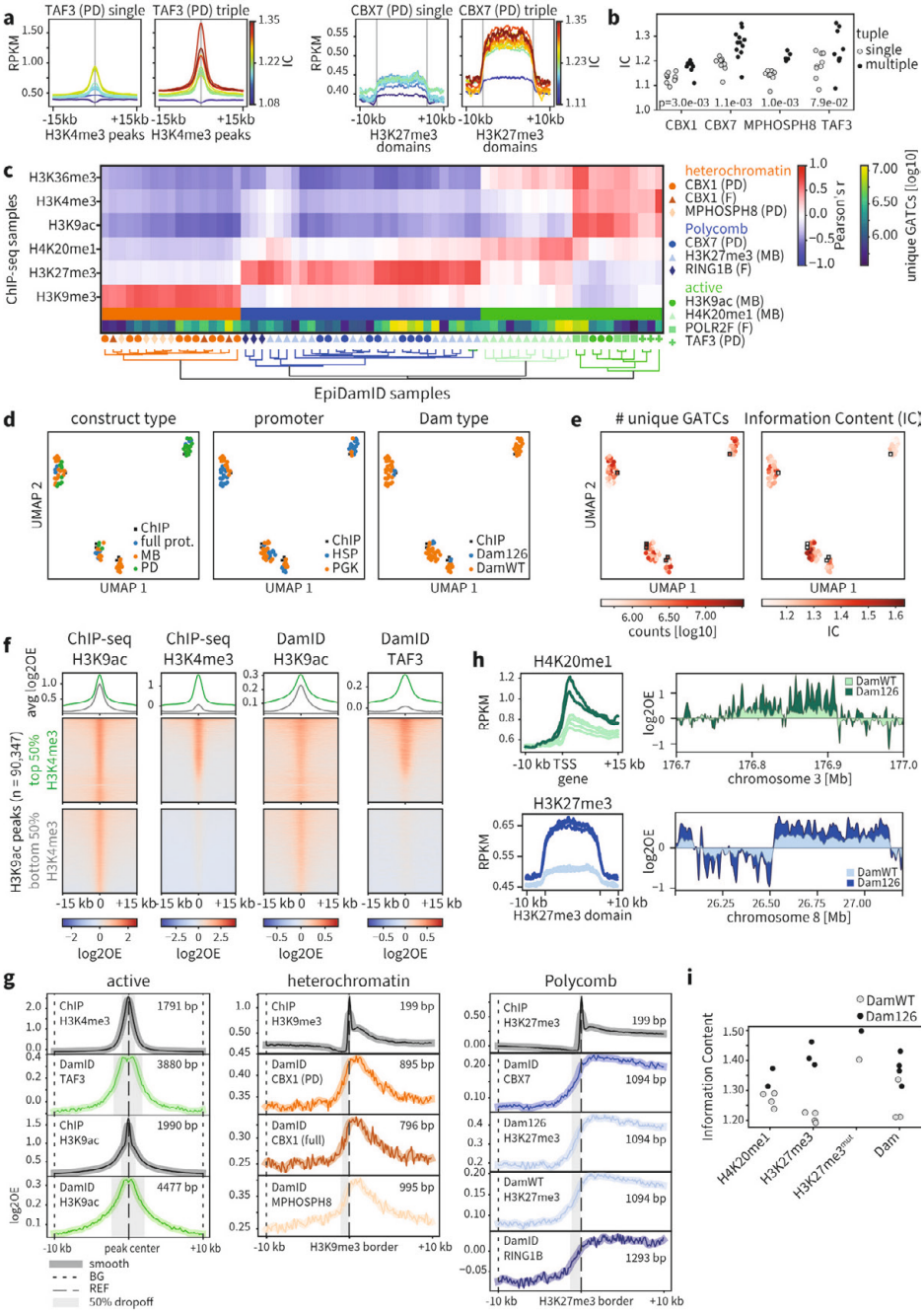


Figure S1: Technical validation of EpiDamID data

a, Average enrichment over genomic regions of interest for TAF3 and CBX7 DamID. Left: data generated by

fusing Dam to a single protein domain; Right: data generated by fusing Dam to a trimer of the same protein domain. Sample lines are colored by their Information Content (IC). **b**, Strip plot of samples comparing the IC of single (grey) and multiple (black) targeting domains. Per construct, the significance was tested with a two-sided Mann-Whitney U test. **c**, Clustered heatmap showing the correlation between ChIP-seq and Dam-normalized DamID. Correlations were computed using Pearson's correlation. Samples are labeled by their targeting domain (colored shapes) and number of unique GATC counts. **d**, UMAPs of samples, colored by construct properties. **e**, UMAPs of samples, colored by the number of unique GATC counts and IC. **f**, ChIP-seq and DamID enrichment at H3K9ac peaks (center +/- 15 kb), split into two categories according to ChIP-seq H3K4me3 occupancy (highest and lowest 50%). The heatmaps show the enrichment for each peak region, while the line plots on top show the average enrichment per H3K4me3 category. **g**, Signal resolution analysis. The plots show ChIP-seq and DamID enrichment at genomic regions of interest +/- 10 kb. Left panel shows active marks; signal is centered around ChIP-seq H3K4me3 and H3K9ac peaks for DamID TAF3 and H3K9ac, respectively. Middle panel shows heterochromatin; signal is centered around ChIP-seq H3K9me3 domain borders. Right panel shows Polycomb; signal is centered around ChIP-seq H3K27me3 domain borders. Solid line indicates the mean signal at these regions, shaded line indicates smoothened signal. Large dashed line indicates the location of the highest signal in ChIP (REF); small dashed line indicates the background measuring point (BG). Grey shaded area indicates the region over which the signal at the REF point drops with 50% relative to the BG point. The size of the drop-off distance is indicated in the top left. **h**, Comparison of DamWT and Dam126 signal. Left: average DamID enrichment plots over genomic regions of interest. Regions are the TSS of the top 25% H3K9ac-enriched genes for H4K20me1 (top), and ChIP-seq domains for H3K27me3 (bottom). Right: genome browser views of DamID enrichment corresponding to left panels. The data shown in H represent the combined data of all samples of a particular targeting domain. **i**, Strip plot of samples comparing the IC of DamWT (grey) and Dam126 (black) targeting constructs.

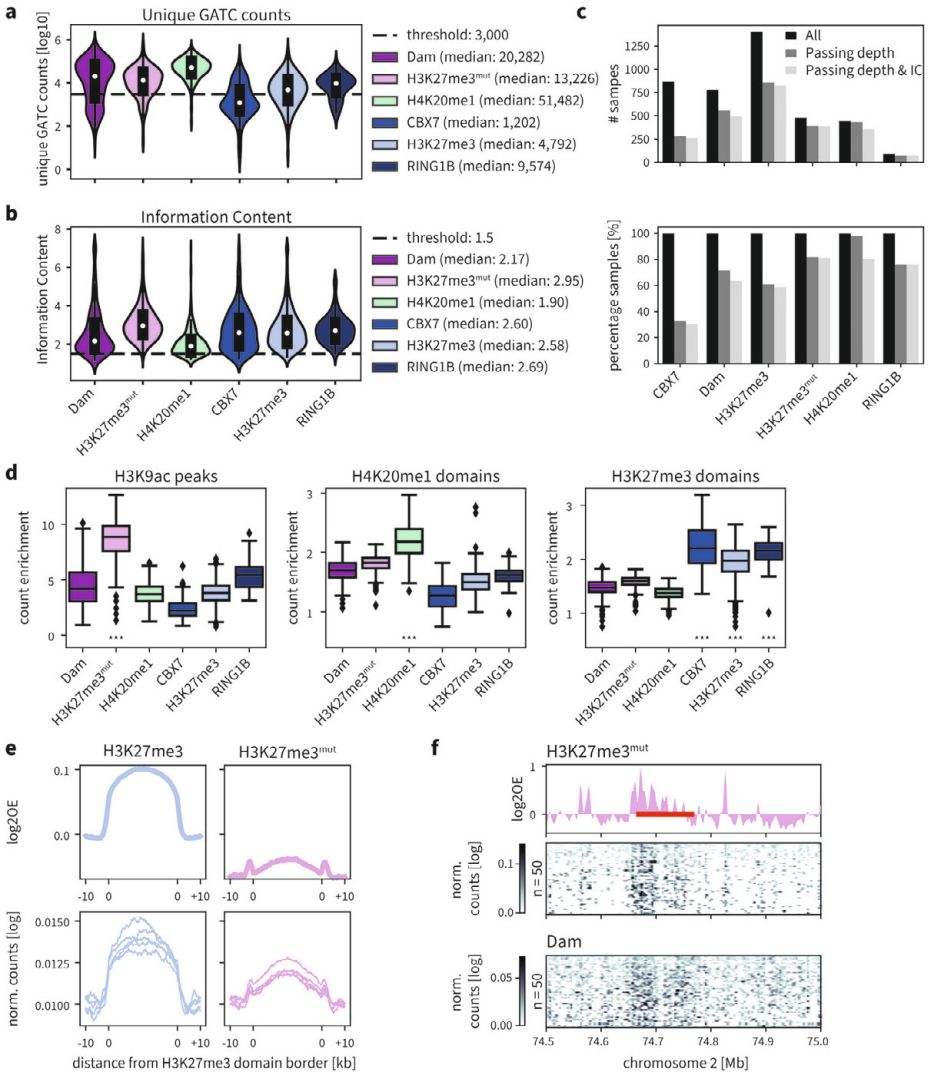


Figure S2: Detection of histone PTMs in single mouse embryonic stem cells with a single-cell implementation of EpiDamID

a, Violin plots indicating the distribution of the number of unique GATCs detected for each cell line. The dashed line indicates the threshold used for data filtering. **b**, Violin plots indicating the distribution of the Information Content (IC) after filtering on depth for each cell line. The dashed line indicates the threshold used for data filtering. **c**, Overview of the number (top) and percentage (bottom) of samples retained after filtering on depth and IC. **d**, Boxplots showing the count enrichment in H3K9ac ChIP-seq peaks (left), H4K20me1 ChIP-seq domains (middle), and H3K27me3 ChIP-seq domains (right) of all single cells per DamID construct. Count enrichment was computed as the fraction of GATC counts that fell within the regions, relative to the total fraction of genomic GATC positions inside these domains. In each plot, the enrichment of constructs of interest are compared to the enrichment in the Dam control. The significance of the difference was tested with a two-sided Mann-Whitney-U test. *** indicates a p-values smaller than 0.001. Constructs without an indication of significance were not tested. **e**, Average signal over H3K27me3 ChIP-seq domains of H3K27me3 and H3K27me3^{mut} mintbodies. Top: in silico populations

normalized for Dam; Bottom: five of the best single-cell samples (bottom) normalized by read depth. **f**, Signal of H3K27me3^{mut} and Dam control over the HoxD cluster and neighboring regions. The DamID track show the Dam-normalized in silico populations of H3K27me3^{mut}, while the heatmaps show the depth-normalized single-cell data of the fifty richest cells for H3K27me3^{mut} and Dam. The red bar around 74.7 Mb indicates the HoxD cluster.

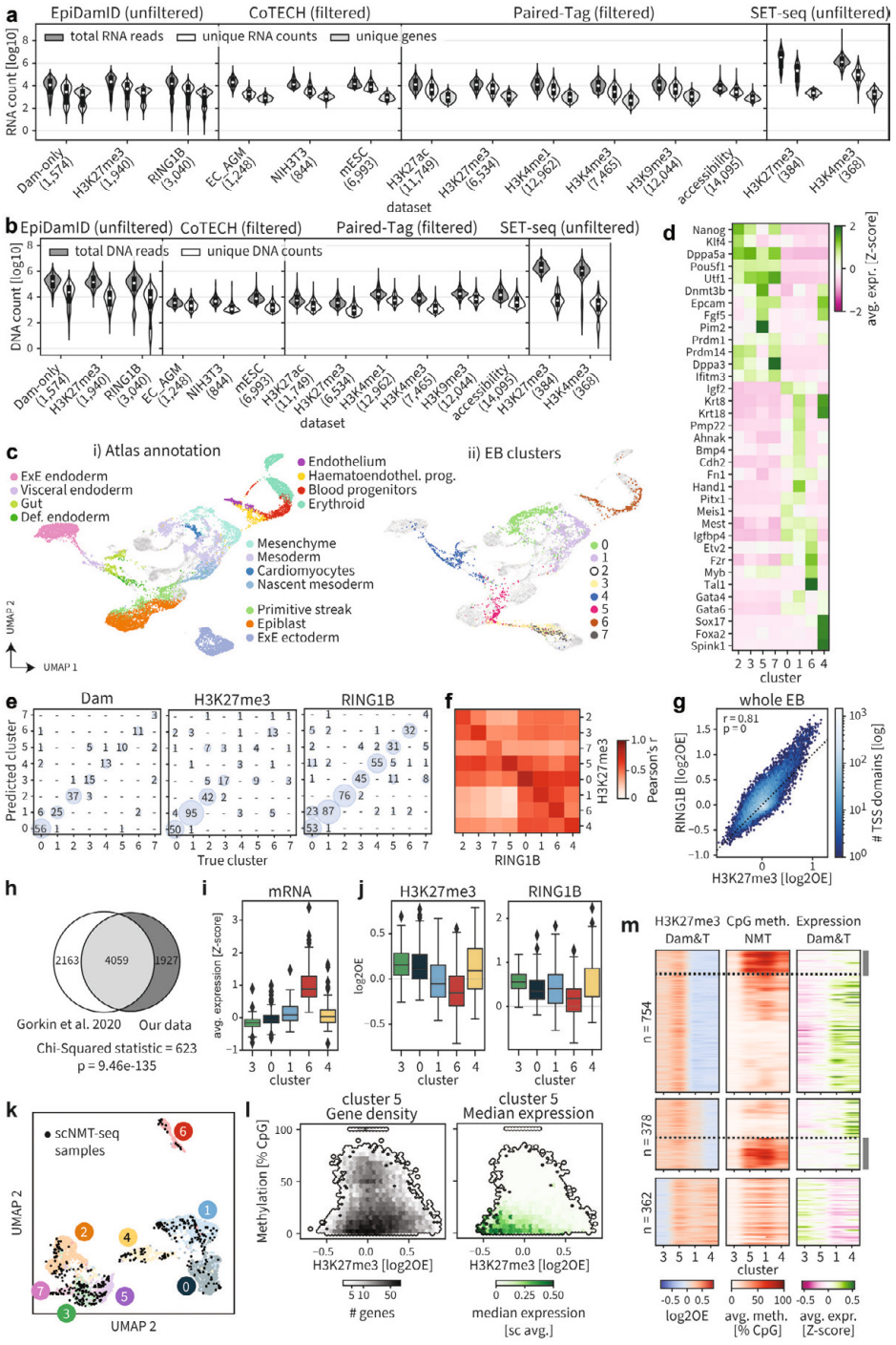


Figure S3: Validation and characterization of scDam&T-seq data in mouse embryoid bodies
a-b, Overview of the RNA and DNA outputs for a number of datasets generated by recent single-cell

multimodal omics techniques. a: The number of raw RNA-derived reads, unique transcripts and unique genes. b: The number of raw DNA-derived reads and unique counts. The included techniques are CoTECH (Xiong et al., 2021), Paired-Tag (Zhu et al., 2021) and SET-seq (Sun et al., 2021). The EpiDamID (scDam&T-seq) data shows the statistics of the embryoid body (EB) dataset. Some techniques show only the statistics of cells that passed quality thresholds ("filtered"), while others show the statistics of all obtained cells ("unfiltered"). The labels on the x-axis indicate the name of the various datasets, with the number of samples shown in paratheses. **c**, UMAPs of samples based on the integration of our EB transcription data with single-cell RNA-seq mouse embryonic data (Pijuan-Sala et al., 2019), colored by reference-annotated cell type (i) and EB-annotated cluster (ii). For atlas integration, the day 0 (i.e., mESC) time point was excluded. **d**, Average expression of known marker genes. Expression was standardized over single-cells and the per-cluster average was computed. **e**, Confusion plots showing the performance of the LDA classifier during training, for each construct. **f**, Pearson correlation between the combined H3K27me3 and RING1B DamID signal at the TSS of all genes per transcriptional cluster. **g**, Correlation of combined H3K27me3 and RING1B DamID signal at the TSS of all genes. Data of all single-cell samples passing DamID thresholds was combined for each construct. The correlation was computed using Pearson's correlation. **h**, Overlap between a published set of PRC targets during mouse development[66] and our PRC targets. Only genes represented in both datasets could be compared. Significance of the overlap was computed with a Chi-squared test. **i**, Boxplots showing the expression (averaged Z-score) of genes identified as significantly upregulated in cluster 6. **j**, Boxplots showing the H3K27me3 (left) and RING1B (right) DamID signal at the TSS of the subset of genes shown in **i** that are PRC targets. **k**, UMAPs of samples based on the integration of our EB transcription data with the transcriptional readout of the EB scNMT-seq data generated by Argelaguet et al.[67]. EpiDamID samples are colored by the transcriptional clusters determine previously; scNMT-seq samples are indicated in black. **l**, Relationship between promoter CpG methylation, promoter H3K27me3 enrichment and gene expression of all genes in cells belonging to cluster 5 (epiblast-like). The left plot shows the relationship between promoter CpG methylation (+/- 2 kb around TSS) and H3K27me3 enrichment (-5 kb/+3 kb around TSS) for all genes. The right plot shows the same relationship, but the color scale indicates the median expression of genes in each region of the plot. **m**, Heatmaps indicating the promoter H3K27me3 enrichment, promoter CpG methylation and gene expression for three groups of PRC targets with variable H3K27me3 enrichment. Rows are genes; columns are transcription clusters. Enrichment is shown for the 4 clusters that contained sufficient scNMT-seq samples (cluster 3: 31 cells; cluster 5: 21 cells; cluster 1: 37 cells; cluster 4: 43 cells). Genes are sorted by hierarchical cluster based on their CpG methylation levels. Examples of genes where H3K27me3 and CpG methylation complementary repress genes are indicated with a dotted line and a grey box.

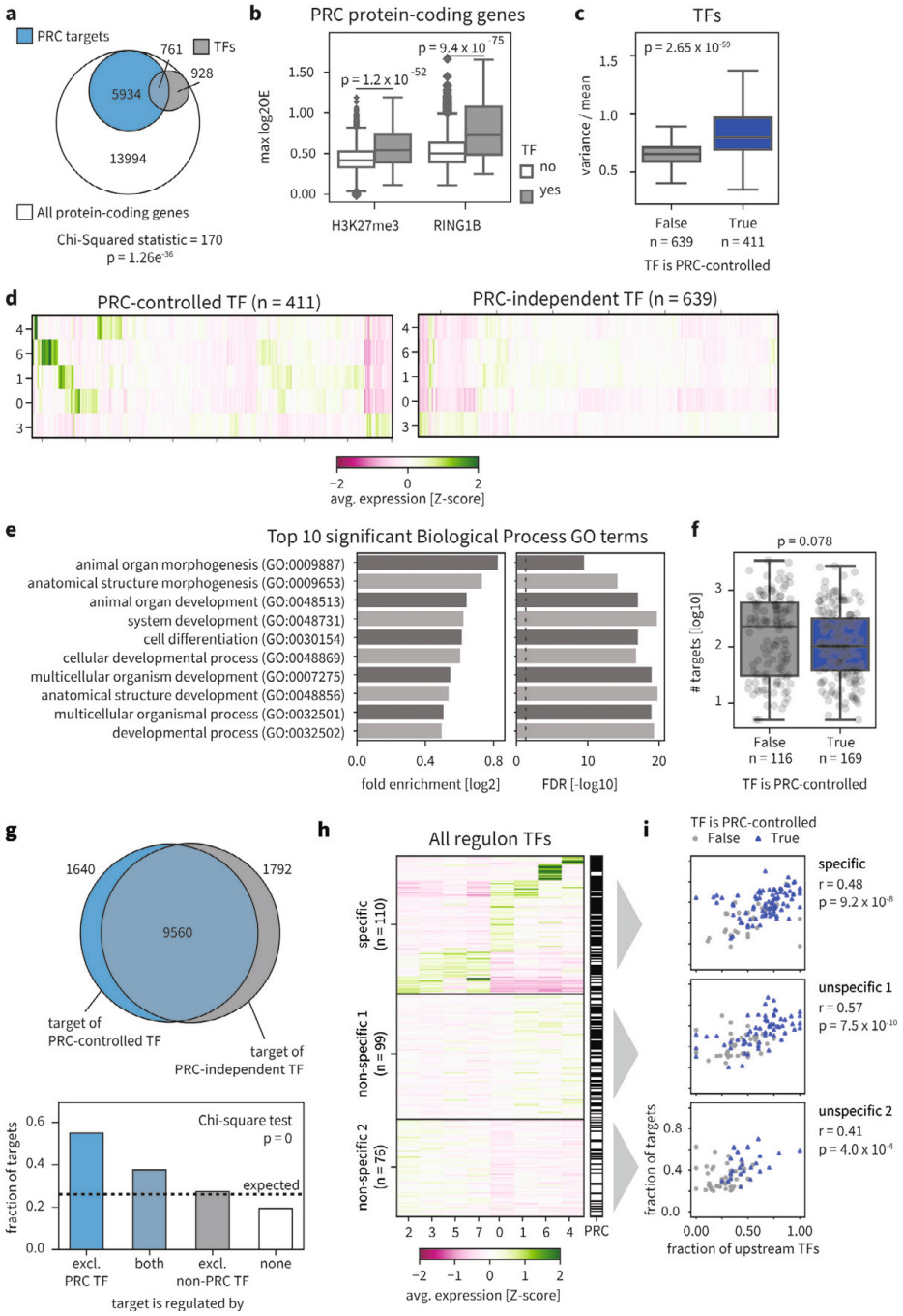


Figure S4: Characterization of the Polycomb-regulated regulatory network
a, Venn diagram showing the overlap between PRC-controlled protein-coding genes (blue) and transcription factors (TF) (grey) in the context of all protein-coding genes (white). The significance of the overlap between PRC targets and TFs was computed using a Chi-squared test. **b**, Boxplots showing the maximum observed

H3K27me3 and RING1B DamID signal across transcriptional clusters for PRC-controlled TFs (grey) and the remaining PRC-controlled protein-coding genes (white). The significance of the difference between TFs and other genes was tested with a two-sided Mann-Whitney-U test. **c**, Quantification of variability in gene expression of PRC-regulated and PRC-independent TFs (only expressed genes are included). Boxplots show variance over mean across all single cells. Significance was computed using a two-sided Mann-Whitney U test. **d**, Clustered heatmaps showing mRNA expression (averaged Z-score) per cluster, of Polycomb-regulated TFs (left) and Polycomb-independent TFs (right). Only expressed genes are included in this plot. **e**, The ten most significant Biological Process GO terms between PRC-controlled and PRC-independent TFs. **f**, Number of targets of each regulon TF, split by whether or not the TF is PRC-regulated. The significance of the difference between the two groups was tested with a two-sided Mann-Whitney U test. **g**, Top: Venn diagram displaying the overlap between genes that are targets of a PRC-controlled TF (blue) and genes that are targets of a PRC-independent TF (grey). Bottom: Bar plot showing the fraction of targets in each category that is PRC-regulated. The dotted line indicates the expected fraction, i.e., the fraction of all genes that is a PRC target. A Chi-square test was performed to evaluate whether the deviation from the expected frequencies is significant. **h**, Clustered heatmap showing mRNA expression (averaged Z-score) per cluster, of all regulon TFs, grouped by lineage-specific or non-specific genes. TFs are annotated as PRC-controlled (black) or PRC-independent (white). **i**, Scatter plot showing the relationship between the fraction of Polycomb-controlled targets and regulators of a regulon TF. Regulon TFs that are PRC controlled are indicated in blue; regulon TFs that are PRC independent are indicated in grey. Regulon TFs are split based on the groups indicated in **h**. Correlation was computed using Pearson's correlation.

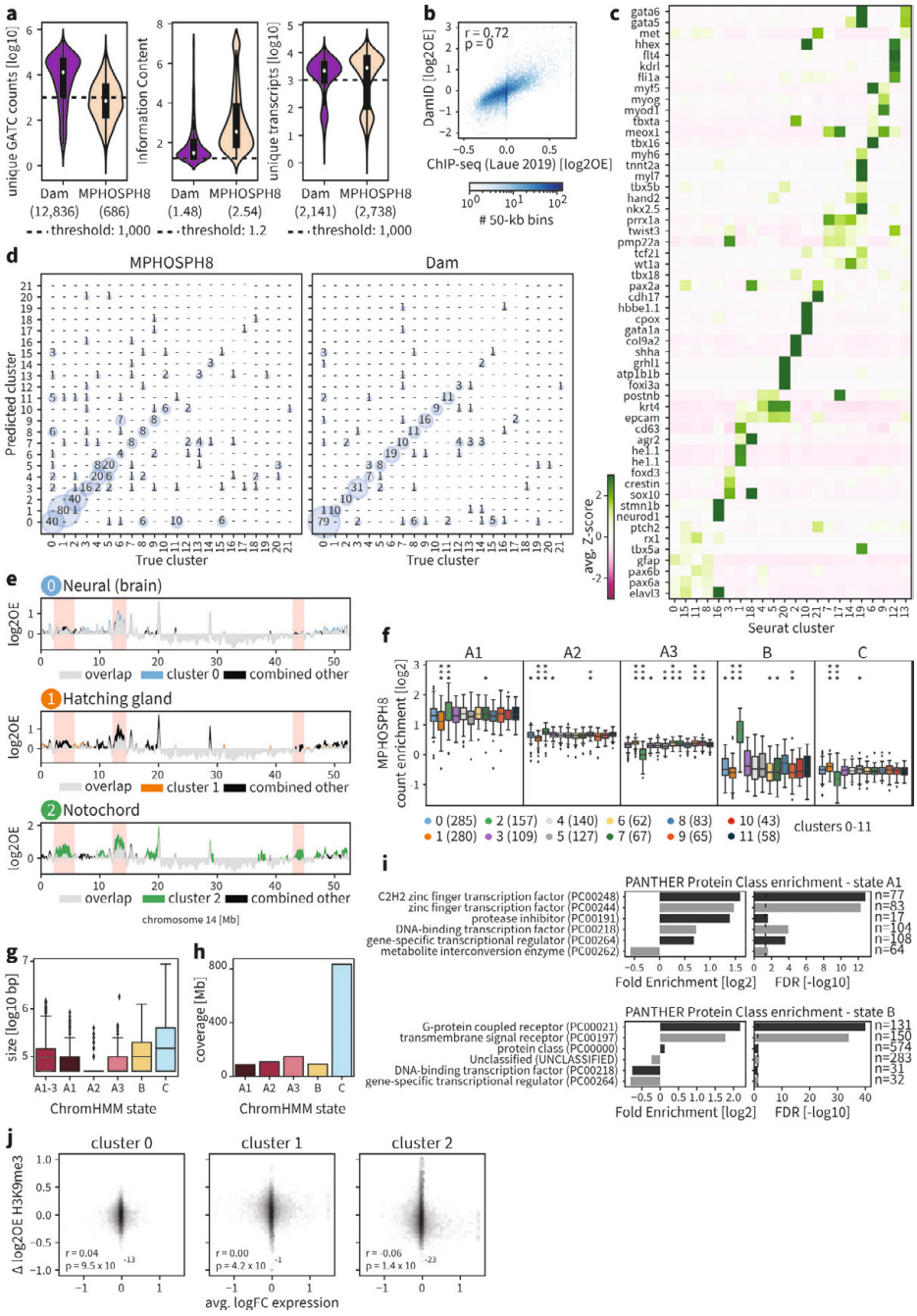


Figure S5: Characterization of transcriptomic clusters and associated genomic H3K9me3 enrichments
a, Violin plots showing the total number of unique GATC counts, the information content (IC) and total number of unique transcripts obtained for all cells in the zebrafish dataset. **b**, Comparison of our data with a published H3K9me3 ChIP-seq dataset of the 6-hpf zebrafish embryo (Laue et al., 2019). All single-

cell MPHOSPH8 and Dam samples were combined to generate an in silico whole-embryo data set; DamID data is the log2OE of MPHOSPH8 signal over Dam is shown; ChIP-seq is the log2OE of H3K9me3 over input control. The correlation was computed using Pearson's correlation. **c**, Expression of marker genes over all clusters, ordered by cell type. The average single-cell Z-scores are shown. **d**, Confusion plots showing the performance of the LDA classifier during training, for each construct. **e**, Genomic H3K9me3 signal over chromosome 14. For clusters 0-2, the cluster-specific signal (color) is compared to the combined signal from all other clusters (black). Each set indicates the overlay, where overlapping regions are colored grey. **f**, Boxplots showing the enrichment of counts within genomic regions belonging to each of the five ChromHMM states for all cells belonging to transcriptional clusters 0-11. Count enrichment was computed as the fraction of GATC counts that fell within the regions, relative to the total fraction of genomic GATC positions inside these domains. Per state, the count enrichment of a cluster was compared to the enrichment of cells in all other clusters using a two-sided Mann-Whitney-U test. *** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.1$; no indication means the result was insignificant. **g**, Distribution of domain sizes per ChromHMM state and for states A1-3 combined. **h**, Total genomic coverage per ChromHMM state. **i**, PANTHER protein-class enrichments (Mi et al., 2013) for genes found in state A1 (top) and B (bottom). **j**, Plot displaying the relationship between differential gene expression and differential H3K9me3 enrichment. The x-axis shows the average log-foldchange in gene expression of cells in one cluster relative to all other cells; the y-axis shows the differential log2OE H3K9me3 at these genes of one cluster relative to all other cells. H3K9me3 at a gene was measured as the log2OE value of the 50-kb genomic window containing the TSS of the gene. The relationship between the variables was tested with a Pearson's correlation test.

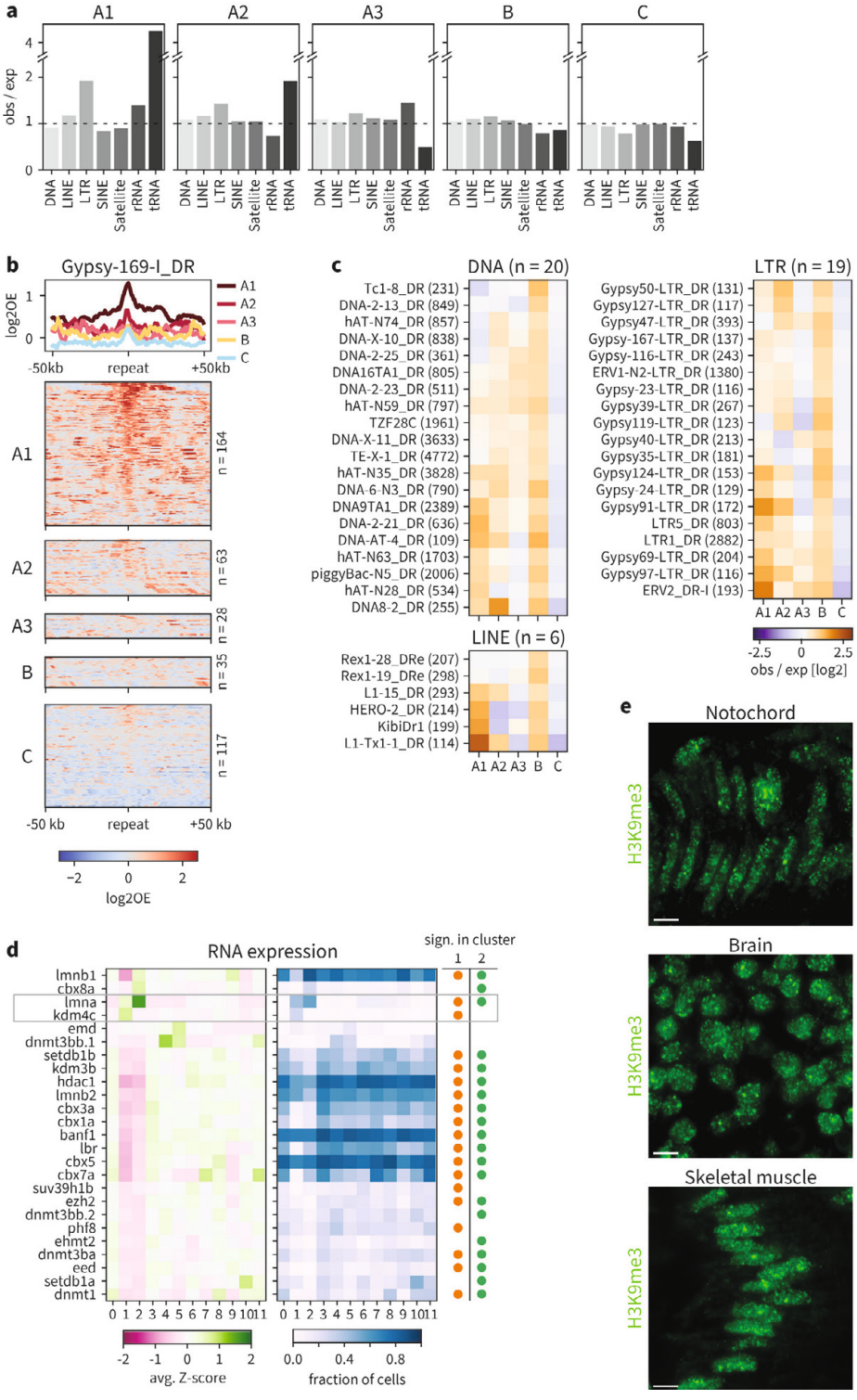

a
A1
A2
A3
B
C
obs / exp
DNA
LINE
LTR
SINE
Satellite
rRNA
tRNA
b
Gypsy-169-I_DR
log2OE
-50kb
repeat
+50kb
n = 164
n = 63
n = 28
n = 35
n = 117
log2OE
c
DNA (n = 20)
Tc1-8_DR (231)
DNA-2-13_DR (849)
hAT-N74_DR (857)
DNA-X-10_DR (838)
DNA-2-25_DR (361)
DNA16TA1_DR (805)
DNA-2-23_DR (511)
hAT-N59_DR (797)
TZF28C (1961)
DNA-X-11_DR (3633)
TE-X-1_DR (4772)
hAT-N35_DR (3828)
DNA-6-N3_DR (790)
DNA9TA1_DR (2389)
DNA-2-21_DR (636)
DNA-AT-4_DR (109)
hAT-N63_DR (1703)
piggyBac-N5_DR (2006)
hAT-N28_DR (534)
DNA8-2_DR (255)
LINE (n = 6)
Rex1-28_DRe (207)
Rex1-19_DRe (298)
L1-15_DR (293)
HERO-2_DR (214)
KibiDr1 (199)
L1-Tx1-1_DR (114)
LTR (n = 19)
Gypsy50-LTR_DR (131)
Gypsy127-LTR_DR (117)
Gypsy47-LTR_DR (393)
Gypsy-167-LTR_DR (137)
Gypsy-116-LTR_DR (243)
ERV1-N2-LTR_DR (1380)
Gypsy-23-LTR_DR (116)
Gypsy39-LTR_DR (267)
Gypsy119-LTR_DR (123)
Gypsy40-LTR_DR (213)
Gypsy35-LTR_DR (181)
Gypsy124-LTR_DR (153)
Gypsy-24-LTR_DR (129)
Gypsy91-LTR_DR (172)
LTR5_DR (803)
LTR1_DR (2882)
Gypsy69-LTR_DR (204)
Gypsy97-LTR_DR (116)
ERV2_DR-I (193)
A1 A2 A3 B C
obs / exp [log2]
d
RNA expression
sign. in cluster
lmnb1
cbx8a
lmna
kdm4c
emd
dnmt3bb.1
setdb1b
kdm3b
hdac1
lmnb2
cbx3a
cbx1a
banf1
lbr
cbx5
cbx7a
suv39h1b
ezh2
dnmt3bb.2
phf8
ehmt2
dnmt3ba
eed
setdb1a
dnmt1
avg. Z-score
fraction of cells
e
Notochord
Brain
Skeletal muscle
H3K9me3

Figure S6: Characterization of repeat content, expression of chromatin factors and nuclear localization of H3K9me3 chromatin

a, Enrichment of repeats per class for all ChromHMM states. Enrichment is computed as the observed number of repeats within a state relative to the expected number based on the genome coverage of each state. **b**, H3K9me3 enrichment at Gypsy-169-I_DR repeats across ChromHMM states. The heatmaps show the enrichment per individual repeat instance, while the line plot shows the average enrichment per state. **c**, Enrichment of repeats in ChromHMM states as in Figure 5I. Only repeats having at least 100 copies throughout the genome and an enrichment $^{3}1.5$ in state B are included. Enrichment is computed as the observed number of repeats in a stated compared to the expected number based on the genome coverage of that state. **d**, RNA expression of various chromatin factors across clusters 0-11. The left heatmap shows the average single-cell expression (Z-score); the right heatmaps shows the fraction of cells in each cluster with at least one transcript of each gene. Only factors that are expressed in at least 10% of cells of at least one cluster are shown. **e**, Representative images of H3K9me3 staining in cryosections of notochord (left), brain (middle), and skeletal muscle (right) in 15-somite embryos. Scale bars represent 4 μm.

Supplementary Tables

Table S1: **Overview of EpiDamID constructs used in RPE-1 DamID experiments**

targeting domain	domain type	target	orientation	promoter	Dam type	# samples passing thresholds
CBX1	protein domain (dimer)	H3K9me3	Dam-X	HSP	DamWT	5
CBX1	protein domain (trimer)	H3K9me3	Dam-X	PGK	DamWT	4
CBX1	full protein	H3K9me3	X-Dam	HSP	DamWT	3
CBX7	protein domain (trimer)	H3K27me3	Dam-X	HSP	DamWT	5
CBX7	protein domain (trimer)	H3K27me3	Dam-X	PGK	DamWT	4
untethered Dam	untethered Dam	accessible chromatin	NA	HSP	DamWT	7
untethered Dam	untethered Dam	accessible chromatin	NA	PGK	DamWT	2
untethered Dam	untethered Dam	accessible chromatin	NA	PGK	Dam126	2
H3K27me3	mintbody	H3K27me3	Dam-X	HSP	DamWT	5
H3K27me3	mintbody	H3K27me3	Dam-X	PGK	DamWT	4
H3K27me3	mintbody	H3K27me3	Dam-X	PGK	Dam126	3
H3K27me3	mintbody	H3K27me3	X-Dam	HSP	DamWT	3
H3K27me3	mintbody	H3K27me3	X-Dam	PGK	DamWT	2
H3K27me3MUT	mintbody	accessible chromatin	Dam-X	PGK	DamWT	1
H3K27me3MUT	mintbody	accessible chromatin	Dam-X	PGK	Dam126	1
H3K27me3MUT	mintbody	accessible chromatin	X-Dam	PGK	DamWT	2
H3K9ac	mintbody	H3K9ac	Dam-X	PGK	DamWT	3
H4K20me1	mintbody	H4K20me1	Dam-X	HSP	DamWT	1
H4K20me1	mintbody	H4K20me1	Dam-X	PGK	DamWT	4
H4K20me1	mintbody	H4K20me1	Dam-X	PGK	Dam126	2
H4K20me1	mintbody	H4K20me1	X-Dam	HSP	DamWT	4
MPHOSPH8	protein domain (trimer)	H3K9me3	Dam-X	HSP	DamWT	5
POLR2F	full protein	PolII binding	Dam-X	PGK	DamWT	5

Table S1: **Continued**

targeting domain	domain type	target	orientation	promoter	Dam type	# samples passing thresholds
RING1B	full protein	RING1B binding	Dam-X	HSP	DamWT	3
TAF3	protein domain (trimer)	H3K4me3	Dam-X	HSP	DamWT	1
TAF3	protein domain (trimer)	H3K4me3	Dam-X	PGK	DamWT	2

Table S2: **Metadata and quality metrics of all single-cell samples of the ESC, EB and zebrafish experiments**

Available in online version: https://doi.org/10.1016/j.molcel.2022.03.009

CHAPTER 5

The role of heterochromatin in 3D genome organization during preimplantation development

Franka J. Rang[1,2], Jop Kind[1,2,3,4]* and Isabel Guerreiro[1,2]*

1: Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, the Netherlands
2: Oncode Institute, the Netherlands
3: Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, Houtlaan 4, 6525 XZ Nijmegen, the Netherlands
4: Lead contact
*Correspondence: J.K. (j.kind@hubrecht.eu) and I.G. (i.guerreiro@hubrecht.eu)

Cell Reports, 2023

Abstract

During the early stages of mammalian development, the epigenetic state of the parental genomes is completely reprogrammed to give rise to the totipotent embryo. An important aspect of this remodeling concerns the heterochromatin and the spatial organization of the genome. While heterochromatin and genome organization are intricately linked in pluripotent and somatic systems, little is known about their relationship in the totipotent embryo. In this review, we summarize the current knowledge on the reprogramming of both regulatory layers. In addition, we discuss available evidence on their relationship and put this in the context of findings in other systems.

Introduction

At the beginning of mammalian development, the sperm and oocyte fuse to give rise to the totipotent zygote. Since the oocyte and sperm are both mature cell types with preexisting and vastly different epigenomes, all layers of epigenetic regulation undergo extensive remodeling to allow for the development of all embryonic and extra-embryonic cell types[1]. Moreover, very soon after fertilization, the embryo undergoes zygotic genome activation (ZGA), in which transcription from the zygotic genome initiates and maternally provided mRNA starts to be degraded. In mouse development, a minor wave of ZGA takes place as early as the zygotic stage, while major ZGA occurs one cleavage later at the 2-cell stage[2]. By this stage, mechanisms of transcriptional regulation have to be in place to ensure the timely activation of the correct subset of genes. The complete remodeling of the epigenome in preimplantation development offers a unique system to study the interactions between different epigenetic layers as they emerge and change. Moreover, the presence of the distinct maternal and paternal epigenetic states within the same cell provides a unique side-by-side comparison of how the initial chromatin state influences its subsequent dynamics. Studying these processes will thus not only provide insight into the foundational events of early development and epigenetic reprogramming, but may also reveal fundamental principles of epigenetic interactions that hold true in any system.

Epigenetic regulation works by making DNA accessible or inaccessible to proteins such as transcription factors and the transcriptional machinery. The inaccessible and inactive fraction of the genome, referred to as heterochromatin, is physically segregated from the accessible and active fraction, or euchromatin: while euchromatin resides in the nuclear interior, the densely compacted heterochromatin tends to locate at the nuclear periphery and around nucleoli[3,4]. This inherent link between heterochromatin and spatial organization was already established decades ago in microscopy studies and research since has further solidified this finding[3,4]. Classically, heterochromatin has been divided into two types: constitutive and facultative. The two types are associated with different types of proteins and are also spatially organized in different ways. Constitutive heterochromatin is mostly consistent across cell types and covers regions rich in repetitive elements, such as the major and minor satellite repeats at centromeres. Facultative heterochromatin, on the other hand, is cell-type specific and is associated with the repression of developmental genes[3]. Both types of heterochromatin play an important role in maintaining cell identity and genome stability through repression of genes and repeats[5-9], but have distinct structural organizations with major differences in spatial distribution, compaction level, and long-range contacts[3]. Despite the many links between heterochromatin and genome organization, relatively little is known about their interactions during the extensive reprogramming that takes place in preimplantation development. Insight into this relationship could further our understanding of epigenetic remodeling during the establishment of totipotency, as well as the complex interactions between different modes of genome regulation in general.

In this review, we summarize and interconnect the current knowledge on the remodeling of the heterochromatin and 3D organization during mammalian preimplantation development. Since most of the research to date has been performed in mouse, we mainly focus on this model organism. In addition, we describe the links between these two modes of genome regulation that have been established in early development and how they relate to observations in other systems. Finally, we suggest avenues of further research to advance our understanding of the relationship between heterochromatin and 3D organization during preimplantation development.

Reprogramming of heterochromatin and 3D organization in mouse preimplantation development

Facultative heterochromatin

The most characteristic facultative heterochromatin marks are the mono-ubiquitination of H2A lysine-119 (H2AK119ub1) and the tri-methylation of H3 lysine-27 (H3K27me3), catalyzed by Polycomb repressive complex 1 (PRC1) and PRC2, respectively. Both complexes exist in multiple forms and they have been shown to work both upstream and downstream of one another[10-13]. In pluripotent stages of development, starting in the blastocyst, H2AK119ub1 and H3K27me3 marks largely overlap at the promoters of developmental genes and serve to repress them. A large fraction of these promoters is also marked by the active histone post-translational modification (PTM) H3K4me3, creating a so-called bivalent domain that represents a reversible repressive state to prevent premature gene activation[14,15]. Once cells commit to a certain lineage, genes specific to the lineage lose H3K27me3/H2AK119ub1 and become active. Conversely, at many genes specific to other cell types, H3K4me3 is lost, H3K27me3 domains broaden, and a permanent repressive state is achieved[16,17].

While in most systems H3K27me3 and H2AK119ub1 are largely located at genic regions, these two marks present an entirely different distribution during mouse preimplantation development and are independently remodeled post-fertilization (Fig. 1i-ii). These atypical profiles arise already during oocyte maturation, where both H3K27me3 and H2AK119ub1 are progressively laid down as unusually broad domains in intergenic regions[18-21]. Consequently, a much larger fraction of the genome is covered by H3K27me3 and H2AK119ub1 in oocyte (~35%) compared to later developmental stages with canonical profiles (<5%). The broad domains of H3K27me3/H2AK119ub1 have a strong overlap with regions with intermediate levels of DNA methylation, i.e. partially methylated domains (PMDs), while being excluded from fully methylated domains (FMDs)[18]. Interestingly, H3K36me3 has been shown to overlap significantly with DNA methylation in oocytes and anti-correlate with H3K27me3. Depletion of H3K36me3 via the knockout (KO) of methyltransferase Setd2 resulted in expansion of H3K27me3 distributions, indicating a role in H3K36me3 in shaping the distribution of the oocyte Polycomb marks[22]. In addition to these atypical broad domains, H3K27me3 and H2AK119ub1 are maintained at promoters of known Polycomb targets[18,20,21,23]. In sperm, most histones have been replaced by protamines to facilitate tight packaging of the chromatin[24,25],

but the remaining histones appear to retain canonical distributions of H3K27me3 and H2AK119ub1[18,20,21].

After fertilization, the broad H3K27me3 and H2AK119ub1 domains are inherited from the oocyte on the maternal allele, while the marks on the paternal allele are rapidly removed and de novo enrichment of start to appear by the late zygotic stage[18,20,21]. Paternal deposition of H3K27me3 is dependent on the activation of the catalytic PRC2 subunit EZH2 via phosphorylation by CDK1 during the G2/M transition[26]. The newly formed paternal H3K27me3 and H2AK119ub1 modifications form very broad domains of low enrichment, mostly in intergenic regions[18-20]. On the maternal allele, H3K27me3 is lost from the promoters of canonical Polycomb targets post-fertilization and is only fully recovered after implantation[18,21,23]. However, Polycomb target genes do retain H2AK119ub1, which likely suffices for gene repression, as they become upregulated upon PRC1 catalytic component disruption but not in the absence of H3K27me3[20,21]. Although the two Polycomb marks display similar distributions in gametes and the zygote, their profiles rapidly start to diverge in subsequent cleavage stages. While H3K27me3 retains its parental asymmetry and broad domains, the two alleles have largely equalized with respect to H2AK119ub1 by the end of the 2-cell stage and start to more closely resemble canonical profiles as seen in mouse embryonic stem cells (mESCs) and the blastocyst[19-21].

Some attempts have been made to elucidate the interdependence of H3K27me3 and H2AK119ub1 during early development[19,20]. Conditional KO of Eed, a core component of PRC2, results in the loss of H3K27me3 in oocytes and in preimplantation embryos until ZGA. However, H2AK119ub1 is only affected at a subset of non-canonical imprinting loci, which temporarily lose H2AK119ub1 in wild-type (WT) embryos, but fail to reestablish the mark in the absence of H3K27me3[20]. Similarly, acute depletion of H2AK119ub1 in early embryos does not lead to big changes in H3K27me3 at the 4-cell stage[20]. However, maternal KO of PRC1 subunits and consequent loss of H2AK119ub1 did lead to a decrease of H3K27me3 at a subset of genes in oocytes and early embryos. These results suggest that PRC1 may directly or indirectly work upstream of PRC2 in this system[19].

Constitutive heterochromatin

Several histone modifications are associated with constitutive heterochromatin, including H3K9me2/3, H4K20me2/3, and H3K64me3. H3K9me3, the most extensively studied of these marks, plays a central role in mediating typical constitutive heterochromatin features. Indeed, its interaction with heterochromatin protein 1a (HP1a) results in phase separation and chromatin compaction[32-34], and also plays a role in localization of heterochromatin to the nuclear periphery[35]. In mammals, methylation of H3K9 is mediated by six histone methyltransferases (SUV39H1, SUV39H2, SETDB1, SETDB2, G9A and GLP) that are only partially redundant and each work in different genomic contexts, as reviewed in Padeken et al. [36]. Interestingly, some of these histone methyltransferases mediate H3K9me enrichment in genic regions to ensure cell-type specific repression[8,9], thus functioning as a form of facultative heterochromatin.

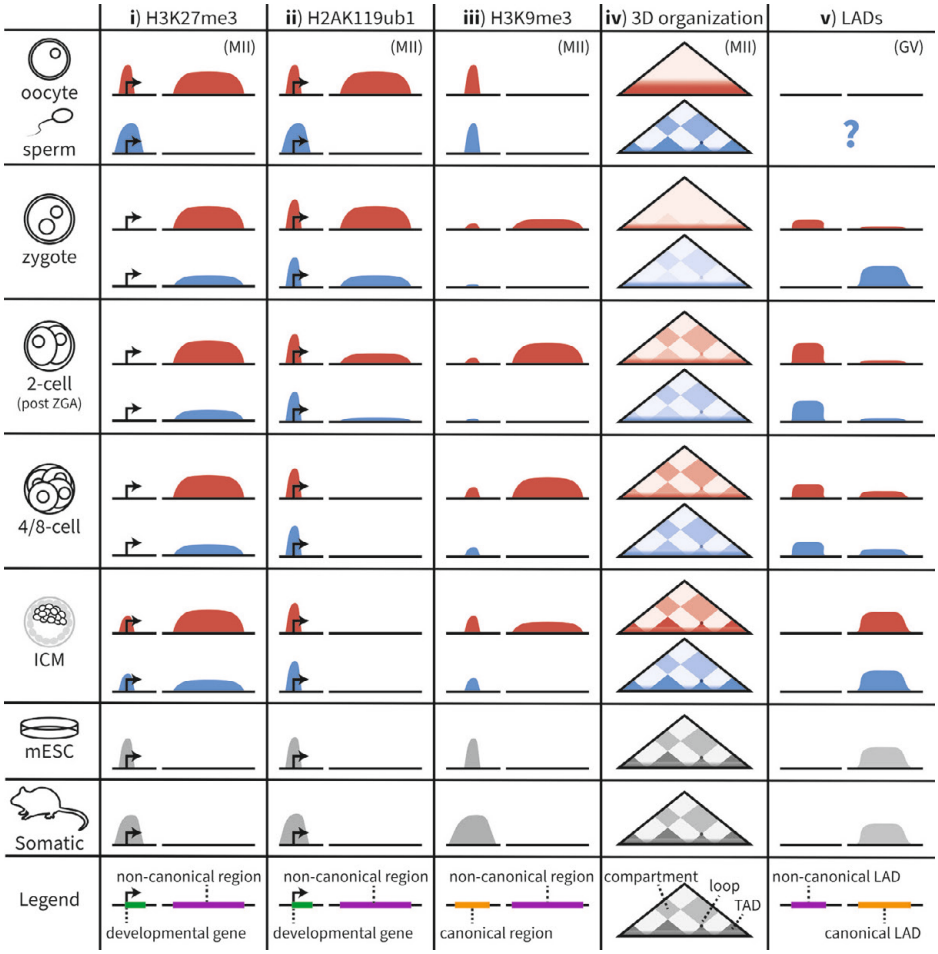


Figure 1: Schematic representation of heterochromatin and genome organization reprogramming during mouse early development

The cartoons display the reprogramming of H3K27me3 (**i**), H2AK119ub1 (**ii**), H3K9me3 (**iii**), 3D organization (**iv**), and LADs (**v**). The legend at the bottom provides information on the type of regions that are affected. Cartoons are based on the descriptions of ChIP-seq data for H3K27me3[18,23], H2AK119ub1[19,20], and H3K9me3[27]; on Hi-C data for 3D organization[28-30]; and on DamID data for LADs[31]. Since the genomics data do not cover repetitive regions, such as centromeric, pericentromeric and telomeric regions, these are not included in the representation.

During preimplantation development, constitutive heterochromatin and its associated histone PTMs are extensively remodeled. Directly after fertilization, there is a strong parental asymmetry in constitutive heterochromatin marks, which are clearly present in the maternal pronucleus, while remaining largely undetected in the paternal pronucleus[37-40]. As is the case for H3K27me3 and H2AK119ub1, the maternal histone marks appear to be inherited from oocytes, while paternal H3K9me2/3 starts to be established in late zygotes by SUV39H2[41].

Recently, the distribution of H3K9me3 in gametes and in early embryos was profiled using chromatin immunoprecipitation and sequencing (ChIP-seq)[27] (Fig. 1iii). Although H3K9me3 could not be detected in the early paternal pronucleus by immunofluorescence microscopy, the ChIP-seq data revealed low levels of paternal H3K9me3 at the PN3 zygote stage. This paternal signal shows some overlap with sperm signal, suggesting that a limited amount of H3K9me3 may be inherited from the father. Interestingly, the paternal H3K9me3 that is laid down de novo during the zygote stage appears to lack repressive qualities, as a knockdown of the responsible histone methyltransferase results in a downregulation of affected genes at the 2-cell stage[41]. This surprising result suggests that paternal H3K9me3 at these genes may be activating rather than repressive. In the maternal pronucleus, the inherited H3K9me3 is remodeled as well, with a loss of H3K9me3 at the promoters of a set of developmental genes and a gain in intergenic regions relative to oocytes. So far, there is no evidence that H3K9me3 on the maternal allele is non-repressive at this stage and, in fact, the genes that lose H3K9me3 in zygotes compared to oocytes are enriched for ZGA genes that are expressed at the 2-cell stage.

Interestingly, the H3K9me3 domains that are gained post-fertilization strongly overlap with the maternally inherited H3K27me3[27], while in pluripotent and differentiated cells these marks only show a limited overlap at gene promoters[8,17,42]. The H3K9me3 enrichment in H3K27me3 domains starts to decrease around the morula stage and is completely lost after implantation, thus slightly preceding the loss of non-canonical H3K27me3[27]. Meanwhile, starting at the 4-cell stage, H3K9me3 gains enrichment at canonical sites, such as long terminal repeat (LTR) retrotransposons. For a number of LTRs, the gain of H3K9me3 at the 4-cell stage coincides with their downregulation, implying that H3K9me3 may play a role in the timely repression of repeats during development[27].

Together, these results show that H3K9me3 is extensively remodeled in the early embryo and displays several unusual characteristics, such as a lack of repression by newly gained paternal domains and an extensive overlap with H3K27me3 at non-canonical sites. While the relevance of these features is still unclear, H3K9me3 does seem to play an important role in the temporal regulation of repeat expression.

3D genome organization

Heterochromatin is intimately linked to the spatial organization of the genome. One important aspect of this organization is the 3D positioning of genomic regions relative to one another in the nucleus. The genome is organized in a multi-layered manner. At the highest level of organization, chromosomes form distinct territories within the nucleus, while interactions between different chromosomes remain limited[43,44]. The chromatin further partitions into two compartments: the active compartment A is associated with higher levels of gene expression, active histone marks and GC-content, while the opposite holds true for the inactive compartment B[44]. Within these compartments, domains of preferential interactions occur, referred to as topologically associating domains (TADs)[45,46]. Mechanistically, TADs are formed by the continuous process of loop-extrusion, whereby the ring-shaped protein cohesin

continually extrudes a loop of DNA until it is halted at a boundary element or dissociates from the DNA[47-51]. The process of loop extrusion within domain boundaries likely plays a role in bringing together regulatory elements with their target genes, while excluding interaction with off-target genes outside the boundaries[52]. The most prominent boundary element in vertebrates is the insulator protein CCCTC-binding factor (CTCF). The binding sites of CTCF determine regions across which loops cannot be extruded and, consequently, pair-wise interactions across boundaries are rarer. In bulk methods such as Hi-C, this phenomenon thus gives rise to contact domains that often have focal points of increased interactions at boundary elements where loops frequently stall[53].

As is the case for heterochromatin, the 3D organization of the genome is extensively remodeled during preimplantation development (Fig. 1iv). The chromatin from the maternal and paternal gametes exists in entirely different organizational states at the moment of fertilization. Since the oocyte is stalled at metaphase of meiosis II prior to fertilization, the maternal chromosomes are strongly condensed and structurally similar to mitotic chromosomes, lacking loops, TADs and compartments, while being enriched for interactions at 1-7 Mb[30]. Meanwhile, the paternal genome is packaged tightly in the sperm nucleus, which has a volume over ten times smaller than that of somatic cells[24,25,54]. Despite this extreme compaction, the overall 3D organization of sperm is similar to mESCs and somatic cells, albeit somewhat enriched for long-range interactions[28,55,56]. Following fertilization, both parental genomes are rapidly remodeled to a state with little consistent 3D architecture in the zygote embryo. At this stage, TADs and loops are barely visible in the Hi-C interaction matrices[28,30] and only become evident when averaging across multiple sites[29,57]. While the paternal genome shows weak compartmentalization, the maternal genome lacks compartments[28-30]. Nevertheless, the paternal chromatin has fewer distal (>2 Mb) interactions, suggesting that it may be in a more relaxed state than the maternal genome during the zygote stage[30].

The allelic differences present in zygote are largely resolved in the 2-cell embryo. In addition, starting at the end of the 2-cell stage (post ZGA), all levels of chromatin organization get progressively stronger and have been completely established by the time of implantation[28,30]. Despite the conspicuous concurrence with transcriptional activation, transcription itself does not appear to be necessary to consolidate the 3D genome architecture[28,30]. Rather, blocking the process of DNA replication appears to prevent further establishment of chromatin organization[28]. Despite these insights, it is still largely unclear what factors cause the lack of organization in the early embryo and which factors subsequently are responsible for its reestablishment.

Lamina Associated Domains

Another important aspect of genome organization involves the spatial positioning of DNA within the nucleus. Most exemplary of this is the segregation of heterochromatic chromatin at the nuclear lamina (NL), a filamentous network at the inner nuclear membrane. Regions of the genome associated with the NL are referred to as lamina-associated domains (LADs).

These broad regions have a median size of ~500 kb and are characterized by a high density of long interspaced nuclear element (LINE) repeats, low gene density, and low levels of gene expression[58,59]. LADs are frequently enriched for H3K9me2/3[42,58,60,61] and to a limited extent for H3K27me3[42,58,62]. In line with their heterochromatic nature, LADs show a strong correspondence with the inactive B compartment identified by Hi-C[53,63].

As is the case of the 3D interactions, NL contacts are extensively reorganized during preimplantation development[31] (Fig. 1v). In zygotes, the paternal genome forms well-defined LADs, while the maternal genome shows weaker and more inconsistent NL interactions in regions with features atypical for LADs. Interestingly, LADs seem to be largely absent in the developing oocyte (germline vesicle [GV]), implying that LADs are established de novo for the maternal genome[31]. At the zygote stage, the genome contacts the NL but electron spectroscopic imaging shows an absence of compacted chromatin in these regions[64], suggesting that these LADs exist in a relatively decondensed state unique to the zygote stage.

The parental asymmetry in NL association is strongest in the zygote and starts to diminish in subsequent stages and is only fully resolved by the time of implantation[31]. At the 2-cell stage, the maternal LADs strengthen, while the paternal LADs are reorganized to resemble the maternal LADs more closely. At this stage, many LADs considered to be constitutive in somatic cells have dissociated from the NL. These regions only partially relocate to the NL at the 8-cell stage, but are fully recovered in the inner cell mass (ICM) of the blastocyst[31]. To date, such extensive reprogramming and loss of constitutive LADs has not been observed in any other system. The extreme remodeling and less condensed state of LADs are unique to the preimplantation embryo, suggesting that NL association may be regulated differently during these early stages of development. However, the exact mechanisms behind these events remain unclear at the moment.

The interactions of heterochromatin and 3D organization in pluripotent and somatic systems

Constitutive heterochromatin and 3D organization

In many biological systems, indirect and direct links have been established between heterochromatin and 3D-genome architecture. Early Hi-C studies revealed that H3K9me3 and H3K27me3 are enriched in separate subcompartments of compartment B[42,53]. Moreover, polymer models of chromatin organization suggest that compartmentalization can be largely explained by homotypic interactions between heterochromatic regions[65,66]. These results suggest that heterochromatin may play an important role in the establishment of compartments. In the case of H3K9me3, phase separation of HP1a is a likely candidate for driving heterochromatin compartmentalization[32-34]. HP1a directly binds H3K9me3 via its chromodomains. In addition, it possesses a shadow domain enabling dimerization and binding of other heterochromatin proteins, as well as unstructured regions that facilitate phase separation[32,34,67]. Direct evidence that H3K9me3 has the potential to impact compartment

status has been provided by experiments showing that ectopic enrichment of H3K9me3 results in the switch of some compartment A regions to compartment B[68]. In addition, H3K9me3 seems to be influenced by and have an influence on the formation of loops. Studies have shown that the process of loop extrusion can disrupt H3K9me3 heterochromatin domains[69] and weaken compartment interactions[48,51]. Heterochromatin marked by H3K9me3/HP1a/HP1b in turn impacts the formation of stable loops by preventing CTCF from binding in these regions[70].

In addition to its role in 3D topology, H3K9me is important for the recruitment of chromatin to the NL via linker proteins[35,71]. In C. elegans, the protein CEC-4 has been shown to be directly responsible for tethering H3K9me-marked DNA to the nuclear periphery during embryonic development[72]. While the mechanisms of tethering chromatin to the NL are less clear in mammalian systems, the NL-associated proteins LBR, LAMIN A and LAP2b seem to play important and partially redundant roles[73-75].

Facultative heterochromatin and 3D organization

The Polycomb marks H3K27me3 and H2AK119ub1 have been associated with specific contacts between distal genomic regions in multiple systems[76-80]. In microscopy experiments, such Polycomb-associated interactions are visible as distinct foci, referred to as Polycomb bodies[81]. The Polycomb interactions may be partially mediated by phase separation of the protein CBX2, a subunit present in some forms of PRC1[82]. However, other mechanisms could also play a role, since the PRC1 protein PHC2 cannot phase separate, but its mutation does lead to the ablation of Polycomb bodies[76]. Based on these results, it seems like PRC1 is more involved in the establishment of 3D interactions than PRC2. The formation of Polycomb interactions is independent from cohesin or CTCF, as their depletion does not lead to the disappearance of such interactions and even appears to strengthen them[78,79,83]. Since chromatin marked by H3K27me3/H2AK119ub1 is enriched in a separate subcompartment[42,53], these interactions may contribute to their compartmentalization away from other chromatin types.

While H3K27me3 is enriched in a subset of LADs[42,58,62], no evidence for H3K27me3-mediated tethering has been found to date. On the contrary, recent work in a human cell line suggests that H3K27me3 could serve as a repellant for NL association[84].

A lack of 3D organization in zygote in the presence of heterochromatic marks

Given the known relationship of Polycomb and H3K9me3 with 3D organization, the observed lack of strong loops, TADs and compartments in the mouse zygote is quite striking. As mentioned earlier, the paternal genome initially has undetectable levels of all heterochromatin modifications and only accumulates low levels of H3K9me3, H3K27me3 and H2AK119ub1 by the end of the zygotic stage[26,37-40,85,86]. Nevertheless, compartments are already clearly present in the paternal pronucleus, albeit weaker than at later stages[28-30]. Meanwhile, the maternal genome is enriched for multiple heterochromatin marks, but

shows little to no compartmentalization. Based on these observations, it would seem that the presence of heterochromatic histone modifications is neither necessary nor sufficient for compartmentalization, as demonstrated by the paternal and maternal states, respectively. However, a role for the very low levels of H3K9me2/3 that accumulate in the paternal pronucleus cannot be excluded.

A plausible reason for the lack of H3K9me3-driven compartmentalization in the maternal pronucleus may be that the HP1a protein is not present in the zygote[40,87,88]. Microscopy studies present conflicting results on the timing of the first appearance of HP1a: One study observed HP1a starting in late S phase of the 2-cell embryo[88], while another study only detected the protein post-implantation[87]. If indeed HP1a is expressed starting from the late 2-cell stage, this would coincide with the increase in compartment strength post-ZGA[28,30] and the start of chromocenter formation[89]. In line with this hypothesis, a recent study in Drosophila embryos showed that HP1a (the fly homolog of HP1a) plays an important role in the establishment of strong B compartments at ZGA[90]. The lack of compartmentalization in the mouse maternal pronucleus could thus potentially be attributed to a lack of heterochromatin phase separation driven by e.g. HP1.

The lack of heterochromatin-driven interactions just after fertilization is in line with the idea that heterochromatin initially exists in an immature state[41]. After ZGA, the expression of additional heterochromatic proteins may contribute to the maturation of heterochromatin[41,91] and in turn promote the consolidation of the B compartment.

H3K27me3-enriched interaction domains at the 2-cell stage

Although heterochromatin mediated interactions seem absent in zygotes, Polycomb-driven interactions do make a unique appearance in the 2-cell embryo (Fig. 2A-D), albeit with different characteristics from Polycomb interactions as observed in mESCs. Two studies independently identified interaction domains that arise specifically on the maternal allele at the 2-cell stage and are strongly enriched for H3K27me3[79,92], which were coined Polycomb-associated domains (PADs) by Du et al. PADs display increased interactions both within and between domains, establishing compartment-like interactions at a smaller scale. The existence of these interactions is rather brief, as the domains are largely lost by the 8-cell stage and completely absent in the 64-cell stage. PADs are initially established during oocyte development between the growing oocyte II (GO II) and fully-grown oocyte (FGO) stages. In metaphase II (MII) oocytes, the interaction domains have disappeared and are only reestablished after fertilization[79].

Conditional maternal KO of the core PRC2 subunit *Eed* resulted in a substantial loss of H3K27me3, but, surprisingly, PADs were largely unaffected (Fig. 2E). However, embryos derived from *Eed* KO oocytes and WT sperm were incapable of reforming PADs at the late 2-cell stage (Fig. 2F). This implies that functional PRC2 may play a role in their reestablishment, either by PRC2-mediated interactions or by bookmarking via H3K27me3. Conditional KO of the catalytic

subunits of PRC1 (*Ring1*/*Rnf2*) resulted in a loss of H2AK119ub1 and a weakening of long-range (2-5 Mb) inter-PAD interactions (Fig. 2G), while intra-PAD and short-range inter-PAD interactions were largely unaffected[79]. The effect of this loss on PAD reestablishment in the late 2-cell embryo could not be determined, as mutant embryos arrest before this stage. Interestingly, H2AK119ub1 is largely lost at regions of maternally biased H3K27me3 by the end of the 2-cell stage[20], calling into question its role in PAD reestablishment. Together, these results suggest that PRC1 and PRC2 both play important roles in PAD formation, but exert their effect at different developmental stages.

After the 2-cell stage, PADs gradually weaken and are lost, potentially due to the loss of PAD regulators[79]. Alternatively, the presence of H3K4me3, which is transiently enriched in PADs at the 4-cell stage[92], could lead to the dissociation of PAD proteins from the chromatin. Another, non-mutually exclusive, explanation is that the regular chromatin architecture mediated by loop extrusion starts to take shape after the 2-cell stage[28,30] and may disrupt PAD interactions. In support of this, loss of loop extrusion via KO of *cohesin* resulted in an increase in PAD interaction strength in oocytes[79].

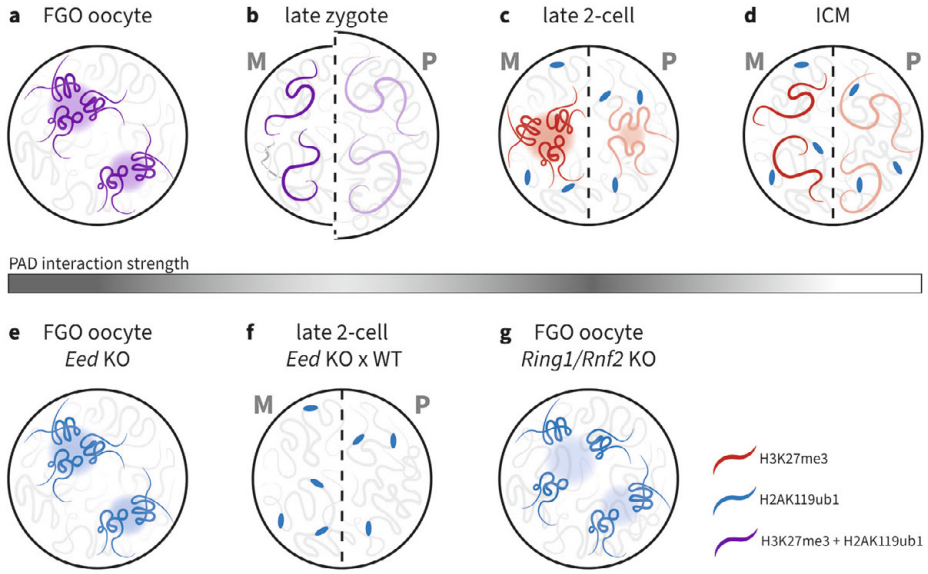


Figure 2: Putative model of changes in PADs during early mouse development
a, In FGO oocytes, broad domains of overlapping H3K27me3/H2AK119ub1 form intra- and inter-PAD interactions. **b**, In the zygote, H3K27me3/H2AK119ub1 is more strongly enriched on the maternal allele, but broad domains with moderate enrichment of both marks also exist on the paternal allele. On both alleles, inter-PAD interactions are lost and intra-PAD interactions are very weak. **c**, By the late 2-cell stage, H2AK119ub1 has been largely lost from broad H3K27me3 domains and is enriched at canonical Polycomb sites. Intra- and inter-PAD interactions are strong on the maternal allele and weak on the paternal allele. **d**, In blastocyst, all PAD interactions have been lost. Broad domains of H3K27me3 remain. **e**, In Eed KO oocytes, H3K27me3 is lost but PAD interactions are unaffected. **f**, 2-cell stage embryos derived from Eed KO oocytes cannot reform PAD interactions. **g**, In Ring1/Rnf2 KO oocytes, H2AK119ub1 is lost and long-range PAD interactions are weakened. Model based on Du et al.[79] and Collombet et al.[92].

The biological relevance of Polycomb interactions is unclear. Loss of maternal EED, and consequently H3K27me3 and PADs, results in minor changes in gene expression during preimplantation stages, problems with non-canonical imprinting by the blastocyst stage, and post-natal overgrowth[93,94]. However, it is unknown whether the loss of EED, H3K27me3 and/or PADs is responsible for the observed phenotype. Finding a way to perturb the interactions without affecting H3K27me3 may give some insight into the extent to which the Polycomb-mediated 3D architecture is instructive at this stage of development.

Heterochromatin association with the NL in early development

Constitutive heterochromatin and NL association

As discussed previously, heterochromatin marks and NL association are mechanistically linked in several systems. While both aspects of chromatin state have been individually studied in mouse preimplantation development, almost no direct comparison has been made between them[31]. Remarkably, there are clearly defined LADs in the paternal pronucleus in the zygote[31], while H3K9me2/3 has been shown to be strongly depleted from the paternal genome at this stage[27,38-40]. These paternal LADs are similar to those observed in mESC and largely overlap a set of LADs that are constitutively present across cell types[31]. Therefore, zygotic and mESC LADs have been theorized to represent the default interactions with the NL that can be further adapted by cell-type specific programs[59,95,96]. If this is the case, the interactions of the paternal genome with the NL may be driven by sequence rather than chromatin state. Indeed, there is some evidence for NL association driven by the presence of a (GA)n or GA-rich motif in other systems[73,75], although constitutive LADs are rather enriched for AT-isochores[95]. In the maternal pronucleus, on the other hand, H3K9me2/3 is present[27,38-40], along with a more unconventional and variable LAD profile[31]. While no direct comparison has been made between available H3K9me3 and NL association profiles, several clues indicate that here also NL-tethering may be independent of this histone mark. Firstly, neither H3K9me2 nor H3K9me3 appears to be localized at the nuclear periphery of the maternal pronucleus[38-41]. This is especially striking for H3K9me2, which almost exclusively localizes to the nuclear periphery in both pluripotent and somatic cells across species[71]. Moreover, overexpression of the lysine-9 specific demethylase KDM4B in the zygote results in a dramatic decrease in H3K9me3, but no subsequent effect is seen on the NL association profile in either the paternal or maternal pronucleus[31]. In addition, electron spectroscopic imaging of the zygote revealed that no condensed chromatin is present at the nuclear periphery[64], unlike at other developmental stages, indicating that zygotic LADs exist in a decondensed state and may be targeted to the NL in a chromatin independent manner. The zygote thus seems to represent a unique system in which interactions between the DNA and NL are entirely independent of H3K9me2/3. This could indicate that NL tethering would be a sequence- rather than chromatin-driven process at this stage. Alternatively, localization at the periphery of the zygotic nucleus could be a passive process rather than an active recruitment in which genomic regions would locate to other nuclear compartments and LADs would arise by exclusion.

Facultative heterochromatin and NL association

Compared to H3K9me2/3, even less research has been done on the relationship between Polycomb and NL association during preimplantation development. However, some features of the newly gained H3K27me3/H2AK119ub1 domains in oocytes and zygotes are reminiscent of regions that constitutively associate with the NL in other systems. For example, the atypical Polycomb domains on both the maternal and paternal allele are broad, located in intergenic regions, and are enriched for genes families such as olfactory receptors[18], all of which are characteristic features of constitutive LADs. The Polycomb domains show an extensive overlap with PMDs in oocytes[18], while LADs also show a strong overlap with the PMDs in somatic tissues[97]. Moreover, constitutive LADs are AT-rich[95] and PRC1 seems to be preferentially targeted to AT-rich regions in the late paternal pronucleus[85,86,98].

Interestingly, both the non-canonical maternal and paternal H3K27me3/H2AK119ub1 domains seem to appear at moments when conventional NL association is lost: For the maternal allele, the Polycomb marks are established in the oocyte[18,19], while LADs are known to be largely absent in GV oocytes[31]. Conversely, paternal Polycomb domains are laid down by the end of the zygote stage[18-20], while paternal LADs are reprogrammed between the zygote and late 2-cell stage[31]. In light of the recent work suggesting H3K27me3 may be inhibitory to NL contacts[84], it would be interesting to investigate whether Polycomb has a role in reprogramming LADs in early development.

NL association and 3D organization

NL association and 3D genome organization are both aspects of the spatial chromatin architecture. However, the exact ways in which the 3D folding and NL localization of the chromatin influence one another are not yet fully understood. In early Hi-C experiments, it was established that the B compartment shows a very strong overlap with LADs[53,63], which is in concordance with the heterochromatic nature of both. In addition, LAD boundaries frequently coincide with TAD boundaries[45], suggesting that regions within a TAD have a shared affinity for NL association and may be targeted to the NL periphery as a unit. This idea is supported by single-cell maps of NL association in a human cell line that show that larger genomic regions usually associate with the NL as a whole rather than having focal and independent points of attachment[63]. So, while the exact relationship between genome folding and NL association is still not entirely clear, these results show that the two modes of spatial organization clearly intersect and influence one another.

To investigate the connection between these two modes of chromatin organization in early development, NL-association profiles obtained from early embryos have been compared to the available Hi-C data of the same stages[31]. This comparison showed that LADs are present in zygotes, prior to the establishment of clear TADs. Moreover, TAD boundaries gain in strength at zygotic LAD boundaries during the early embryonic stages, suggesting that LADs precede TADs as a form of genome organization and may even serve as a starting point for further

maturation of the 3D structure. In line with a structuring role for the NL, a genomic tiling imaging study showed that interaction domains form at the nuclear periphery in single paternal pronuclei, despite an absence of a clear genome structure in aggregate profiles[99]. Based on these observations and the results from human single-cell LAD profiles[63], it would be interesting to determine whether LADs still associate with the NL in a coordinated manner in the absence of a strong TAD structure, or whether each locus now independently contacts and dissociates from the NL.

Another interesting observation came from the comparison of LADs with compartments in the preimplantation embryo. While LADs that are constant during early development show consistent overlap with the B compartment, a large part (39%) of LADs at the 2-cell stage belong to compartment A[31]. Interestingly, LADs that are established de novo in the 2-cell embryo and persist throughout development (11%) typically fall in the A compartment at this stage, but switch to the B compartment by the 8-cell stage. This suggests that, at least in some cases, NL association may prime regions for a switch to the B compartment. Together, these results show that, although LADs correlate highly with the B compartment in most systems, this is not necessarily the case during the first stages of embryogenesis. Moreover, in the preimplantation embryo, changes in LAD structure may direct, or at least indicate, future changes in compartmentalization. The stronger role of NL association in compartmentalization during early development could potentially be due to the absence of conventional heterochromatin. Early development can thus provide new insight into the mechanisms by which TADs get established, as well as the possible influence of NL association in shaping nuclear structure.

Conclusion

During the early stages of embryonic development, all layers of epigenetic regulation are extensively reprogrammed, including the heterochromatin and genome organization. While the intricate relationship between these two modalities is starting to be unraveled in pluripotent and somatic systems, little is known about these interactions during preimplantation development. Here, we have reviewed the current knowledge on constitutive chromatin, facultative heterochromatin and nuclear organization from the moment of fertilization until implantation in mouse embryogenesis. In addition, we have discussed the available data on their interconnectedness. From this, it seems like both heterochromatin and nuclear organization exist in immature states in the early embryo. Both start to mature by the end of the 2-cell stage, after ZGA, with conventional features such as TADs, compartments and chromatin compaction emerging. Before this moment, particularly in the zygote stage, the immature chromatin state appears to result in a weaker relationship between heterochromatin marks, 3D organization and NL association. The immature state and atypical relationship between the different modes of genome regulation could be the result of the absence of important effector proteins involved in processes such as phase separation and chromatin compaction. Further research into the epigenetics of early development will be necessary to

fully understand these processes, as well as their relevance to the establishment of totipotency and subsequent development.

Acknowledgements

We would like to thank the members of the Kind group for their feedback and support. This work was funded by an ERC Consolidator grant (ERC-CoG 1010002885-FateID) to J.K.. The Oncode Institute is partially funded by the KWF Dutch Cancer Society. IG is supported by a NWO-ENW Veni grant VI.Veni.202.073.

Author contributions

F.R., I.G. and J.K. conceived the topic. F.R. wrote the manuscript and made figures. All authors reviewed and edited the manuscript.

References

1 Xia, W. & Xie, W. Rebooting the Epigenomes during Mammalian Early Embryogenesis. *Stem Cell Reports* **15**, 1158-1175 (2020).

2 Aoki, F., Worrad, D. M. & Schultz, R. M. Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. *Dev Biol* **181**, 296-307 (1997).

3 Penagos-Puig, A. & Furlan-Magaril, M. Heterochromatin as an Important Driver of Genome Organization. *Front Cell Dev Biol* **8**, 579137 (2020).

4 Akhtar, A. & Gasser, S. M. The nuclear envelope and transcriptional control. *Nat Rev Genet* **8**, 507-517 (2007).

5 Peters, A. H. *et al.* Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107**, 323-337 (2001).

6 Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci Adv* **1**, e1500737 (2015).

7 Leeb, M. *et al.* Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes Dev* **24**, 265-276 (2010).

8 Bilodeau, S., Kagey, M. H., Frampton, G. M., Rahl, P. B. & Young, R. A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-2489 (2009).

9 Nicetto, D. *et al.* H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification. *Science* **363**, 294-297 (2019).

10 Wang, L. *et al.* Hierarchical recruitment of polycomb group silencing complexes. *Mol Cell* **14**, 637-646 (2004).

11 Blackledge, N. P. *et al.* Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell* **157**, 1445-1459 (2014).

12 Cooper, S. *et al.* Targeting polycomb to pericentric heterochromatin in embryonic stem cells reveals a role for H2AK119u1 in PRC2 recruitment. *Cell Rep* **7**, 1456-1470 (2014).

13 Kalb, R. *et al.* Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nat Struct Mol Biol* **21**, 569-571 (2014).

14 Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome research* **31**, 2170-2184 (2021).

15 Zhang, J. *et al.* Highly enriched BEND3 prevents the premature activation of bivalent genes during differentiation. *Science* **375**, 1053-1058 (2022).

16 Pauler, F. M. *et al.* H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res* **19**, 221-233 (2009).

17 Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-491 (2010).

18 Zheng, H. *et al.* Resetting Epigenetic Memory by Reprogramming of Histone Modifications in Mammals. *Mol Cell* **63**, 1066-1079 (2016).

19 Mei, H. *et al.* H2AK119ub1 guides maternal inheritance and zygotic deposition of H3K27me3 in mouse embryos. *Nat Genet* **53**, 539-550 (2021).

20 Chen, Z., Djekidel, M. N. & Zhang, Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. *Nat Genet* **53**, 551-563 (2021).

21 Zhu, Y. *et al.* Genomewide decoupling of H2AK119ub1 and H3K27me3 in early mouse development. *Science Bulletin* **66**, 2489-2497 (2021).

22 Xu, Q. *et al.* SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. *Nat Genet* **51**, 844-856 (2019).

23 Liu, X. *et al.* Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* **537**, 558-562 (2016).

24 Ward, W. S. & Coffey, D. S. DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. *Biol Reprod* **44**, 569-574 (1991).

25 Johnson, G. D. *et al.* The sperm nucleus: chromatin, RNA, and the nuclear matrix. *Reproduction* **141**, 21-36 (2011).

26 Meng, T. G. *et al.* PRC2 and EHMT1 regulate H3K27me2 and H3K27me3 establishment across the zygote genome. *Nat Commun* **11**, 6354 (2020).

27 Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620-631 (2018).

28 Ke, Y. *et al.* 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381 e320 (2017).

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

30 Du, Z. *et al.* Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235 (2017).

31 Borsos, M. *et al.* Genome-lamina interactions are established de novo in the early mouse embryo. *Nature* **569**, 729-733 (2019).

32 Larson, A. G. *et al.* Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin. *Nature* **547**, 236-240 (2017).

33 Wang, L. *et al.* Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism. *Mol Cell* **76**, 646-659 e646 (2019).

34 Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241-245 (2017).

35 See, K. *et al.* Histone methyltransferase activity programs nuclear peripheral genome positioning. *Dev Biol* **466**, 90-98 (2020).

36 Padeken, J., Methot, S. P. & Gasser, S. M. Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. *Nat Rev Mol Cell Biol* (2022).

37 Probst, A. V., Santos, F., Reik, W., Almouzni, G. & Dean, W. Structural differences in centromeric heterochromatin are spatially reconciled on fertilisation in the mouse zygote. *Chromosoma* **116**, 403-415 (2007).

38 Santos, F., Peters, A. H., Otte, A. P., Reik, W. & Dean, W. Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. *Dev Biol* **280**, 225-236 (2005).

39 Liu, H., Kim, J. M. & Aoki, F. Regulation of histone H3 lysine 9 methylation in oocytes and early pre-implantation embryos. *Development* **131**, 2269-2280 (2004).

40 van der Heijden, G. W. *et al.* Asymmetry in histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote. *Mechanisms of development* **122**, 1008-1022 (2005).

41 Burton, A. *et al.* Heterochromatin establishment during early mammalian development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. *Nat Cell Biol* **22**, 767-778 (2020).

42 Zheng, X., Kim, Y. & Zheng, Y. Identification of lamin B-regulated chromatin regions based on chromatin landscapes. *Mol Biol Cell* **26**, 2685-2697 (2015).

43 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301 (2001).

44 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

45 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

46 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385 (2012).

47 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

48 Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56 (2017).

49 Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).

50 Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049 (2016).

51 Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707 e614 (2017).

52 Cavalheiro, G. R., Pollex, T. & Furlong, E. E. To loop or not to loop: what is the role of TADs in enhancer function and gene regulation? *Curr Opin Genet Dev* **67**, 119-129 (2021).

53 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

54 Ooi, S. L. & Henikoff, S. Germline histone dynamics and epigenetics. *Curr Opin Cell Biol* **19**, 257-265 (2007).

55 Jung, Y. H. *et al.* Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. *Cell Rep* **18**, 1366-1382 (2017).

56 Battulin, N. *et al.* Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol* **16**, 77 (2015).

57 Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J* **36**, 3600-3618 (2017).

58 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).

59 Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-613 (2010).

60 Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* **41**, 246-250 (2009).

61 Kind, J. *et al.* Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178-192 (2013).

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

63 Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134-147 (2015).

64 Ahmed, K. *et al.* Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS One* **5**, e10531 (2010).

65 Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* **570**, 395-399 (2019).

66 Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences* **115**, E6697-E6706 (2018).

67 Jacobs, S. A. *et al.* Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. *EMBO J* **20**, 5232-5241 (2001).

68 Feng, Y. *et al.* Simultaneous epigenetic perturbation and genome imaging reveal distinct roles of H3K9me3 in chromatin architecture and transcription. *Genome Biol* **21**, 296 (2020).

69 Haarhuis, J. H. I. *et al.* A Mediator-cohesin axis controls heterochromatin domain formation. *Nat Commun* **13**, 754 (2022).

70 Spracklin, G. *et al.* Heterochromatin diversity modulates genome compartmentalization and loop extrusion barriers. *bioRxiv* (2021).

71 Poleshko, A. *et al.* H3K9me2 orchestrates inheritance of spatial positioning of peripheral heterochromatin through mitosis. *Elife* **8** (2019).

72 Gonzalez-Sandoval, A. *et al.* Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. *Cell* **163**, 1333-1347 (2015).

73 Zullo, J. M. *et al.* DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell* **149**, 1474-1487 (2012).

74 Solovei, I. *et al.* LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584-598 (2013).

75 Ottaviani, A. *et al.* Identification of a perinuclear positioning element in human subtelomeres that requires A-type lamins and CTCF. *EMBO J* **28**, 2428-2436 (2009).

76 Wani, A. H. *et al.* Chromatin topology is coupled to Polycomb group protein subnuclear organization. *Nat Commun* **7**, 10291 (2016).

77 Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet* **47**, 1179-1186 (2015).

78 Kundu, S. *et al.* Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Mol Cell* **71**, 191 (2018).

79 Du, Z. *et al.* Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Mol Cell* **77**, 825-839 e827 (2020).

80 Tolhuis, B. *et al.* Interactions among Polycomb domains are guided by chromosome architecture. *PLoS Genet* **7**, e1001343 (2011).

81 Buchenau, P., Hodgson, J., Strutt, H. & Arndt-Jovin, D. J. The distribution of polycomb-group proteins during cell division and development in Drosophila embryos: impact on models for silencing. *J Cell Biol* **141**, 469-481 (1998).

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

83 Rhodes, J. D. P. *et al.* Cohesin Disrupts Polycomb-Dependent Chromosome Interactions in Embryonic Stem Cells. *Cell Rep* **30**, 820-835 e810 (2020).

84 Siegenfeld, A. P. *et al.* Polycomb-lamina antagonism partitions heterochromatin at the nuclear periphery. *Nature Communications* **13**, 4199 (2022).

85 Eid, A. & Torres-Padilla, M. E. Characterization of non-canonical Polycomb Repressive Complex 1 subunits during early mouse embryogenesis. *Epigenetics* **11**, 389-397 (2016).

86 Puschendorf, M. *et al.* PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. *Nat Genet* **40**, 411-420 (2008).

87 Wongtawan, T., Taylor, J. E., Lawson, K. A., Wilmut, I. & Pennings, S. Histone H4K20me3 and HP1alpha are late heterochromatin markers in development, but present in undifferentiated embryonic stem cells. *J Cell Sci* **124**, 1878-1890 (2011).

88 Meglicki, M., Teperek-Tkacz, M. & Borsuk, E. Appearance and heterochromatin localization of HP1alpha in early mouse embryos depends on cytoplasmic clock and H3S10 phosphorylation. *Cell Cycle* **11**, 2189-2205 (2012).

89 Martin, C. *et al.* Genome restructuring in mouse embryos during reprogramming and early development. *Dev Biol* **292**, 317-332 (2006).

90 Zenk, F. *et al.* HP1 drives de novo 3D genome reorganization in early Drosophila embryos. *Nature* **593**, 289-293 (2021).

91 Guthmann, M., Burton, A. & Torres-Padilla, M. E. Expression and phase separation potential of heterochromatin proteins during early mouse development. *EMBO Rep* **20**, e47952 (2019).

92 Collombet, S. *et al.* Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature* **580**, 142-146 (2020).

93 Prokopuk, L. *et al.* Loss of maternal EED results in postnatal overgrowth. *Clin Epigenetics* **10**, 95 (2018).

94 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).

95 Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* **23**, 270-280 (2013).

96 Rullens, P. M. J. & Kind, J. Attach and stretch: Emerging roles for genome-lamina contacts in shaping the 3D genome. *Curr Opin Cell Biol* **70**, 51-57 (2021).

97 Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet* **50**, 591-602 (2018).

98 Tardat, M. *et al.* Cbx2 targets PRC1 to constitutive heterochromatin in mouse zygotes in a parent-of-origin-dependent manner. *Mol Cell* **58**, 157-171 (2015).

99 Payne, A. C. *et al.* In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371** (2021).

CHAPTER 6

Antagonism between H3K27me3 and genome lamina-association drives atypical spatial genome organization in the totipotent embryo

Isabel Guerreiro[1,2,6*], Franka J. Rang[1,2,6], Yumiko K. Kawamura[3], Carla Kroon-Veenboer[1,2], Jeroen Korving[1,2], Femke C. Groenveld[1,2,4], Ramada E. van Beek[1,2], Silke J. A. Lochs[1,2], Ellen Boele[1,2], Antoine H. M. F. Peters[3,5], Jop Kind[1,2,4*]

1: Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and University Medical Center Utrecht, Utrecht, the Netherlands
2: Oncode Institute, the Netherlands
3: Friedrich Miescher Institute for Biomedical Research (FMI), Basel, Switzerland
4: Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, the Netherlands
5: Faculty of Sciences, University of Basel, Basel, Switzerland
6: These authors contributed equally
*Correspondence: J.K. (j.kind@hubrecht.eu) and I.G. (i.guerreiro@hubrecht.eu)

Nature Genetics, in press

Abstract

The first days of mammalian embryonic development are accompanied by major changes in the chromatin landscape and nuclear organization. In particular, the positioning of genomic regions with respect to the nuclear lamina is highly unusual during this developmental time and shows major differences between parental alleles. The mechanisms and implications of this atypical genome organization remain, however, elusive. Here, we generated single-cell profiles of lamina-associated domains (LADs) coupled with transcriptomics throughout the first stages of mouse development. We find that regions that are uniquely dissociated from the lamina in the 2-cell embryo strongly overlap broad domains of non-canonical H3K27me3. Loss of H3K27me3 through a maternal knock-out of *Eed* results in a restoration of canonical LAD profiles and resolves the allelic asymmetry, suggesting an antagonistic relationship between lamina association and H3K27me3. Furthermore, through expression of a specialized tether, we successfully recruit H3K27me3 domains to the nuclear lamina, which is especially effective when the mark coincides with canonical LADs. Based on these results, we propose a model in which the atypical organization of LADs at the 2-cell stage is the result of a tug-of-war between the intrinsic affinity of genomic regions for the nuclear lamina and H3K27me3, constrained by the available space at the nuclear periphery. This study provides detailed insight into the molecular mechanisms regulating nuclear organization during early mammalian development.

Introduction

Mammalian development begins with the fusion of two differentiated cells, the gametes, that give rise to a totipotent zygote. The embryo subsequently undergoes multiple cycles of cell division, with inner cells progressively transitioning to a pluripotent state, while outer cells commit to the extra-embryonic lineage by the time of implantation in the uterus. After fertilization, maternal transcripts are progressively degraded as embryonic genes become active. In mouse, all these events occur within the first three days of development and are accompanied by extensive epigenetic reprogramming, as well as major changes in spatial genome organization (reviewed in refs[1,2]).

One important feature of nuclear organization is the localization of genomic regions at the nuclear lamina (NL). These genomic regions, termed lamina-associated domains (LADs), have been extensively studied and are characterized by a low gene density, low gene expression and high repeat content, as well as other features of constitutive heterochromatin. Moreover, LADs have been shown to play an important role in genome architecture and gene expression in various systems (reviewed in refs[3,4,5]). LADs are typically detected using the DamID technique[6], which is based on the expression of the *E. coli* DNA adenine-methyltransferase (Dam) to a protein-of-interest and subsequent methylation *in vivo*. Fusing Dam to a component of the NL, typically Lamin B1, thus results in the specific methylation of LADs that can subsequently be sequenced and mapped to the genome.

Previous work studying LADs in the context of preimplantation development has shown atypical patterns of lamina association at these stages. Maternal LADs have been found to be established *de novo* following fertilization, while paternal LADs undergo massive rearrangements between the zygote and 2-cell stages. Consequently, maternal and paternal genomes show differences in lamina association up until the 8-cell stage[1].

In addition to LADs, other chromatin features have been demonstrated to undergo extensive rewiring during early stages of development. Recently, several studies profiling histone post-translational modifications (PTMs) in early embryogenesis have started to shed light on the epigenetic features and dynamics of preimplantation development[7-13]. Trimethylation at histone 3 lysine 27 (H3K27me3), a histone PTM associated with the repression of developmental genes, is deposited by the Polycomb repressive complex (PRC) 2. After fertilization, H3K27me3 has been shown to lose its typical distribution at promoters of developmental genes at maternal and paternal genomes while retaining non-canonical broad distal domains in regions devoid of developmental genes in the maternal allele[13]. Trimethylation at histone 3 lysine 9 (H3K9me3), a mark found in LADs in several cell types[14], also shows unusual enrichment and allelic asymmetry in early mouse development. ChIP-seq data has shown that H3K9me3 and H3K27me3 extensively overlap during early developmental stages across the genome, in contrast to what has been reported in other biological systems[11].

Although low-input technologies have recently shed light on the chromatin state and nuclear architecture of the early mouse embryo, the underlying mechanisms and the relationship between different epigenetic layers remain largely unexplored. Additionally, while cell-to-cell variability in gene expression and chromatin modifiers is proposed to contribute to early cell fate choices[15-17], the extent and role of variability in genome-lamina association during preimplantation development remains largely unexplored. Here, we profile single-cell allele-specific LADs throughout a range of early developmental stages to provide unprecedented mechanistic insight into the different aspects of atypical genome-nuclear lamina associations that characterize the nucleus of the totipotent embryo.

Genome–lamina associations are highly variable between single blastomeres at the 2-cell stage

Our previous work on LADs in preimplantation development suggested that cell-to-cell variability in LADs may be particularly high between single cells in early developmental stages[1]. To address this finding in greater detail we have made use of scDam&T-seq, a newly developed single-cell DamID technique[18] that: 1) provides superior signal-to-noise ratio, 2) increases throughput, 3) allows to register the embryo-of-origin for each cell, 4) is coupled to transcriptomics in the same cell (Fig. S1a-b).

Using this technique with Dam fused to Lamin B1 (Dam-LMNB1), we obtained a total of 755 single-cell LAD profiles that passed quality control thresholds (see Methods): 107 zygote cells, 197 2-cell stage cells, 183 8-cell stage cells and 268 mES cells (Fig. S1b, Supplementary Table 1-2). Genomic and transcriptional outputs showed a median of ~10’000 unique DamID fragments and ~10’000 unique transcripts per cell across all stages (Fig. S1c-d). Average LAD profiles per stage showed high concordance with previously published genomic and imaging-based data (Fig. S1e-g), and transcriptional profiles had the expected patterns of gene expression (Fig. S1h).

For each collected cell, we obtained the LAD profile and the corresponding gene expression read-out (Fig. 1a, Fig. S2a). Representation of the DamID data by uniform manifold approximation and projection (UMAP) shows clear clustering according to stage based on lamina association (Fig. 1b), demonstrating that single-cell LAD profiles are stage specific. Similarly, UMAP representation of the transcription data displayed clear separation on developmental stage (Fig. 1c).

To understand whether there were differences in cell-to-cell LAD variability across stages, we converted single-cell LMNB1 values to an aggregate contact frequency (CF), which represents the fraction of cells in which a genomic bin is associating with the lamina[19]. CF profiles suggested substantial differences in genome-lamina association across stages consistent with previous work[1] (Fig. S2a). CF distributions per stage further revealed more intermediate CF values for 2-cell and 8-cell stages, indicating a bigger proportion of genomic loci with

variable lamina association (Fig. S2b). To quantify genome-wide LAD variability between cells, we calculated the overlap in lamina association (Yule's Q coefficient) between all pairs of cells, which provides a measure of similarity between single-cell LAD profiles. This showed that 2-cell genome-lamina associations are particularly heterogeneous between single cells compared to other stages (Fig. 1d). Moreover, we found that cells from the same embryo tend to have more similarity in genome-lamina association, especially at the 2-cell stage (Fig. S2c). This result suggests that LADs are partially inherited from zygote, potentially due to restraints dictated by nuclear organization and chromatin states. However, we cannot completely exclude a technical component since the volume of injected Dam construct per embryo may slightly vary.

Lastly, to determine whether the level of LAD variability was constant along the linear chromosome, we examined the distribution of CFs across all autosomal chromosomes at the 2-cell stage and found that genomic regions proximal to the centromere showed unusually high CF values (Fig. S2d). The centromeric enrichment was less pronounced in other developmental stages and absent in mESCs (Fig. S2d-e). These results suggest that despite high levels of LAD variability, centromeric regions tend to associate with the lamina in a more uniform manner across cells. This is a feature uniquely observed at the totipotent 2-cell stage, which coincides with the dramatic change in centromere organization in the nucleus, with relocation from the borders of nucleolar precursor bodies and clustering into chromocenters[20,21].

Cell-to-cell LAD variability at the 2-cell stage is higher for the paternal allele

Previous work has reported allelic differences in genome-lamina association up to the 8-cell stage[1]. This prompted us to investigate single-cell LAD profiles on the maternal and paternal alleles by using a hybrid cross between mice of two different strains (B6CBAF1/J females and CAST/EiJ males). For mESCs, a hybrid strain of CAST/EiJx129Sv was used. These crosses yielded high-quality single-cell profiles of allele-specific lamina association, which revealed that the allelic asymmetry is present across single cells (Fig. 1e, Fig. S3a-c). Moreover, in zygote, 2-cell and 8-cell embryos, the paternal genome occupies a bigger portion of the nuclear periphery compared to the maternal genome, while in mESCs LAD coverage was comparable between the two alleles (Fig. S3d).

We next examined whether the high level of LAD variability was observed on both alleles, again using the metric of global cell-to-cell similarity (Fig. 1f). Strikingly, this revealed a profound discrepancy in the level of variability on the two alleles: While at the zygote stage maternal LADs are more variable than paternal LADs, this trend is clearly inverted at the 2-cell and 8-cell stages (Fig. 1f). This is in contrast with canonical mESC LADs that show similar levels of cell-to-cell variability between both alleles.

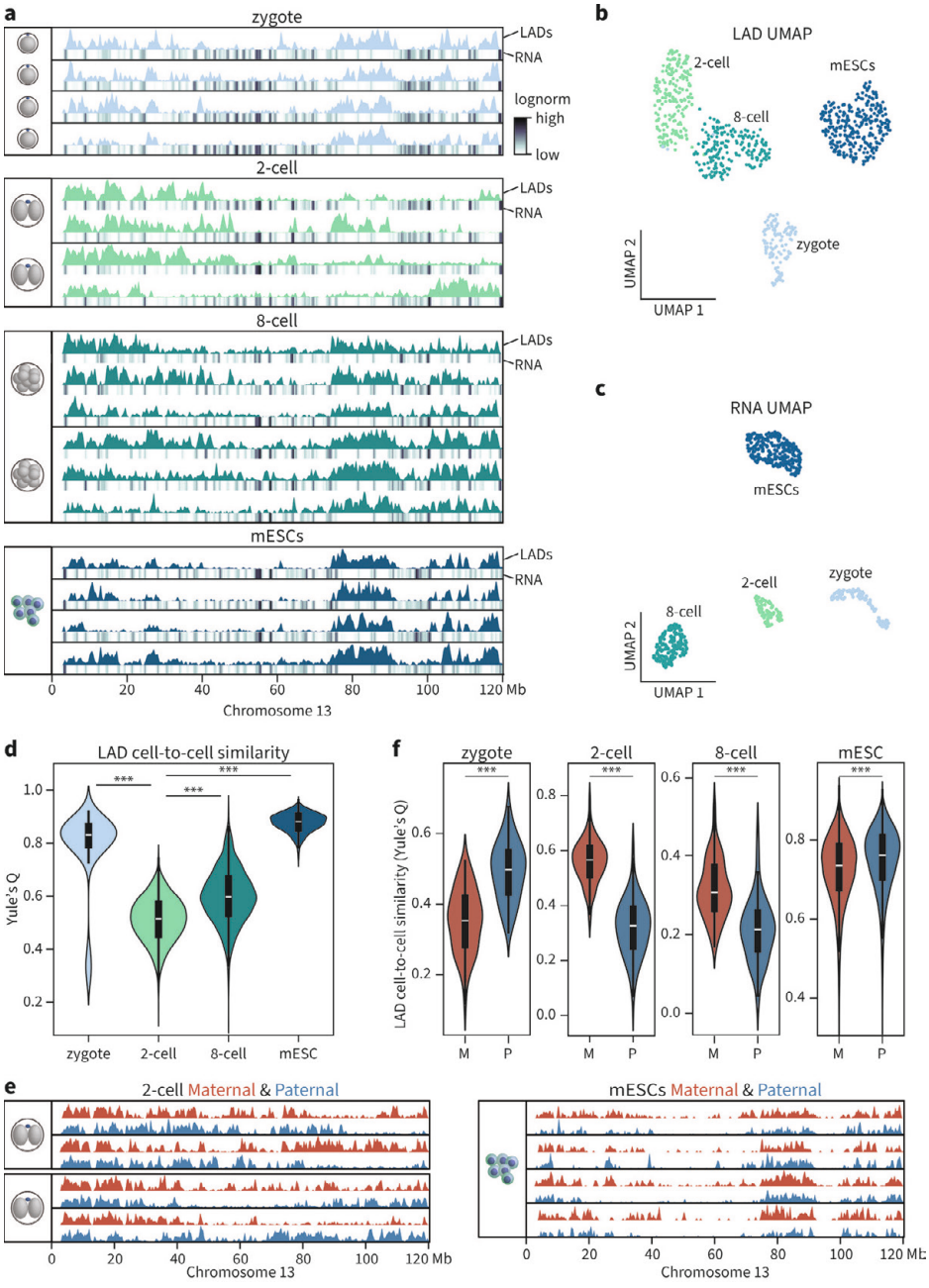


Figure 1: Genome-lamina contacts at the 2-cell stage are highly variable between single cells
a, Examples of LAD single-cell profiles (RPKM) and corresponding gene expression track (log-transformed depth-normalized values scaled to maximum value per sample) across the entire chromosome 13 at different developmental stages and in mESCs. Single-cell profiles derived from the same embryo are grouped. **b**, UMAP based on Dam-LMNB1 single-cell readout (n = 755). **c**, Single-cell UMAP based on transcriptional readout of the cells in (b) passing transcriptional thresholds (n = 482). **d**, Distributions of cell-to-cell similarity (Yule's Q) of binarized single-cell Dam-LMNB1 data for zygote (n = 253 cell pairs, p <

1e-100), 2-cell (n = 17,020 cell pairs, p = NA), 8-cell (n = 4,950 cell pairs, p < 1e-100), and mESC (n = 15,356 cell pairs, p < 1-100). **e**, 2-cell and mESC example single-cell LAD profiles split in maternal (red) and paternal (blue) alleles. For the 2-cell stage, allele-specific profiles from the same embryo are grouped. **f**, Distributions of cell-to-cell similarity (Yule's Q) of binarized allelic single-cell Dam-LMNB1 data for zygote (n = 78 cell pairs, p = 8.4e-16); 2-cell stage, (n = 210 cell pairs, p = 4.8e-62); 8-cell stage (n = 136 cell pairs, p = 2.2e-14) and mESC (n = 29,161 cell pairs, p < 1e-100). Statistical testing was performed with a two-sided Mann-Whithney U test. Boxplots included in (d) and (f) indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).

Despite the differences in lamina association patterns and variability at the 2-cell stage, both alleles showed higher CF values at centromeric regions of chromosomes, indicating that these regions are more consistently associated with the NL (Fig. S3e). The difference in CF between centromeric and non-centromeric regions was, however, less noticeable in the paternal genome (Fig. S3f), presumably reflecting generally higher genome-lamina associations along the paternal chromosomes.

Overall, these results indicate that the parental genomes do not evenly associate with the NL and that paternal LADs contribute the most to the unusually high cell-to-cell variability in the 2-cell and 8-cell embryo.

LAD variability at the 2-cell stage is not accompanied by major changes in chromatin state and transcription

The localization of genomic regions at the NL is typically associated with heterochromatic features and low gene expression. Since LAD cell-to-cell variability is unusually high at the 2-cell stage, we hypothesized that heterogeneous lamina association may be related to changes in chromatin state or gene expression. To test whether differential lamina association resulted in effects on gene expression, we made use of the combined genomic and transcriptomic read-out of our single-cell data (Fig. 1a). When comparing the expression of genes in cells where they associate with the lamina to the expression in cells where the same genes do not associate with the NL, we observed no differences in transcript counts (Fig. S4a), arguing that the variable lamina association has no detectable consequences for the underlying gene expression.

To determine whether, similar to LADs, other chromatin features display variability across cells, we employed EpiDamID[22], a single-cell DamID-based technique that has been adapted to detect histone marks through the fusion of either single-chain variable fragments (scFv) or chromatin reader domains to Dam. We chose to profile 1) H3K9me3, which is often enriched in LADs, 2) H3K27me3, which plays an essential role in repressing genes during embryonic development, and 3) open chromatin, which demarcates euchromatin and tends to anti-correlate with LADs. To generate single-cell profiles of all three genomic features, the fusion constructs Dam-$Cbx1_{CD}$ (a tuple of the Cbx1 chromodomain), Dam-αH3K27me3 (scFv)[22], and the untethered Dam were used, respectively (Fig. 2a, Fig. S4b). The aggregate single-cell profiles displayed high genome-wide correlations with the corresponding publicly available datasets (Fig. S4c)[22].

In contrast to the high variability of genome-lamina association, H3K27me3, H3K9me3 and open chromatin appeared to have a more uniform distribution across single cells (Fig. 2a highlighted regions, Fig. S4d). In order to quantify and compare the levels of cell-to-cell variability of the different chromatin features, we controlled for construct-specific sparsity and noise by normalizing the similarity scores of our data to the same metric from a simulated dataset, which mimics the influence of technical artefacts on variability (Fig. S4b and Methods). The normalized cell-to-cell similarity measurement confirmed that LADs are more variable compared to H3K9me3, H3K27me3 and accessible chromatin (Fig. 2b). These results indicate that unlike genome-lamina associations, these chromatin features are rather constant between individual cells, although some variability may still exist at a finer resolution. It is therefore unlikely for heterogeneity in chromatin state to be the cause or the consequence of LAD cell-to-cell variability at the 2-cell stage.

Regions that dissociate from the NL at the 2-cell stage are high in H3K27me3

In addition to high levels of cell-to-cell LAD variability, overall patterns of genome-lamina associations are highly atypical during the first cleavage stages: a large proportion of genomic regions associate or dissociate with the lamina uniquely in early embryos[1].

Interestingly, by visual inspection, we observed a striking relationship between regions of unusual lamina dissociation and non-canonical broad H3K27me3 domains (Fig. 2a). This prompted us to further investigate the relationship between LADs and H3K27me3 at the 2-cell stage.

To more clearly visualise atypical genome-lamina association patterns, we compared the 2-cell and mESC LADs in the context of 2-cell non-canonical H3K27me3 (ncH3K27me3) domains obtained from publicly available ChIP-seq data[13]. Interestingly, we found that regions with high H3K27me3 levels are depleted in genome-lamina associations at the 2-cell stage, whereas these same regions strongly associate with the NL in mESCs (Fig. 2c-d).

Having identified a possible relationship between genome-lamina association and H3K27me3 enrichment, we clustered all genomic regions based on allele-resolved genome-lamina association and H3K27me3 values across all stages, which resulted in the identification of 6 genomic clusters (Fig. 2e, Fig. S4e, Supplementary Table 3). To gain insight into how these clusters relate to different chromatin states, we projected data of multiple ChIP-seq published datasets onto the clustering[10-13,23] (Fig. S4e, Supplementary Table 4).

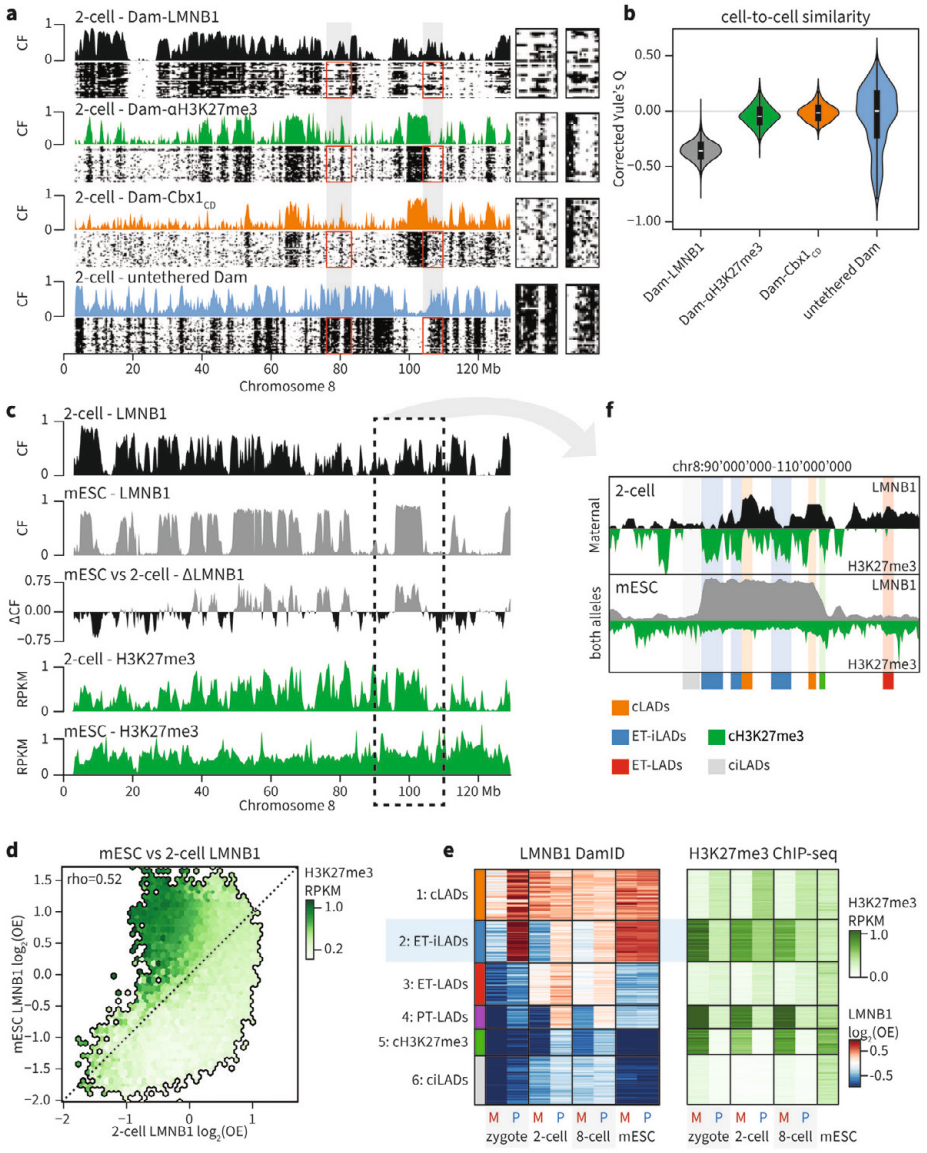


Figure 2: ET-iLADs that specifically detach from the NL in early embryos are enriched in H3K27me3
a, Binarized single-cell profiles for Dam-LMNB1, Dam-aH3K27me3, Dam-Cbx1$_{CD}$ in 2-cell embryos across the entire chromosome 8, ordered by decreasing unique number of GATCs. The 30 richest cells are shown per condition. Above each heatmap, corresponding CF tracks are shown. Grey boxes highlight example regions with variable genome-lamina association, but uniform enrichment of the other chromatin features. Cut-outs on the right show the same regions at greater magnification. **b**, Distributions of cell-to-cell similarity scores (corrected Yule's Q) per construct. A value of zero indicates a level of similarity that is expected based on technical noise, greater and smaller values indicate higher and lower similarity, respectively. Boxplots indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers). **c**, LMNB1 CF profiles and differential CF profile between mESC and 2-cell stage along chromosome 8. Publicly available H3K27me3 ChIP-seq data profiles of the 2-cell stage

and mESCs are also shown[13]. The dashed box highlights an example with low CF and high H3K27me3 at the 2-cell stage. **d**, Genome-wide comparison between mESC and 2-cell LMNB1 DamID CF values in 100-kb bins. Color intensity refers to average H3K27me3 RPKM values obtained from publicly available ChIP-seq data[13]. **e**, Clustering of 100-kb genomic bins based on their allelic LMNB1 and H3K27me3 ChIP-seq values[13] at the different embryonic stages and in mESCs. Left: LMNB1 CF values from both the maternal (M) and paternal (P) alleles across the different stages. Right: H3K27me3 ChIP-seq values are plotted. **f**, Example genomic region with examples of five of the genomic clusters identified in (e). Mirror profiles show LMNB1 values (DamID) on top and H3K27me3 (ChIP-seq[13]) on the bottom for the 2-cell stage (maternal allele) and mESCs (both alleles).

Consistent with our previous observations, we confirmed the presence of genomic regions that are strongly associated to the lamina in mESCs, but lack lamina association in the early embryo, particularly on the maternal allele (cluster 2, Fig. 2e-f, Fig. S4e). Lack of genome-lamina associations specific to early developmental stages was strongly mirrored by consistently high levels of non-canonical H3K27me3 and H3K9me3. These chromatin states are unique to early development, as these regions are not enriched for H3K27me3 and H3K9me3 in mESCs when genome-lamina associations are restored (Fig. 2e, Fig. S4e). Moreover, these regions feature characteristics of typical LADs, such as low gene density and high density of LINE L1 repeats and A/T content (Fig. S4f-h). We therefore termed these regions Embryonic Transient inter-LADs (ET-iLADs).

In addition to ET-iLADs, we identified five additional clusters with distinct lamina association and H3K27me3 enrichment. Constitutive LADs (cLADs, cluster 1) are characterised by strong lamina association and low H3K27me3 across all stages. Embryonic Transient LADs (ET-LADs, cluster 3) contain genomic regions that display most genome-lamina associations unique to cleavage-stage embryos (Fig. 2e-f, Fig. S4e). Similarly, paternal-specific LADs (pET-LADs, cluster 4) are exclusively enriched on the paternal allele at 2-cell and 8-cell stages, while having high levels of H3K27me3 on the maternal allele, further demonstrating the antagonistic relationship between these chromatin types in the early embryo. Both pET-LADs and ET-LADs represent genomic regions that associate with the lamina only in the context of preimplantation development and score lower on classical LAD features, such as low LINE L1 density, low A/T content, and high gene density compared to more canonical LADs (Fig. S4f-h). Finally constitutive inter-LADs (ciLADs, cluster 6) and canonical H3K27me3 (cH3K27me3, cluster 5) have little to no lamina association throughout development. Whereas cH3K27me3 regions are enriched for known Polycomb-regulated genes and show high levels of H2AK119ub1 (Fig. S4f-e), a histone PTM laid down by Polycomb Repressive Complex 1 (PRC1), ciLADs are enriched instead for features of euchromatin, such as H3K4me3, H3K27ac, SINEs and high gene density (Fig. S4e-g).

Altogether these results indicate that genome-lamina association is extensively rewired during early development and show unique relationships to other chromatin features. Most notably, ET-iLADs specifically detach from the NL during the first days of embryonic development, coinciding with the presence of transiently high occupancy of non-canonical H3K27me3 to these same genomic regions. These observations point towards an antagonistic relationship between genome-lamina association and Polycomb regulation during preimplantation development.

H3K27me3 sequesters genomic regions away from the NL at the 2-cell stage

Having observed an apparent inverse relationship between genome-lamina association and ncH3K27me3 domains, we aimed to test this relationship directly. To this end, we proceeded to deplete H3K27me3 in early development using a maternal knockout of *Eed* (*Eed* mKO), an essential component of Polycomb repressive complex 2 (PRC2) that deposits H3K27me3. This mutation is acquired in growing oocytes of $Eed^{fl/fl}$;$Gdf9^{Cre}$ female mice and results in H3K27me3 loss from the oocyte up to the 8-cell stage[24,25], when the mark is restored by expression of paternal *Eed*. Embryos obtained from crosses with $Eed^{fl/fl}$ mothers are used as a control. As previously reported[25,26], H3K27me3 is vastly reduced in *Eed* mKO embryos compared to control (Fig. S5a).

We performed scDam&T-seq with the Dam-LMNB1 construct on both *Eed* mKO and control 2-cell embryos to uncover the effect of H3K27me3 loss on LADs. To obtain allelic-resolved data, we used hybrid crosses between C57BL/6J females and JF1/MsJ males. Comparison of LAD profiles between the two conditions showed extensive differences in genome-lamina association upon H3K27me3 depletion (Fig. 3a, Fig. S5b). Notably, gains in genome-lamina associations upon *Eed* mKO correspond to patterns of H3K27me3 enrichment at the 2-cell stage (Fig. 3a, Fig. S5c). To further confirm this observation, we computed allele-specific LMNB1 enrichment of both *Eed* mKO and control embryos over H3K27me3 domains. While control embryos displayed clear depletion of genome-lamina association at H3K27me3 domains, *Eed* mKO embryos showed LMNB1 enrichment similar to neighboring regions, indicating that H3K27me3 regions are relocated to the NL in the *Eed* mKO condition (Fig. 3b).

We then asked how the different genomic clusters identified above were affected by Eed depletion. Indeed, the clusters that gained genome-lamina associations corresponded to regions rich in H3K27me3 in the WT 2-cell embryo (Fig. S5d). Particularly, the ET-iLADs, which are characterised by canonical LAD features, underwent the largest increase in genome-lamina associations, while cH3K27me3 regions and pET-LADs showed more modest increases in LMNB1 values. Strikingly, the resulting 2-cell stage LAD patterns in the *Eed* mKO condition strongly resembled genome-lamina associations as they are found in mESCs (Fig. S5d), suggesting that the loss of H3K27me3 reverts LADs back to its canonical state. Therefore, our findings show that H3K27me3 or PRC2 play a key role in determining the atypical LAD organizations during early developmental stages.

Allelic LAD differences are mostly resolved in the absence of H3K27me3

Given the antagonistic effect of H3K27me3 on genome-lamina association and the strong allelic asymmetry of this mark, we hypothesized that H3K27me3 could be related to the pronounced allelic differences in LADs typical of early developmental stages[1]. Comparison of maternal and paternal Dam-LMNB1 values in control embryos showed that regions with a paternal bias in lamina association were enriched for H3K27me3 on the maternal allele (Fig. 3c, left panel). A similar trend was seen in the converse situation although to a lesser extent, likely due to overall low levels of H3K27me3 on the paternal allele (Fig. 3c, right panel).

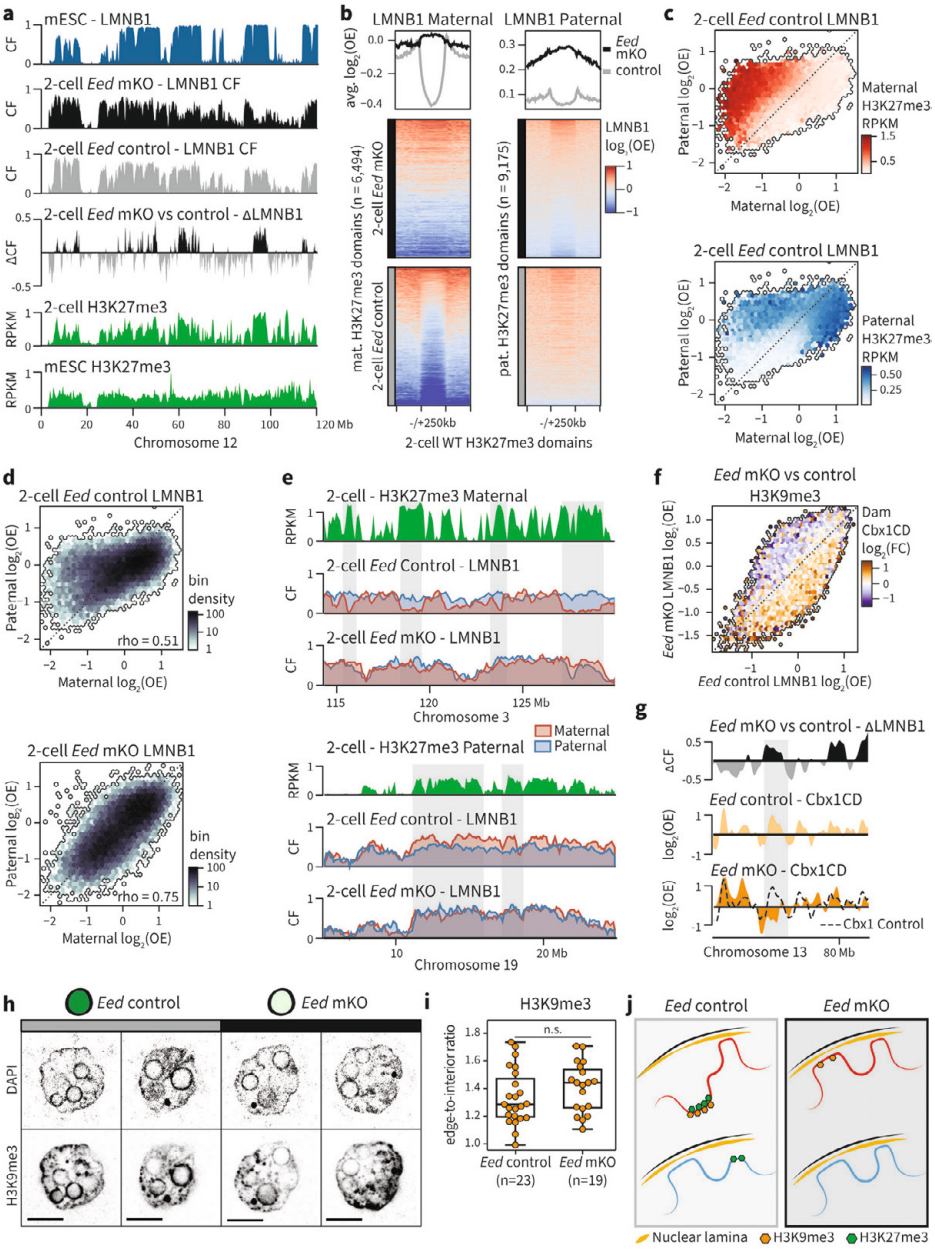


Figure 3: H3K27me3 antagonizes genome-lamina association during early development
a, LMNB1 profiles of control and *Eed* mKO 2-cell embryos and mESC, differential LMNB1 enrichment of *Eed* mKO vs control, and H3K27me3 profiles (ChIP-seq[13]) over chromosome 12. **b**, Enrichment plot showing maternal (left) and paternal (right) LMNB1 enrichment of *Eed* mKO and control 2-cell embryos over H3K27me3 domains from the same allele and surrounding 250 kb. Heatmaps show LMNB1 signal per domain, while line plots show average enrichment over all domains. **c**, Comparison of maternal and paternal LMNB1 $\log_2$(OE) in 100-kb bins . The color indicates the average maternal (left, red) and

paternal (right, blue) H3K27me3 RPKM values[13]. **d**, Comparison of paternal and maternal LMNB1 $\log_2$(OE) in control (left, rho = 0.51, p < 1e-100) or *Eed* mKO (right, rho = 0.75, p < 1e-100) embryos. Correlations were computed using Spearman's rank-order correlation. **e**, Example genomic regions where a reduction of LMNB1 allelic asymmetry is observed in regions enriched for maternal-specific H3K27me3 (top) or paternal-specific H3K27me3 (bottom). **f**, Comparison of LMNB1 $\log_2$(OE) in the *Eed* mKO and control embryos. Color intensity refers to the $\log_2$(fold change) in Dam-Cbx1$_{CD}$ (H3K9me3) enrichment between the two conditions. **g**, Example genomic region where the differential LMNB1 enrichment between the *Eed* mKO and control is plotted, as well as the Dam-Cbx1$_{CD}$ (H3K9me3) $\log_2$(OE) of both conditions separately. The Dam-Cbx1$_{CD}$ of the control is also plotted as a dashed line on the *Eed* mKO profile for reference. A shaded box highlights a region that gains genome-lamina associating in the *Eed* mKO, while showing a loss in Dam-Cbx1$_{CD}$ enrichment. **h**, DAPI and immunostaining of H3K9me3 in 2-cell *Eed* mKO or *Eed* control embryos (scale bar = 10 mm). **i**, Quantification of the radial (1 mm) enrichment of H3K9me3 relative to the rest of the nucleus for the immunostaining in (**h**). Significance was computed using Welch's two-sided t-test (p = 0.23). Boxplots indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers) **j**, Model that illustrates the effect of H3K27me3 depletion in *Eed* mKO embryos.

We next examined the effect of removing H3K27me3 on allelic differences in LADs. There was a clear increase in correlation between maternal and paternal Dam-LMNB1 values in *Eed* mKO embryos (Spearman's rho = 0.75) compared to control conditions (Spearman's rho = 0.51) (Fig. 3d). Visual inspection of Dam-LMNB1 profiles also showed clear reduction of allelic LAD asymmetry upon H3K27me3 loss (Fig. 3e). This increase in allelic concordance was only apparent in regions enriched in H3K27me3, while regions with low H3K27me3 levels displayed high allelic concordance in *Eed* mKO and control embryos (Fig. S5e).

These results show that non-canonical distributions of H3K27me3 prevent conventional lamina association from forming and dictate allelic LAD asymmetry during early mouse development.

H3K9me3 is unlikely to play a role in relocating of genomic regions to the NL upon loss of H3K27me3

Given the extensive overlap between H3K27me3 and H3K9me3 in preimplantation embryos (Fig. 2a, Fig. S4e) we wondered whether in the absence of H3K27me3, H3K9me3 could be involved in the mechanism that repositions genomic regions to the NL. To this end, we performed scDamID&T using the Dam-Cbx1$_{CD}$ construct, which labels H3K9me3 enriched regions, in the *Eed* mKO condition. Surprisingly we found that the genomic regions that gain lamina association in the absence of H3K27me3 tend to have reduced H3K9me3 levels compared to controls (Fig. 3f-g, Fig. S5f). To confirm this observation, we stained 2-cell *Eed* mKO and control embryos for H3K9me3 and found no increase in signal at the nuclear periphery, consistent with decreased H3K9me3 levels at genomic regions that repositioned to the NL (Fig. 3h-i). The observed reduction in H3K9me3 at regions that lost ncH3K27me3 precludes this mark from being involved in the reestablishment of a canonical LAD configuration in the absence of H3K27me3. This notion is further supported by the fact that the paternal genome also gains genome-lamina association in *Eed* mKO embryos in H3K27me3-rich regions, even though paternal H3K9me3 is largely absent at this stage[11] (Fig. S4e).

Altogether these results show that *Eed* and H3K27me3 depletion causes reestablishment of canonical LADs and resolves allelic LAD asymmetries through an H3K9me3-independent mechanism (Fig. 3j).

Association with the nuclear lamina alone is not sufficient to remove H3K27me3

Having uncovered the strong antagonistic relationship between H3K27me3 and nuclear localization of the genome, we wondered whether forcing ncH3K27me3 regions to the NL would have an effect on chromatin state and embryonic development. To this end, we designed constructs where a triplet of Cbx7 chromodomains, which specifically bind H3K27me3[27], was fused to different inner nuclear membrane proteins: Lap2β, Emerin (Emd) or Lamin B Receptor (Lbr). These constructs were injected in the zygote stage together with Dam-LMNB1 to test for successful tethering of H3K27me3 regions via scDam&T-seq readout at the 2-cell stage. All three constructs showed increased lamina association of H3K27me3 domains (Fig. S6a). However, the Cbx7-Lap2β fusion showed the most pronounced tethering effect and was therefore selected for further experiments. Lap2β without the Cbx7 chromodomain triplet was used as a control and showed a comparable LAD profile to untreated 2-cell embryos (Fig. 4a, Fig. S6a). Surprisingly, embryos with tethered H3K27me3 showed no large-scale changes in chromatin accessibility relative to genome-lamina association changes at the 2-cell stage (Fig. S6b).

In order to disentangle the response of the maternal and paternal genome to the tethering, we repeated the experiment in embryos originating from a B6CBAF1/J x CAST/EiJ cross and calculated allele-specific LMNB1 enrichment over H3K27me3 domains. While control Lap2b embryos had low LMNB1 values over H3K27me3-rich regions, the tethering construct showed a clear enrichment (Fig. 4b, Fig. S6c-d), similar to the effect observed upon *Eed* mKO (Fig. 3b, Fig. S5c).

Strikingly, the previously defined genomic regions with increased genome-lamina association in the *Eed* mKO had a similar albeit even more pronounced increase in Dam-LMNB1 signal. Specifically, the ET-iLADs characterized by high H3K27me3 enrichment and typical LAD features were most consistently tethered to the NL (Fig. 4c).

Like the *Eed* mKO embryos, *Cbx7-Lap2β*-injected embryos show higher correlation between maternal and paternal LADs (Spearman's rho = 0.68) compared to control (Spearman's rho = 0.32), indicating decreased LAD allelic asymmetry. However, a higher degree of allelic asymmetry remained, due to unequal tethering of some regions with allele-specific H3K27me3 enrichment (Fig. S6e-f).

Having validated the successful tethering of H3K27me3 regions by Cbx7-Lap2b, we next tested whether H3K27me3 presence was compatible with NL localization. We found a clear increase in H3K27me3 localization at the nuclear periphery in *Cbx7-Lap2b*-injected 2-cell embryos in comparison to uninjected or *Lap2β*-injected embryos (Fig. 4d, Fig. S6g), suggesting that

H3K27me3 is retained upon forced relocation to the NL. We previously observed that H3K9me3 levels are reduced in the *Eed* mKO condition (Fig. 3f-g, Fig. S5f). However, it is unclear whether this effect is related to the loss of H3K27me3 or a consequence of the genomic repositioning of ET-iLADs to the NL. To disentangle these variables, we performed IF staining for H3K9me3 in embryos injected with *Cbx7-Lap2β*, in which ET-iLADs are also relocated to the NL while retaining H3K27me3. Unlike in the *Eed* mKO embryos, we observed a strong enrichment of H3K9me3 at the nuclear periphery upon tethering H3K27me3 domains to the NL (Fig. 4e, Fig. S6h). Therefore, H3K9me3 reduction at ET-iLADs upon *Eed* mKO does not appear to be caused by the change in nuclear localization of these regions, indicating that the mechanism for this loss is most likely Eed/H3K27me3-dependent. *Eed* mKO has also been described to result in the reduction of interactions between PADs, transient topological structures that are H3K27me3-rich. To investigate whether H3K27me3 tethering at the NL could disrupt PAD-PAD interactions we made use of our single-cell Dam-LMNB1 data to infer chromatin organization of regions by using a LAD coordination metric with which we can infer genome topological organization (Methods and Kind, 2015). We first confirmed that 2-cell LAD coordination reflects three-dimensional chromatin organization by comparing it to publicly available Hi-C (genome-wide chromatin conformation capture) data which showed a good correspondence (Fig. S6i). Since PADs tend to have low Dam-LMNB1 counts in control 2-cell embryos, we compared LAD coordination values exclusively between PADs in Cbx7-Lap2β and *Eed* mKO conditions which are enriched in Dam-LMNB1 values. First, we indeed could recapitulate the previously reported reduced PAD-PAD interactions in the *Eed* mKO condition which validated this approach. Next, in the Cbx7-Lap2β condition we observed clear PAD-PAD interactions (Fig. S6j). Collectively, these results indicate that it is the PRC2 and H3K27me3 depletion and not the change in nuclear localization that causes the H3K9me3 decrease or the loss of higher order chromatin structures in *Eed* mKO conditions.

Finally, we wished to test the effect of forced tethering of the ncH3K27me3 regions towards the NL on embryonic development and found no apparent effect on embryo development or gene expression up to the blastocyst stage (Fig. S6k-l).

In summary, these results show that we have developed a strategy that successfully tethers ncH3K27me3 to the NL. With this system, we could disentangle the roles of H3K27me3 loss and LAD rewiring observed in *Eed* mKO embryos. We find that, unlike *Eed* mKO embryos, tethering H3K27me3 to the NL does not result in H3K9me3 reduction in newly formed LADs and that allelic asymmetry in genome-lamina association is present in some regions with parental-specific H3K27me3 (Fig. 4f)

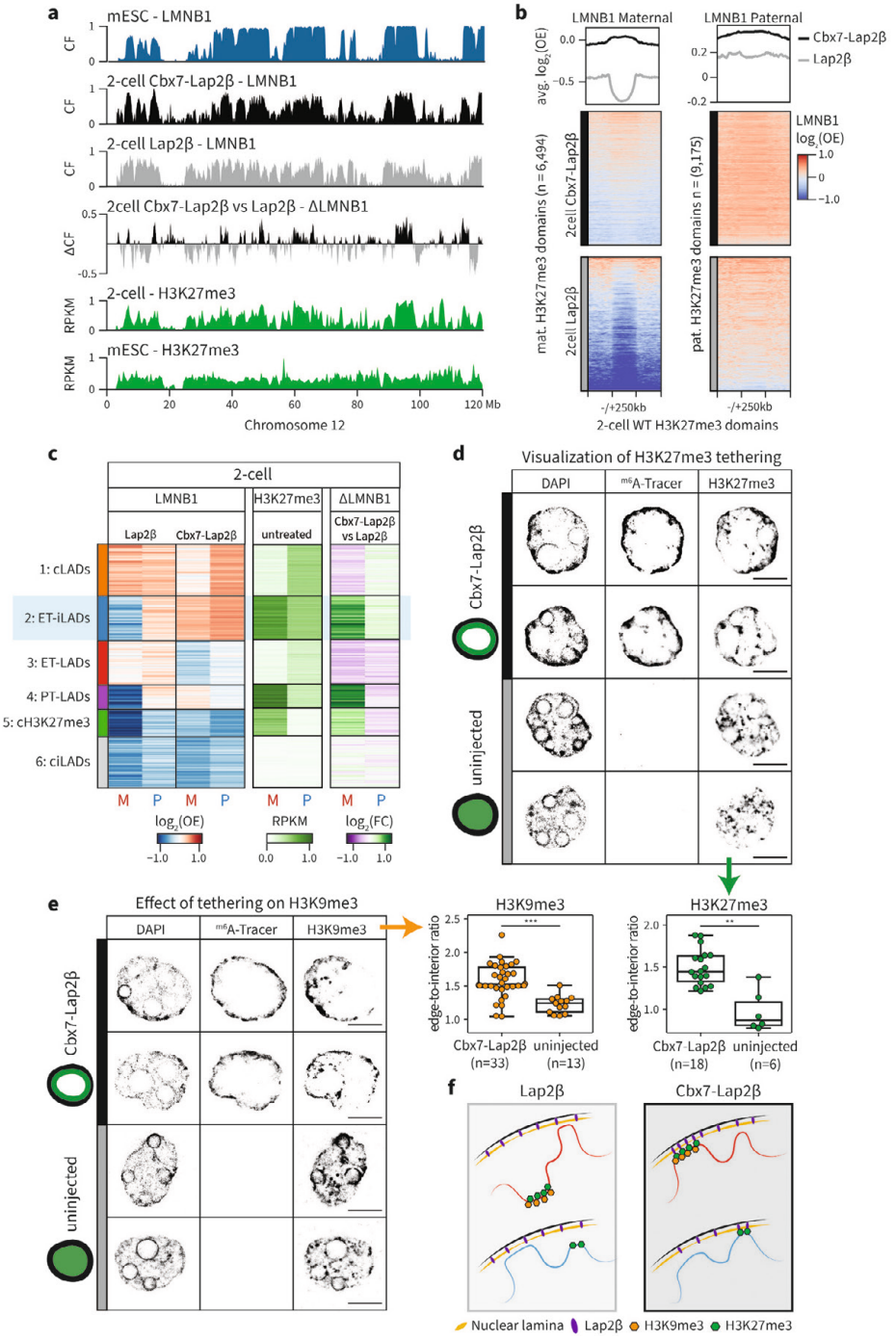


Figure 4: H3K27me3 is retained when forced to associate to the lamina

a, LMNB1 profiles of *Lap2b* and *Cbx7-Lap2b*-injected 2-cell embryos, differential LMNB1 enrichment between

the two conditions, mESC LMNB1 profile and H3K27me3 profiles (ChIP-seq[13]) over chromosome 12. **b**, Enrichment plot showing maternal (left) and paternal (right) LMNB1 enrichment for Cbx7-Lap2b and Lap2b conditions over H3K27me3 domains from the same allele and surrounding 250 kb. Heatmaps show LMNB1 signal per domain, while line plots show average enrichment over all domains. **c**, Heatmap showing allelic LMNB1 values across genomic clusters (as in Fig. 2e) for *Lap2b* and *Cbx7-Lap2b*-injected 2-cell embryos. H3K27me3 from ChIP-seq data[13] at the 2-cell stage is shown for comparison. **d-e**, DAPI, immunostaining of the ^{m6}A-Tracer, and immunostaining of H3K27me3 (d) or H3K9me3 (e) in 2-cell Cbx7-Lap2b and Lap2b embryos (scale bar = 10mm). Detection of ^{m6}A-Tracer signal indicates successful injection of the Dam-LMNB1 and Cbx7-Lap2b constructs (see Methods). Middle: Quantification of the radial (1 mm) enrichment of H3K27me3 (top) and H3K9me3 (bottom) relative to the rest of the nucleus for the immunostaining in (h). Significance was computed using Welch's two-sided t-test (H3K27me3, p = 1.6e-3; H3K9me3, p = 3.6e-7). j, Model that illustrates the effect of H3K27me3 tethering during early development. Boxplots included in (d) and (e) indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).

Antagonism between ncH3K27me3 levels and NL affinity cause unusual nuclear localization plasticity of the totipotent genome

Eed mKO and Cbx7-Lap2β embryos both showed the most gain in genome-lamina association on ET-iLADs (Fig. 4c, Fig. S5d). These regions are characterized by 1) being at the NL in other cell types, 2) having high levels of H3K27me3 and 3) displaying typical sequence traits of LADs (Fig. S4f-h). This would suggest the presence of an underlying affinity encoded in the DNA sequence that is counteracted by the presence of H3K27me3. We thus postulated that H3K27me3 and intrinsic NL affinity act as opposing 'forces' that give rise to the atypical LAD landscape characteristic of the totipotent nucleus. Having manipulated this relationship by either removing H3K27me3 or forcing it to locate to the NL, we have the tools to test this hypothesis.

As a first step, we set out to define a metric that can represent genome-lamina affinity inherent to the DNA sequence. Previously, it has been shown that Lamin proteins bind A/T-rich sequences *in vitro*[28] and that cLADs are A/T-rich[29]. These results suggest a link between A/T content and high NL affinity. Just after fertilization the paternal genome is largely devoid of histone marks and yet shows very strong canonical lamina association with high A/T content[1]. As previously suggested, the zygotic LAD profile thus likely represents a 'default' lamina association state that is encoded in the DNA sequence itself[1]. Indeed, the correlation between A/T-content and genome-lamina association was very high (Spearman's rho = 0.92) in our data, which further confirms that A/T content is a good metric to infer intrinsic NL affinity (Fig. S7a).

To evaluate the joint effects of intrinsic NL affinity and H3K27me3 on LADs at the 2-cell stage, we divided the genomic bins in nine categories based on their A/T content (low/mid/high) and H3K27me3 level (low/mid/high) (Fig. S7b-d). In the case of H3K27me3, maternal and paternal genomes were considered separately. Interestingly, the paternal allele displays a strong overlap of NL affinity and H3K27me3 enrichment making it difficult to disentangle the respective contributions of NL affinity and H3K27me3 (Fig. S7d). For this reason, we first considered the maternal allele.

When considering maternal genome-lamina association across the different categories, we found that genomic regions with low H3K27me3 and high NL affinity show the strongest genome-lamina association. Centromere-proximal regions are strongly represented in this category, explaining their unusually strong genome-lamina association at the 2-cell stage (Fig. S2d-e). Conversely, regions with high H3K27me3 and low NL affinity show the weakest genome-lamina association. In general, we find that at each level of NL affinity, genome-lamina association tends to diminish with increasing levels of H3K27me3 (Fig. 5a, left). These results corroborate our hypothesis that genome-lamina association at the 2-cell stage is determined by the presence of two antagonistic forces: the NL affinity intrinsic to the DNA sequence countered by a repellent effect of H3K27me3 presence.

Next, we analysed the changes in genome-lamina association across the different genome categories upon H3K27me3 depletion (*Eed* mKO) and peripheral tethering (Cbx7-Lap2β). As expected, genome-lamina association in the *Eed* mKO 2-cell nucleus is no longer influenced by H3K27me3 and is instead dictated by NL affinity (Fig. 5a, right). In the *Cbx7-Lap2β* injected embryos, on the other hand, lamina association became strongest for regions with both high NL affinity and high H3K27me3, suggesting that the two features reinforced each other in an additive manner to bring genomic regions to the NL (Fig. 5b, right). This indicates that expression of Cbx7-Lap2β abrogated the repellent properties of H3K27me3 and converted it to a positive force in NL tethering.

To further dissect the impact of different H3K27me3 levels on LADs, we took a closer look at the relationship between H3K27me3 and genome-lamina association for different levels of NL affinity (as measured by A/T content). We found that the antagonistic effect of H3K27me3 in control conditions is gradual and is sensitive to the level of H3K27me3 present (Fig. 5c-d). In *Eed* mKO embryos, genome-lamina association remains mostly stable per NL affinity category due to the loss of H3K27me3 (Fig. 5c). Conversely, tethering of H3K27me3 to the NL results in progressively increasing lamina association as H3K27me3 levels increase (Fig. 5d). These results indicate that the intrinsic affinity for the NL encoded in the DNA sequence is progressively countered by increasing levels of H3K27me3.

While both the *Eed* mKO and Cbx7-Lap2β directly affect the lamina association of regions with H3K27me3, we also observe changes in CF in regions where this mark is absent (Fig. 5a-d). This is likely due to competition between regions for the limited space available at the NL. Indeed, we see that the genomic fraction associating with the lamina is fairly constant across all conditions, despite drastic changes in LAD profiles (Fig. S7f). Since both conditions increase the lamina association of regions with H3K27me3, regions lacking this mark will be outcompeted and consequently have reduced lamina association.

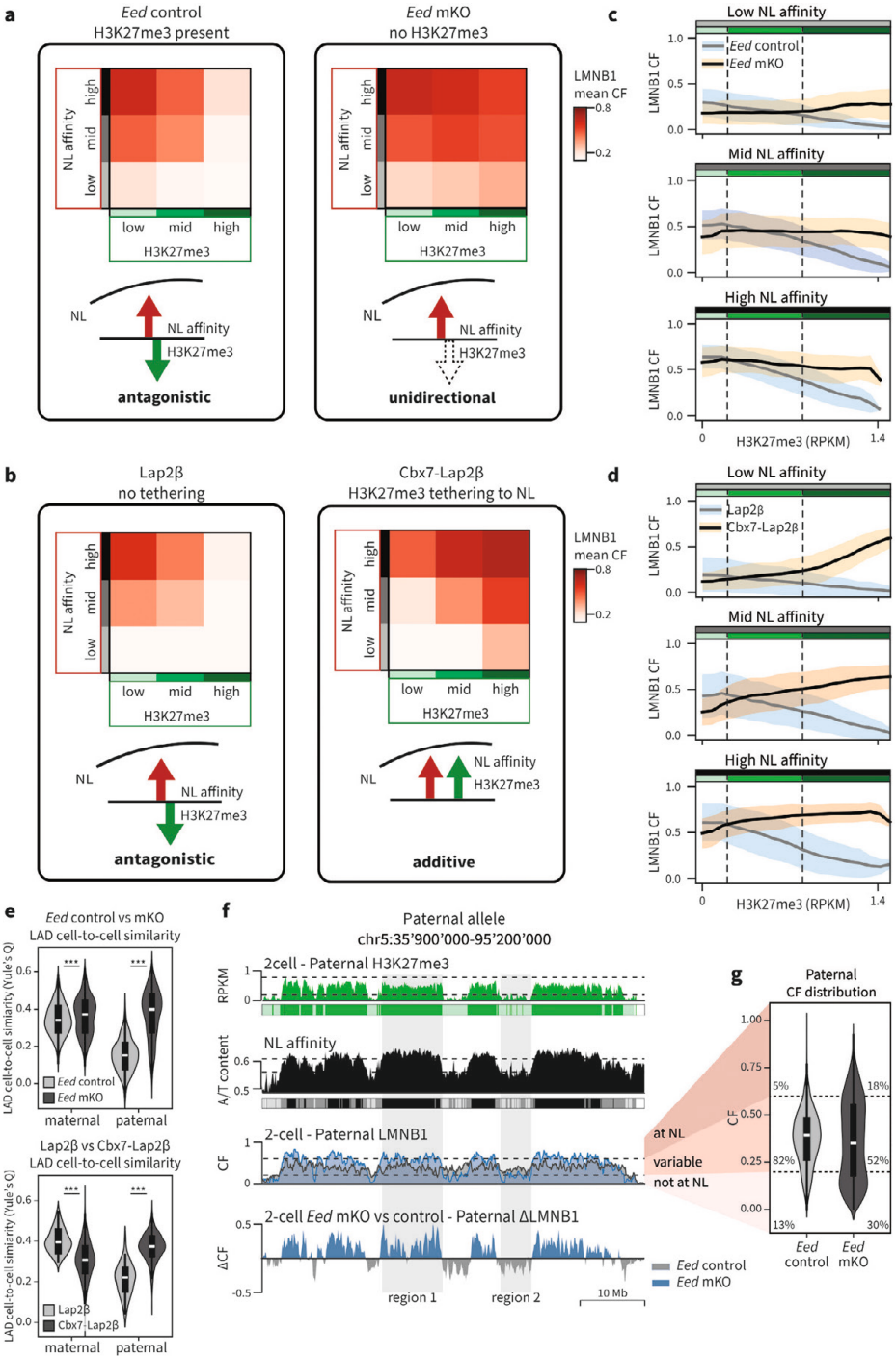

a
Eed control
H3K27me3 present
Eed mKO
no H3K27me3
NL affinity
high
mid
low
H3K27me3
NL
antagonistic
unidirectional
LMNB1
mean CF
0.8
0.2
b
Lap2β
no tethering
Cbx7-Lap2β
H3K27me3 tethering to NL
additive
c
Low NL affinity
Mid NL affinity
High NL affinity
Eed control
Eed mKO
LMNB1 CF
H3K27me3 (RPKM)
d
Lap2β
Cbx7-Lap2β
e
Eed control vs mKO
LAD cell-to-cell similarity
LAD cell-to-cell simiarity (Yule's Q)
maternal
paternal
Lap2β vs Cbx7-Lap2β
f
Paternal allele
chr5:35'900'000-95'200'000
2cell - Paternal H3K27me3
RPKM
NL affinity
A/T content
2-cell - Paternal LMNB1
CF
2-cell Eed mKO vs control - Paternal ΔLMNB1
ΔCF
region 1
region 2
10 Mb
g
Paternal
CF distribution
at NL
variable
not at NL
5%
18%
82%
52%
13%
30%

Figure 5: Antagonism between intrinsic NL affinity and H3K27me3 levels dictates genome-lamina association and leads to high LAD heterogeneity between cells of early embryos
a-b, Heatmap of average maternal LMNB1 CF values of *Eed* control and *Eed* mKO 2-cell embryos (**a**) or *Lap2b* and *Cbx7-Lap2b*-injected embryos (**b**) across nine categories of varying NL affinity and H3K27me3 level (defined in Fig. S7d). **c-d**, LMNB1 CF values of *Eed* control and *Eed* mKO (**c**) or Lap2b and Cbx7-Lap2b conditions (**d**) across increasing 2-cell H3K27me3 RPKM values. The line indicates the mean while the shaded area indicates standard deviation. **e**, Allele-specific distribution of the cell-to-cell similarity (Yule's Q) in *Eed* control and *Eed* mKO (left, maternal p = 3.2e-4, paternal p < 1e-100) and *Lap2b* and *Cbx7-Lap2b*-injected embryos (right, maternal p = 3.1e-13, paternal p = 1.7e-23). *Eed* control, n = 1,275 cell pairs; *Eed* mKO, n = 465 cell pairs; Lap2b, n = 45 cell pairs; Cbx7-Lap2b, n = 741 cell pairs. Significance between conditions was computed with the Mann-Whitney U test. **f**, Example profile on chromosome 5 of paternal H3K27me3, NL affinity, and paternal LMNB1 profiles of *Eed* control and mKO. Color-coded boxes under H3K27me3 and NL affinity refer to the different levels pictured in (a) and Figure S7d. Region 1 highlights an example where genome-lamina association is increased in *Eed* mKO, while region 2 highlights an example where it is decreased. Both regions show reduced variability. **g**, Distribution of paternal LMNB1 CF values in *Eed* control and *Eed* mKO embryos. Percentages indicate, from top to bottom, the fraction of 100-kb genomic bins with high, intermediate, and low CF values. Boxplots included in (e) and (g) indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).

LADs are particularly variable during early developmental stages which can mostly be attributed to the paternal genome (Fig. 1d,f). We thus wondered whether the opposing effects of NL affinity and H3K27me3 could play a role in the plasticity of genome-lamina association characteristic of the preimplantation embryo. We have established that we can manipulate this antagonism by either removing H3K27me3 or tethering H3K27me3 regions to the NL. We therefore tested the effect of either condition on LAD cell-to-cell variability. Surprisingly, both conditions showed a pronounced reduction of paternal LAD variability while variability on the maternal genome remains largely unaffected (Fig. 5e).

This was in first instance a surprising result given the rather low levels of H3K27me3 in the paternal allele compared to the maternal allele[13] (Fig. 2e). We thus wondered why removing H3K27me3 would have such a strong effect on paternal LAD variability. As mentioned previously, paternal genome-lamina association strongly overlap with H3K27me3 domains (example profile) at the 2-cell stage. Indeed, the vast majority of paternal regions with mid or high intrinsic NL affinity is enriched for H3K27me3 (71% and 99%, respectively; Fig. S7d). This means that a large part of the paternal genome (58%) is directly affected by the antagonism between intrinsic NL affinity and H3K27me3, explaining the particularly high level of cell-to-cell variability observed in the paternal allele during early embryo development and corroborating a tug-of-war model between NL affinity and H3K27me3 levels. The maternal genome, on the other hand, showed more evenly distributed H3K27me3 across regions of varying NL affinity (Fig. S7d). Therefore, the antagonistic effects are less prevalent in the maternal genome, resulting in reduced variability in genome-lamina association compared to the paternal genome.

Since LAD cell-to-cell variability is particularly high in the paternal genome during early embryogenesis, we decided to focus on this allele to understand how H3K27me3 impacts on the consistency with which a genomic region is found at the nuclear periphery. As previously

mentioned, we can summarize single-cell LAD information into CF scores, which reflect the fraction of cells in which a genomic region associates with the lamina. CF can thus be used as a measure of LAD variability[19], with intermediate CF values indicating higher variability. As expected, *Eed* mKO results in an increased CF in regions with mid/high NL affinity specifically, indicating that those LADs are now more consistently NL-associated between individual cells (Fig. 5f, region 1). On the other hand, regions with low intrinsic NL affinity and low H3K27me3 levels alter from intermediate CF levels to low CF values, indicating that they are now consistently excluded from associating with the NL, likely being outcompeted by higher NL affinity regions (Fig. 5f, region 2). As these two scenarios encompass the vast majority of the paternal genome (85%), there is an overall reduction in the number of regions with intermediate CF values and thus LAD variability (Fig. 5g).

Together these results show that, the overlap of two antagonistic forces, inherent NL affinity and H3K27me3, constrained by the space available at the nuclear periphery, cause the unusually high variability of lamina association across cells of the totipotent embryo.

Discussion

Here, we have profiled LADs across preimplantation stages and mESCs in single cells and identified a potential mechanism for the atypical genome-lamina interactions during early development.

High cell-to-cell variability of genome-lamina association at the 2-cell stage

We show that at 2-cell stage, the localization of the genome with respect to the nuclear periphery varies extensively across cells (Fig. 1d). While some level of LAD single-cell variability is observed in other systems[18,19], the unusually high LAD heterogeneity in early development could be related to the totipotent nature of these stages. Interestingly, paternal zygotic LADs are not as variable, potentially due to either *de novo* establishment of LAD patterns in the absence of paternally inherited chromatin modifications, or to maintenance of LAD patterns carried over from sperm, which are yet to be profiled.

Uncoupling between 3D-nuclear genome organization and gene expression at the 2-cell stage

Upon examining various chromatin marks, we did not observe variability at the same scale as for LADs (Fig. 2b), although we do not exclude that there might be heterogeneity at a smaller genomic scale (for example at the level of promoters or genes). However, these results indicate that LAD variability is neither caused by or causes heterogeneity in these genomic features. Surprisingly, we also did not observe an effect of LAD variability on transcription (Fig. S4a). We hypothesize that the totipotent cells of the early embryo show unusual uncoupling between 3D-genome organization and gene regulation.

Non-canonical iLADs at the 2-cell stage are instead enriched for H3K27me3

Maternal non-canonical H3K27me3 broad domains are located in gene-distal regions and form during oogenesis, persisting until post-implantation stages[13]. Here, we find that these non-canonical H3K27me3 regions have a strong correspondence with canonical LADs (Fig. 2c), but show decreased genome-lamina association at the 2-cell stage compared to mESCs or zygotic paternal genomes. The overlap between H3K27me3 and canonical LADs is particularly strong on the paternal allele. Interestingly, a study on the deposition of H2AK119ub1 and H3K27me3 at paternal peri-centromeric heterochromatin (PCH) showed that this process is dependent on the A/T-binding capacity of PRC1 component Cbx2 and could be inhibited by the presence of H3K9me3/HP1β, resulting in its recruitment specifically to the A/T rich paternal PCH[30]. As canonical LADs are very A/T rich and the paternal allele lacks H3K9me3 in zygote[11] (Fig. S4f), the same mechanism could explain the deposition of H3K27me3 in these regions.

H3K27me3 inhibits genome-lamina interactions

Depletion of H3K27me3 in the 2-cell embryo via maternal KO of *Eed*, a component of PRC2, prompts ncH3K27me3 regions to become LADs, demonstrating that H3K27me3 limits lamina-association. In support of this finding, a recent study reported increased lamina association of B compartment regions rich in H3K27me3 upon inhibition of another PRC2 component – Ezh2 – in K562 cells[31]. Interestingly, this inhibitory effect could explain the massive rearrangements of paternal LADs between zygote and 2-cell stages[1], as this coincides with the establishment of paternal H3K27me3[13].

Differential inheritance of H3K27me3 from sperm and oocyte results in a large degree of allelic asymmetry in this mark[13,32]. Since H3K27me3 antagonizes genome-lamina interactions, this asymmetry could, in turn, cause allelic differences in LAD profiles. Indeed, the loss of H3K27me3 restores allelic symmetry in genome-lamina association in *Eed* mKO 2-cell embryos, confirming this histone PTM is responsible for allelic LAD asymmetry (Fig. 3j).

The histone PTM H3K9me3 is typically enriched in canonical LADs and strongly overlaps H3K27me3 in the early embryo. Therefore, we hypothesized that this mark may be involved in restoring genome-lamina association in *Eed* mKO embryos. Instead, we found that *Eed* embryos display reduced H3K9me3 levels in regions relocating to the nuclear lamina, making it unlikely to play a role in the observed reorganization. Rather, H3K27me3 could be associated with the formation of H3K9me3 domains during early embryonic development, as suggested previously[11].

Lamina association is not sufficient to cause H3K27me3 depletion

Nuclear envelope proteins have been shown to interact with histone modifiers[33] and the NL environment is able to alter the activity of some promoters[33,34]. Given the clear antagonism between ncH3K27me3 and LADs in early embryos, tethering this histone PTM to the nuclear periphery could bring these genomic regions to an environment that promotes H3K27me3 depletion. However, we found that H3K27me3 was maintained upon tethering (Fig. 4d), demonstrating that this mark is not inherently incompatible with lamina association. As such,

it provided us with a tool to disentangle the effects of H3K27me3 loss and LAD reorganization observed in *Eed* mKO embryos. Through various experiments, we could thus conclude that H3K9me3 loss and transcriptional changes, observed in *Eed* mKO embryos, is a result of a PRC2/H3K27me3-mediated process rather than an effect of NL repositioning (Fig. 4e-f).

Tug-of-war between H3K27me3 and NL intrinsic affinity causes paternal LAD variability
Typically, LADs present clear sequence signatures, the most prominent being high A/T content[29]. This observation has led to the suggestion that A/T content could determine lamina association of constitutive LADs[5]. The fact that histone modifications are largely absent from the paternal genome[10-13,23] (Fig. 2e, Fig. S4e) offers a unique system where 'default' lamina association in the near absence of histone PTMs can be observed. We could thus use our zygote LAD data to confirm A/T content as a good metric for NL affinity of the DNA sequence. Categorizing genomic regions based on their NL intrinsic affinity and H3K27me3 levels allowed us to analyze genome-lamina association at the 2-cell stage with two distinct manipulations of the H3K27me3/LAD interplay. Our results showed a clear antagonistic dose-dependent relationship between NL-intrinsic affinity and 'NL-repellent' H3K27me3 that accounts for the atypical genome-lamina association unique to the early embryo. We also found the extensive overlap between high NL affinity regions and H3K27me3 in the paternal allele to explain the high levels of cell-to-cell variability that are so particular to this allele. The absence of H3K27me3 in *Eed* mKO embryos thus results in the predominance of only NL affinity-driven mechanisms, leading to more typical LAD conformations and reduced cell-to-cell heterogeneity.

Conclusion

The study of epigenetics, nuclear organization and transcription during the first days of embryonic development has greatly benefited from the development of novel low-input technologies. These have brought much needed insight into the nuclear events that occur from the moment gametes fuse to form a totipotent zygote until implantation[8-13]. An overall view of non-canonical epigenetic features and major restructuring of genomic organization has emerged, but little is known about how these processes are connected and what their role in embryonic development may be. Here, we propose a model whereby H3K27me3 antagonizes genome-lamina association during preimplantation development, thereby causing atypical organization, allelic asymmetry and cell-to-cell heterogeneity of genome-lamina association. In the present study, we demonstrate that this interplay between Polycomb and genome-lamina association is mechanistically involved in the processes that so dramatically reorganize the nuclear architecture during preimplantation development.

Acknowledgements

We would like to thank all the members of the Kind laboratory for their comments throughout the project and their critical reading of the manuscript. We thank Evgeniy A. Ozonov for advice

on data analysis and Grigorios Fanourgakis for collection of additional *Eed* mKO embryos. This work was supported by an ERC Starting grant EpiID (ERC Stg EpiID-678423) and ERC Consolidator grant FateID (ERC CoG-101002885) and an NWO-ENW VIDI grant (161.339). The Oncode Institute is partially funded by the KWF Dutch Cancer Society. I.G. was supported by an EMBO Long-Term Fellowship (ALTF1214-2016), Swiss National Science Fund grant (P400PB_186758) and NWO-ENW Veni grant (VI.Veni.202.073). The lab of A.H.M.F.P. has received funding from the Novartis Research Foundation and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement ERC-AdG 695288 - Totipotency). In addition, we would like to thank the Hubrecht Sorting Facility as well as the Utrecht Sequencing Facility (USEQ), subsidized by the University Medical Center Utrecht and the Netherlands X-omics Initiative (NWO project no. 184.034.019).

Author contributions

I.G., F.J.R. and J.K. designed the study. F.J.R. performed all data analysis with input from I.G. All embryo and scDam&T-seq experiments were performed by I.G. unless otherwise stated with assistance from C.K.V, F.C.G. and R.E.v.B.. Zygote injections were performed by C.K.V. and J.Ko. except when using the *Eed* mKO mouse line. Y.K.K. performed the embryo experiments for the *Eed* mKO line and corresponding control with supervision from A.H.M.F.P.. S.J.A.L. and E.B. generated the mESC line expressing Dam-LMNB1 and performed the mESC scDam&T-seq experiment. I.G. and F.J.R wrote the manuscript with editing by J.K. All authors reviewed and edited the manuscript.

Methods

RNA synthesis

All constructs were cloned into an *in vitro* transcription vector previously described in Borsos *et al.* (2019)[1], linearized, purified using the QIAquick PCR Purification Kit (Qiagen) and transcribed using the T3 mMessage mMachine kit (Invitrogen, AM1348) according to manufacturer instructions. The synthesized RNA was purified using the MEGAclear kit (Invitrogen, AM1908) or the Lithium Chloride RNA precipitation method and eluted in10 mM Tris-HCl pH 7.5 and 0.1 mM EDTA.

Animal care and zygote injection

All animal experiments were approved by the animal ethics committee of the Royal Netherlands Academy of Arts and Sciences (KNAW) under project license AVD801002016728 and study dossiers HI173301 and HI213301. Embryos were collected from B6CBAF1/J females crossed with CAST/EiJ males for the hybrid experiments and B6CBAF1/J males for non-hybrid experiments. The *Eed* floxed mouse ($Eed^{fl/fl}$) was provided by Prof. Stuart H. Orkin. To obtain *Eed* maternal KO embryos, we crossed $Eed^{fl/fl}$;$Gdf9^{Cre}$ females (on a C57BL/6J background) with WT T males.

For all crosses, 7- to 10-week-old females were superovulated by injecting pregnant mare serum gonadotropin (PMSG, 5IU, MSD, Cat#A207A01) and human chorionic gonadotropin (hCG, 5IU, MSD, Cat#A201A01). For *in vitro* fertilization (IVF), spermatozoa from JF1/MsJ males were capacitated in Human Tubal Fluid medium (Merck Millipore, Cat#MR-070-D) supplemented with 10 mg ml^{-1} Albumin (Sigma, Cat#A-3311) (HTF-BSA) for 1h preceding insemination. MII oocytes were collected in the insemination medium (HTF-BSA) and capacitated spermatozoa were added for fertilization. The insemination starting time point was termed as 0 hours post-fertilization (hpf).

When using standard mating, zygotes were injected about 24h post-hCG.
For both IVF and normal mating mRNA was microinjected in the cytosol of the zygote at 10hpf. A full description of the used constructs, corresponding concentrations and induction conditions is described in Supplementary Table 5.

Injected zygotes were cultured in KSOM for hybrid crosses (Sigma, Cat#MR-106-D) or M16 medium (Sigma, Cat#M7292) for non-hybrid crosses covered with mineral oil (Sigma Cat# M8410) at 37°C with 5% CO_2 and 5% O_2 air.

To increase the quality of DamID signal, untethered Dam and Dam-$Cbx1_{CD}$ constructs were fused to an ERT2 domain so that the fusion protein would be translocated to the nucleus upon 4-Hydroxytamoxifen addition (4-OHT, Sigma, Cat#SML1666) (Supplementary Table 5).

Embryo collection, dissociation and sorting
Embryos were collected by mouth pipetting at 29-31 hours post-hCG for the zygote stage, 52-55 hours post-hCG for the 2-cell stage and 75-78 hours post-hCG for the 8-cell stage (Supplementary Table 5). The zona pellucida was removed using Tyrode's acid (Sigma, Cat#T1788), washed in M2 medium (Sigma, Cat#MR-015) and placed in TryplE (Gibco, Cat#12605010) where embryos were dissociated into single cells one by one and placed in M2 medium before single-cell collection into a 384-well plate containing 5uL of mineral oil and 100nL of barcoded polyadenylated primers.

scDam&T-seq processing
All scDam&T-seq steps were performed as previously described[35]. Briefly, cells were lysed and reverse transcription was performed followed by second-strand synthesis in order to convert the RNA of the cell into cDNA. After a proteinase K step, methylated GATCs resulting from Dam enzyme activity were specifically digested with DpnI (scDam&T) and 25nM (scDam&T) double-stranded adapters containing cell-specific barcodes were ligated.

At this point, cells with non-overlapping barcodes were pooled together to undergo in vitro transcription which amplifies both the transcriptional and genomic product in a linear manner due to a T7 promoter to both the double-stranded DamID adapters and the polyadenylated primers. The resulting amplified RNA (aRNA) was reverse transcribed and

library preparation was performed as previously described[36]. Libraries were sequenced on the Illumina NextSeq500 (75-bp reads) or NextSeq2000 (100-bp reads) platform. For scDam&T-seq processing of mESC cells specifically, half volumes were used in all reactions to reduce overall processing costs.

Immunofluorescence staining

After removal of the zona pellucida with Tyrode acid, embryos were fixed in 4% paraformaldehyde (PFA) in PBS at room temperature for 15 minutes and permeabilized in PBS with 0.5% Triton X-100 for 20 min, at room temperature. Embryos were then incubated with blocking solution (2% bovine serum albumin (BSA) in PBS) for 1 h or more. Incubation with primary antibody and m6A-Tracer protein[37] were performed overnight at 4°C. When m6A-Tracer protein was used, the overnight incubation was followed by a 1h incubation at room temperature with anti-GFP antibody. For all stainings, secondary antibodies incubations were also performed for 1h at room temperature followed by DAPI staining 3 µg/mL for 20 min. Samples were mounted on glass slides using a spacer and VECTASHIELD Antifade mounting medium (Vector Laboratories). Primary antibodies used were: rabbit anti-H3K27me3 (Cell Signalling Technology, C36B11, lot 19) at 1:200, rabbit anti-H3K9me3 (abcam, ab8898, lot GR3281994-1) at 1:300, chicken anti-GFP (Aves, GFP-1020, lot GFP697986) at 1:1000, mouse anti-Flag (Merck, F1804, SLCK5688) at 1:500. Secondary antibodies were all used at 1:500: Alexa Fluor 488 Goat anti-Chicken (Invitrogen, A11039, lot 2180688), 532 Alexa Fluor Goat anti-Mouse (Invitrogen, A32727, lot SF251136) and Alexa Fluor 647 Goat anti-Rabbit (Invitrogen, A21244, lot 2179230). Purified m6A-Tracer protein[37] was used at 1:1000. Embryos were scanned with a 0.3 µm distance between optical sections. Imaging was performed on a Leica TCS SP8 laser scanning confocal microscope with a 63X oil-immersion objective. Images were processed using Fiji.

Image quantification

Image quantification was performed in Python 3.8 using scipy (v. 1.7.1) and scikit-image (v. 0.19.2). First, nuclei were identified based on DAPI across stacks using Multi-Otsu thresholding and the z-stack with the largest nuclear surface was selected automatically, after which the selection was manually inspected and adjusted in case of artefacts. Nuclei with condensed chromatin or dim DAPI staining were excluded. To determine the relative levels of H3K27me3 (Fig. S5a), the average H3K27me3 signal in the nucleus was compared to the average signal in the background (excluding other nuclei and artefacts). To determine the enrichment of H3K9me3 (Fig. 3i, Fig. 4e, Fig. S6h) or H3K27me3 (Fig. 4d, Fig. S6g) at the nuclear periphery, the average signal within 1 µm of the nuclear edge was compared to the average signal in the rest of the nucleus.

Viability assay

Number of embryos that reached the blastocyst after *Cbx7-Lap2β* or *Lap2β* injection were assessed at embryonic day 3.5.

Cell culture

Cell lines were grown in a humidified chamber at 37 °C in 5% CO_2 and were routinely tested for mycoplasma. Mouse F1 hybrid Cast/EiJ (paternal) x 129SvJae (maternal) embryonic stem cells (mESCs; a gift from the Joost Gribnau laboratory) were cultured on 6-well plates with irradiated primary mouse embryonic fibroblasts (MEFs) in mESC culture media (CM) defined as follows: Glasgow's MEM (G-MEM, Gibco, 11710035) supplemented with 10% FBS, 1% Pen/Strep, 1x GlutaMAX (Gibco, 35050061), 1x MEM non-essential amino acids (Gibco, 11140050), 1 mM sodium pyruvate (Gibco, 11360070), 0.1 mM β-mercaptoethanol (Sigma, M3148) and 1000 U/mL ESGROmLIF (EMD Millipore, ESG1107). mESCs were alternatively cultured in feeder-free conditions on gelatin coated plates (0.1% gelatin, in house) in 60%-BRL medium, defined as a mix of 40% CM medium (as defined) and 60% conditioned CM medium (incubated 1 week on Buffalo Rat Liver cells), supplemented with 10% FBS, 1% Pen/Strep, 1x GlutaMAX, 1x MEM non-essential amino acids, 0.1 mM β-mercaptoethanol and 1000 U/mL ESGROmLIF. Cells were split every 2-3 days and medium was changed every 1-2 days. This mESC cell line does not contain a Y chromosome.

Generation of mouse embryonic stem cell lines

The stable clonal F1 hybrid mESC line expressing the Dam-LaminB1 fusion protein was generated from an EF1α-Tir1-IRES-neo expressing mother line (generated with lentiviral transduction)[18]. The Dam construct was CRISPR targeted into this line by knocking in mAID-Dam in the N terminus of the LMNB1 locus. The donor vector (designed in house, generated by GeneWiz) carried the Blasticidin-p2A-HA-mAID-Dam cassette, flanked on each side by 1000-bp homology arms of the endogenous LMNB1 locus (pUC57-BSD-p2A-HA-mAID-Dam). The Cas9/guide vector was the p225A-LmnB1-spCas9-gRNA vector, with a guide RNA inserted to target the 5'UTR of the LMNB1 locus (sgRNA: 5' CACGGGGGTCGCGGTCGCCA 3'). For transfections in general, cells were cultured on gelatin-coated 6-well plates in 60% BRL-medium at 70%–90% confluency. Cells were transfected with Lipofectamin2000 (Invitrogen, 11668030) according to the supplier protocol with 1.5 µg donor vector and 1.5 µg Cas9/guide vector. At 24 hours after transfection, GFP positive cells were sorted on a BD FACsJazz Cell sorter and seeded on gelatin-coated plates in 60% BRL-medium. 48 hours after sorting, cells were started on antibiotic selection with 60% BRL-medium containing 3.0 µg/mL Blasticidin (ThermoFisher, A1113903) and 0.5 mM indole-3-acetic acid (IAA, Sigma, I5148) and cells were refreshed every 2-3 days. From this point onwards, 0.5 mM IAA is added to the medium during normal culturing conditions to degrade the mAID-Dam-Lamin B1 fusion protein via the auxin protein degradation system[38]. After 6 days of antibiotic selection, single cells were sorted into 96-well plates containing MEFs using the BD FACsInflux Cell sorter and grown without antibiotic selection in CM medium with 0.5 mM IAA. Clones grew out in approximately 10 days and were screened for correct integration by PCR with primers from Dam to the LMNB1 locus downstream of targeting construct; fw-TTCAACAAAAGCCAGGATCC and rev-TAAGGAATCTGGTGCACAGAACACC. The heterozygous expression of the Dam-Lamin B1 fusion protein was further confirmed by Western blot using an anti-HA antibody at 1 in 5000 dilution (Abcam, ab9110) and an anti-LaminB1 antibody at 1 in 5000 dilution (Abcam, ab16048). To prevent silencing of the EF1α-

Tir1-IRES-neo construct due to the flanking lentiviral construct sequences, the Tir1 construct was additionally knocked-in into the TIGRE locus using CRISPR targeting. This integration was generated by co-transfection of the donor vector pEN396-pCAGGS-Tir1-V5-2A-PuroR TIGRE (Addgene plasmid, #92142) and Cas9-gRNA plasmid pX459-EN1201 (backbone from Addgene plamid #62988, guide from Addgene plasmid #92144[39], sgRNA: 5' ACTGCCATAACACCTAACTT 3'). Cells were transfected with Lipofectamine3000 (ThermoFisher, L3000008) according to the supplier protocol with 2 µg donor vector and 1 µg Cas9-gRNA vector. At 24 hours after transfection, GFP positive cells were sorted on a BD FACsJazz Cell sorter and seeded on gelatin-coated plates in 60% BRL-medium. 48 hours after sorting, cells were started on antibiotic selection with 60% BRL-medium containing 0.8 µg/mL Puromycin (Sigma, P9620) and 0.5 mM IAA. Cells were refreshed every 2-3 days and selected for 5-10 days. The Tir1-puro clones were screened for the presence of Tir1 by PCR from the CAGG promoter to Tir1 with the primers fw-CCTCTGCTAACCATGTTCATG and rev-TCCTTCACAGCTGATCAGCACC, followed by screening for correct integration in the TIGRE locus by PCR from the polyA to the TIGRE locus with primers fw-GGGAAGAGAATAGCAGGCATGCT and rev-ACCAGCCACTTCAAAGTGGTACC. The Tir1 expression was further confirmed by Western blot using a V5 antibody (Invitrogen R960-25). Upon further characterization of the best clone, a 70-bp deletion was found directly after the transcription start site the wildtype LMNB1 allele, causing frameshift, which was most likely the result of the CRISPR targeting. Cell viability and growth rates were not visibly affected. This deletion was repaired using 200 bp ssDNA utramere oligo's (IDT) with 65 bp homology arms on each side of the deletion as donor and a p225A-LmnB1-repair-spCas9-gRNA vector (sgRNA: 5' GCGGGGGCGCTACAAACCAC 3'). Cells were transfected with Lipofectamine 3000 according to the supplier protocol with 1.5 µg donor oligo and 1 µg Cas9-gRNA vector. At 24 hours after transfection, GFP positive cells were sorted into 96-well plates containing MEFs using the BD FACsJazz Cell sorter and grown without antibiotic selection in CM medium. Clones were screened for correct repair of the wildtype LMNB1 allele by PCR around the original deletion with the primers fw- ACTCACAAAGGGCGTCTGGC and rev- GTGACAATCGAGCCGGTACTC. Correct expression of the mAID-Dam-Lamin B1 fusion protein as well as the wildtype Lamin B1 protein was confirmed using Immunofluorescence staining using an anti-HA antibody at 1 in 500 dilution (Cell Signaling Technologies, C29F4) and an anti-LaminB1 antibody at 1 in 500 dilution (Abcam, ab16048), followed by confocal imaging. All successfully repaired clones were subsequently screened for their level of induction upon IAA removal by m6A-PCR, evaluated by gel electrophoresis[6,40], followed by DamID2 sequencing in bulk[35,40], to select a heterozygous clone with a correct karyotype with the best signal-to-noise ratio of enrichment over LAD regions. This clone is labelled as F1ES mAID-Dam-LaminB1 #2B4.

mESC Dam-Lamin B1 induction and FACS sorting for single-cell experiments

Expression of the mAID-Dam-Lamin B1 fusion protein in the F1ES cell line was suppressed by addition of 0.5 mM IAA during standard culturing. When plated for scDam&T-seq experiments, the cells were passaged at least two times in feeder-free conditions on 6-well plates coated with 0.1% gelatin in 60%-BRL medium. Cells were kept at 1 mM IAA for the final 48 hours before the start of the experiment. 6 hours before harvesting of cells, the IAA was removed by washing

three times with PBS and refreshing with 60%-BRL medium without IAA. FACS was performed on BD FACSJazz or BD FACSInflux Cell Sorter systems with BD Sortware. mESCs were harvested by trypsinization, centrifuged at 300 g, resuspended in 60%-BRL medium containing 10 mg/mL Hoechst 34580 (Sigma, 63493) per 1x10^6 cells and incubated for 45 min at °C in 5% CO_2. Prior to sorting, cells were passed through a 40-mm cell strainer. Propidium iodide (1 mg/mL) was used as a live/dead discriminant. Single cells were gated on forward and side scatters and Hoechst cell cycle profiles. Index information was recorded for all sorts. One cell per well was sorted into 384-well hard-shell plates (Biorad, HSP3801) containing 5 µL of filtered mineral oil (Sigma, 69794) and 50 nL of 1.5 mM barcoded CEL-Seq2 primer[18,35].

Processing of scDamID and scDam&T-seq data

Data generated by the DamID and scDam&T-seq protocols was largely processed with the workflow and scripts described in Markodimitraki et al. (2020)[35] (see also www.github.com/KindLab/scDamAndTools). The procedure is described in short below.

Demultiplexing

All reads are demultiplexed based on the barcode present at the start of R1 using a reference list of barcodes. In the case of scDam&T-seq data, the reference barcodes contain both DamID-specific and CEL-Seq2-specific barcodes. In the case of the scDamID data, the reference barcodes only contain DamID-specific barcodes. Zero mismatches are allowed between the observed barcode and reference. The UMI information, also present at the start of R1, is appended to the read name.

DamID data processing

DamID reads are aligned using bowtie2 (v. 2.3.3.1)[41] with the following parameters: "--seed 42 --very-sensitive -N 1" to the mm10 reference genome. In the case of paired-end data (scDam&T-seq), only R1 is used as that contains the digested GATC site. The resulting alignments are then converted to UMI-unique GATC counts by matching each alignment to known strand-specific GATC positions in the reference genome. Any reads that do not align to a known GATC position or have a mapping quality smaller than 10 are removed. Up to 4 unique UMIs are allowed for single-cell samples to account for the maximum number of alleles in G2. Finally, counts are binned at the desired resolution.

CEL-Seq2 data processing

CEL-Seq2 reads are aligned using hisat2 (v. 2.1.0)[42] with the following parameters: "--mp '2,0' --sp '4,0'". For the alignment, only R2 is used, as R1 contains the sample barcode, UMI and poly-A tail, which have already been processed during demultiplexing. As reference, the mm10 reference genome and the GRCm38 (v. 89) transcript models are used. Alignments are subsequently converted to transcript counts per gene with custom scripts that assign reads to genes similar to HTSeq's[43] htseq-count with mode "intersection_strict".

Allele-specific alignment of DamID and CEL-seq2 reads

In the case of samples derived from hybrid crosses, we used strain-specific SNPs to assign reads to a parent. For this, we obtained SNP information from the Mouse Genomes Project of the Sanger Wellcome Institute for all used strains except for JF1/Ms, which were obtained from the MoG+ website of the RIKEN BioResource Center (https://molossinus.brc.riken.jp/mogplus/#JF1). These SNPs were subsequently substituted in the mm10 reference genome to generate strain-specific reference files. DamID and CEL-seq2 reads were subsequently aligned to the reference files of both strains as described above. Since all hybrid data was generated with scDam&T-seq, paired-end data was available for the DamID readout and both R1 and R2 were used in aligning to maximize SNP coverage. Using a custom script, the alignments of each read to the two genotypes were subsequently evaluated w.r.t. number of mismatches and alignment score. The read was then attributed to the better scoring genotype. In the case of a tie (i.e. equal number of mismatches and same alignment score), the read was considered to be ambiguous. This procedure results in three files for each sample: one alignment file for each genotype and one file with ambiguous reads. For the samples derived from the B6CBAF1/J x CAST/EiJ cross, SNPs from three different backgrounds can be present: CBA/J and C57BL/6J SNPs from the B6CBAF1/J mother and CAST/EiJ from the father. Reads derived from this cross were thus aligned to the three reference genomes representing these strains and split based on their alignment scores as described. Reads attributed to CBA/J and/or C57BL/6J were considered as maternal reads, reads attributed to CAST/EiJ as paternal reads, and reads tying between CAST/EiJ and CBA/J or C57BL/6J as ambiguous.

Processing allele-specific DamID and CEL-seq2 read to UMI-unique counts

For both CEL-seq2-derived and DamID-derived reads information from R2 is used to attribute them to a genotype. However, in both cases, the IVT and fragmentation steps in the scDam&T-seq protocol can result in copies of the original mRNA/DNA molecule of different lengths and thus different R2 sequence content. As a result, different copies of the same molecule sometimes overlap SNPs and sometimes do not. For this reason, it is import to perform UMI flattening per gene or GATC position for all alignment files (both genotypes and ambiguous) simultaneously. Per gene or GATC position, only one unique UMI is allowed across the genotypes. If a UMI was observed for one genotype and in the ambiguous reads, the unique count was attributed to the genotype. If a UMI was observed for both genotypes, the unique count was considered to be ambiguous. For this procedure, modified versions of the DamID and CEL-seq2 counting scripts were used that consider all three alignment files in parallel. Counting was otherwise performed as described above.

Filtering of DamID data

Samples were filtered w.r.t. their DamID readout based on the number of observed unique GATCs and their information content (IC). The IC is a measure for the amount of true signal relative to the amount of background in the sample. The background is determined based on a comparison of the observed signal with the density of mappable GATCs in the genome. The procedure is explained in detail in Rang and de Luca (2022)[22] and the code can be found

on GitHub (https://github.com/KindLab/EpiDamID2022). Since the fraction of the Dam-methylated genome varies per Dam-construct and per embryonic stage, the thresholds for the number of unique GATCs and IC were fine-tuned per dataset (Supplementary Table 6).

In the case of samples derived from hybrid crosses, DamID data was additionally filtered based on the presence of both a maternal and paternal allele. Specifically, at least 25% of allele-specific counts should come from each parent. In practice, this resulted in the removal of samples that exclusively had maternal-derived material, likely due to the presence of unfertilized oocytes undergoing spontaneous parthenogenesis. We observed no samples containing >75% paternal-derived material. For analyses using data of the combined alleles, the same filtering on unique GATCs and IC was applied. For analyses using allele-specific data, only samples were used that had a total number of allele-specific GATC counts equal to the general depth threshold of that condition. We performed this additional selection to prevent high levels of noise due to sparsity in allele-specific data. The numbers of cells that passed the aforementioned thresholds are documented in Fig. S1b and Supplementary Tables 1-2 and unique number of GATC distribution of Dam-LMNB1-expressing cells that passed those thresholds is illustrated in Fig. S1c.

For genome-wide analyses, we additionally performed filtering on the genomic bins that were included. For analyses that were not allele-specific, we excluded all genomic bins that contain fewer than 1 mappable GATC per kb. For allele-specific analyses, we additionally removed bins for which less than 10% of the contained GATCs could be attributed to an allele. In addition, we removed bins for which we empirically observed that 98% of allele-specific DamID counts were attributed to only one allele. Finally, for all analyses, we excluded chromosome X and Y.

Filtering of CEL-seq2 data

Samples were filtered w.r.t. their CEL-seq2 readout based on the observed number of unique transcripts (^{3}3,000), the percentage of mitochondrial transcripts (<15%), and the percentage of ERCC spike-in derived reads (<0.5%). For all stages and constructs these thresholds were the same. In addition, hybrid samples that were suspected to have undergone spontaneous parthenogenesis based on their DamID readout (see above) were also excluded from transcriptional analyses. This filtering could not be performed based on the transcriptional readout, since the vast majority of transcripts at the zygote and 2-cell stage are maternally contributed.

Computing DamID binary contacts

Single-cell count tables were further processed to binary contacts, which give an indication for each genomic bin whether a sample had an observed contact with the Dam construct. To determine binary contacts, samples were binned at 100,000-bp resolution and depth normalized by log(counts/Scounts * 10,000 + 1). Allele-specific files were normalized for the total number of counts attributed to *either* allele. Subsequently, the samples were smoothened with a Gaussian kernel (s.d. 150 kb). Both the large bin size and smoothing minimize noise that

may occur in sparse single-cell samples. The observed signal was subsequently compared with a control, which has been depth normalized and smoothed in a similar manner. In the case of the Dam-LMNB1 and Dam-only constructs, the control is the density of mappable GATCs. In the case of Dam-αH3K27me3 and Dam-Cbx1$_{CD}$, the control is the average single-cell signal of all Dam-only samples of the same embryonic stage, since these constructs are free-floating in the nucleus and are more prone to accessibility biases[22]. Contacts were then called when the difference between the observed signal and the control was bigger than 0 for Dam-LMNB1 and Dam-only, or bigger than log(1.1) » 0.95 for the remaining constructs.

Contact frequency and in silico population profiles

Contact frequency (CF) was determined for each genomic bin as the fraction of single-cell samples with an observed contact in that bin. As a result, CF values range between 0 and 1. Since binary contacts are only determined at a 100-kb resolution, CF profiles are only available at this resolution. For analyses requiring higher resolutions, we generated in silico population profiles by combining the count data of all single-cell samples per condition. The in silico population profiles were subsequently depth normalized (RPKM), normalized for a control, and log2-transformed to give log2 observed-over-expected (log2OE) values. Allele-specific files were depth normalized for the total number of counts attributed to *either* allele. In the case of Dam-LMNB1 and Dam-only, the control is the density of mappable GATCs. For the other constructs, the control is the in silico population data of the Dam-only samples.

Processing of published data

Accession numbers of all public datasets used are described in Supplementary Table 4.

ChIP-seq and CUT&RUN

Reads were aligned using bowtie2 (v. 2.3.3.1) with the following parameters: "--seed 42 --very-sensitive -N 1". Indexes for the alignments were then generated using "samtools index" and genome coverage tracks were computed using the "bamCoverage" utility from DeepTools (v. 3.3.2)[44] with the following parameters: "--ignoreDuplicates --minMappingQuality 10". For samples derived from hybrid crosses, a similar strategy was used as for our own scDam&T-seq data to attribute reads to alleles: SNPs were incorporated into the mm10 reference genome to generate parental-specific references. Reads were aligned to both genomes, after which reads were attributed to a specific parent or ambiguous alignment files based on the number of mismatches and alignment scores. These three alignment files were then separately processed with DeepTools. RPKM depth-normalization was performed for all samples, where allele-separated data was normalized for the total number of allelic reads. Samples were normalized for input-control when available and log2-transformed to give log2OE values. To correct for differential levels of allele assignment across the genome, allele-specific the RPKM value of each bin was divided by the fraction of reads in the bin for which allele-assignment was possible.

DamID

Data from our previous study (Borsos et al., 2019; available on GEO under GSE112551) was reprocessed using the same procedures as used for the current data. Since this data was generated with the first version of the scDamID protocol[19], no UMIs are present to identify PCR duplicates. To limit amplification artefacts only 1 count per strand-specific GATC position was maintained.

Hi-C

Published Hi-C data was obtained from GEO (GSE82185) and processed from raw sequencing reads to interaction matrices using Hi-C Pro (v. 2.11.4) using the recommended workflows for non-allelic and allele-separated data. The obtained interaction matrices were subsequently converted to ".cool" format using the "hicConvertFormat" command from HiCExplorer (v. 2.2.1.1). The matrices were subsequently normalized and corrected for biases using the "cooler balance" functionality from Cooler (v. 0.8.11)[45]. Further processing and visualization of the normalized Hi-C matrices was performed using CoolTools (v. 0.5.1)[46]. Compartment scores were computed with cooltools using normalized interaction matrices at a resultion of 100 kb.

Single-cell DamID analyses

Single-cell DamID UMAP

The UMAPs based on the single-cell DamID readout in Figure 1b and Figure S3c were generated by performing a PCA on the data and selected the top PCs based on the explained variance ratio (PC1-10). These PCs were used as an input to compute the UMAP. In the case of Fig. S3c, the maternal and paternal readouts of all samples were treated as separate samples. Consequently, each cell appears twice in the UMAP: once with the maternal readout and once with the paternal readout.

Cell-to-cell LAD similarity

Cell-to-cell LAD similarity was computed based on the binary contact data of all autosomal chromosomes. To minimize the influence of differential sparsity and noise between LMNB1 datasets, we downsampled all samples being compared to a common threshold and binned data at 1-Mb resolution. These thresholds were: 40,000 total unique GATCs (Fig. 1d) and 10,000 allele-specific GATCs (Fig. 5e). For comparison between alleles of the same sample (Fig. 1f), no downsampling was performed, as depth-related artefacts are identical for the two alleles. We used Yule's Q as a metric of similarity between cells: $(N_{00}N_{11} - N_{01}N_{10})/ (N_{00}N_{11} + N_{01}N_{10})$, where N_{11} is the number of genomic bins where both samples had a contact, N_{00} the number of bins where neither sample had a contact, and N_{01} and N_{10} the number of bins where one sample had a contact and the other did not.

Cell-to-cell similarity comparison between Dam-constructs

Different Dam-constructs result in data with vastly different distributions, sparsity and noise levels that will impact the similarity score (Yule's Q), for which downsampling cannot correct.

To mitigate the influence of these factors on the similarity score for different Dam-constructs, we devised a control for each condition that simulated the expected Yule's Q scores based on technical variability alone. For this, we combined the data of each condition per batch (i.e. per sequencing library) and subsequently downsampled the data to generate mock single-cell samples with a number of unique GATCs equal to the actual single-cell data. We subsequently normalized and binarized the simulated single-cell data in an identical manner to the original data. The simulated dataset should thus represent samples that display the same level of technical variability as the original sample, without showing any true biological variation. Finally, we normalized the similarity score of each pair of cells by subtracting their score in the simulated dataset.

LAD coordination

To quantify the coordinated association of genomic bins with the NL, we computed the Pearson correlation between all pairs of genomic bins, as previously described[19]. To control for the influence of technical factors (e.g. depth and CF distributions), we compared the observed correlations to those observed when using shuffled binary contact tables. This shuffling was performed in such a way that both marginals (i.e. the CF of each 100-kb bin and the total number of contacts in each cell) remained intact, using a published algorithm[47]. For each condition, we performed 1,000 randomizations of the binary contact matrix and computed the bin-bin correlations of the resulting matrices. The observed mean and standard deviation of the correlation matrices were then used to standardize the true bin-bin correlation matrix. The standardized LAD coordination values were compared to normalized Hi-C interaction frequencies (Fig. S6i) by correlating their values up to a distance of 50 Mb. The average LAD coordination over pairs of PADs (Fig. S6j) was based on code from Coolpup.py[48].

Clustering of genomic bins

Datasets used for genomic bin clustering (100 kb) was based on LMNB1 data of hybrid zygote, 2-cell, and 8-cell embryos, LMNB1 data of mESC, H3K27me3 ChIP-seq data of hybrid PN5 zygote, late 2-cell and 8-cell embryos, and H3K27me3 ChIP-seq data of mESC (H3K27me3 data from Zheng et al. (2016), GSE76687). Allele-separated data was used for all samples, except H3K27me3 mESC. CF values were used for DamID data, log10-transformed RPKM values were used for ChIP-seq data. Only data of autosomal chromosomes was included. Genomic bins were further filtered based on the criteria described in DamID and ChIP-seq methods sections. Finally, bins were removed if they overlapped >10% a "High Signal Region" or >80% a "Low Mappability Region" as defined in the ENCODE mm10 blacklist. For the autosomal chromosomes, this left ~88.2% of the genomic bins to be included in the clustering.

Prior to clustering, values were standardized, clipped to a range from –2.5 to 2.5, and a PCA was performed to remove redundancy in the data. The top PCs were selected based on the explained variance ratio (PC1-5), which collectively accounted for 84.5% of variance in the data. These PCs were subsequently used to compute UMAPs representing the genomic bins, as well as for K-means clustering of all bins. For the K-means clustering, a number of 6 clusters was

chosen. Decreasing the number of clusters resulted in the merging of distinct clusters, while increasing the number of clusters resulted in two or more clusters with very similar behaviors.

Definition of H3K27me3 categories and intrinsic NL affinity categories
For the analyses presented in Figure 5 and Figure S7, different categories of H3K27me3 enrichment and NL affinity were defined. For H3K27me3 levels, maternal and paternal genomes were considered separately. Genomic bins (100 kb) were divided into H3K27me3 low (RPKM < 0.2), mid (0.2 ≤ RPKM < 0.8), and high (RPKM ≥ 0.8) (see Fig. S7c). Intrinsic NL affinity was defined based on A/T content (fraction of A/T bases in 100-kb bins) and thresholds were chosen by comparison to zygote CF values (Fig. S7a): low (A/T content < 0.56), mid (0.56 ≤ A/T content < 0.61), and high (A/T content ≥ 0.61) (Fig. S7b).

Single-cell transcription analyses

Single-cell CEL-seq2 UMAP
To generate the transcriptional UMAP (Fig. 1c), single-cell transcript tables were processed in R (v. 4.1.2) using Seurat (v. 4.1.0)[49]. Only samples passing transcription and DamID thresholds were included; only scDam&T-seq LMNB1 samples of embryos from homozygous crosses and the mESC samples were used. Genes with counts observed in fewer than 10 cells were excluded and data was normalized using the "NormalizeData" and "ScaleData" commands. The UMAP was then generated using the "FindVariableFeatures", "RunPCA" and "RunUMAP" commands.

Cbx7-Lap2β versus Lap2β differential expression
Differential expression between Lap2β and Cbx7-Lap2β conditions was tested for transcriptional data derived from the Dam-LMNB1 scDam&T-seq experiment in hybrid 2-cell embryos, and the whole-embryo CEL-seq2 data from embryos collected at the end of the viability experiment. Data was processed using Seurat's "NormalizeData" and "ScaleData". Differentially expressed genes between were identified using the "FindMarkers" command. Only genes with an adjusted p-value < 0.05 were considered as differentially expressed. No DE genes were detected.

Transcription of genes in cells with versus without lamina association
To determine the effect of lamina association on the expression of a gene, we determined for each gene the group of cells in which the 100kb genomic bin containing the gene TSS was in contact with the lamina ("contact") and the group of cells in which it was not ("free"). This was done separately per embryonic stage. The transcript counts of the gene and the total transcript counts were then combined for the two groups, and the expression value (as log(RPM + 10)) was determined for each group. Genes were excluded from the analysis if either the contact or free group contained fewer than 10 cells; if the gene was expressed in fewer than 10 cells across both groups; if the gene was located on chrX or chrY; or if the gene was annotated as a maternal mRNA transcript by Park et al. (2013)[50]. In the case of allele-specific data, genes were also excluded if their TSS fell within a genomic bin that did not show sufficient allele separation (see above, *Filtering of DamID data*). The correlation in gene expression values between contact and free states was computed using Spearman's correlation.

Data availability

All genomic and transcriptomic data generated in this study has been deposited at the Gene Expression Omnibus under accession number GSE218598.

Code Availability

All data analysis code is available upon request.

References

1 Borsos, M. *et al.* Genome-lamina interactions are established de novo in the early mouse embryo. *Nature* **569**, 729-733 (2019).

2 Burton, A. & Torres-Padilla, M. E. Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. *Nat Rev Mol Cell Biol* **15**, 723-734 (2014).

3 Guerreiro, I. & Kind, J. Spatial chromatin organization and gene regulation at the nuclear lamina. *Curr Opin Genet Dev* **55**, 19-25 (2019).

4 Kind, J. & van Steensel, B. Genome-nuclear lamina interactions and gene regulation. *Curr Opin Cell Biol* **22**, 320-325 (2010).

5 van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780-791 (2017).

6 Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat Protoc* **2**, 1467-1478 (2007).

7 Chen, Z., Djekidel, M. N. & Zhang, Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. *Nat Genet* **53**, 551-563 (2021).

8 Dahl, J. A. *et al.* Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548-552 (2016).

9 Liu, X. *et al.* Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* **537**, 558-562 (2016).

10 Mei, H. *et al.* H2AK119ub1 guides maternal inheritance and zygotic deposition of H3K27me3 in mouse embryos. *Nat Genet* **53**, 539-550 (2021).

11 Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620-631 (2018).

12 Zhang, B. *et al.* Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**, 553-557 (2016).

13 Zheng, H. *et al.* Resetting Epigenetic Memory by Reprogramming of Histone Modifications in Mammals. *Mol Cell* **63**, 1066-1079 (2016).

14 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).

15 Biase, F. H., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res* **24**, 1787-1796 (2014).

16 Shi, J. *et al.* Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell RNA-seq. *Development* **142**, 3468-3477 (2015).

17 Torres-Padilla, M. E., Parfitt, D. E., Kouzarides, T. & Zernicka-Goetz, M. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445**, 214-218 (2007).

18 Rooijers, K. *et al.* Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. *bioRxiv*, 529388 (2019).

19 Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134-147 (2015).

20 Borsos, M. & Torres-Padilla, M. E. Building up the nucleus: nuclear organization in the establishment of totipotency and pluripotency during mammalian development. *Genes Dev* **30**, 611-621 (2016).

21 Puschendorf, M. *et al.* PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. *Nat Genet* **40**, 411-420 (2008).

22 Rang, F. J. *et al.* Single-cell profiling of transcriptome and histone modifications with EpiDamID. *Molecular Cell* **82**, 1956-1970.e1914 (2022).

23 Wang, M., Chen, Z. & Zhang, Y. CBP/p300 and HDAC activities regulate H3K27 acetylation dynamics and zygotic genome activation in mouse preimplantation embryos. *Embo j* **41**, e112012 (2022).

24 Du, Z. *et al.* Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235 (2017).

25 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).

26 Du, Z. *et al.* Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Mol Cell* **77**, 825-839 e827 (2020).

27 Rang, F. J. *et al.* Single-cell profiling of transcriptome and histone modifications with EpiDamID. *Mol Cell* **82**, 1956-1970.e1914 (2022).

28 Luderus, M. E., den Blaauwen, J. L., de Smit, O. J., Compton, D. A. & van Driel, R. Binding of matrix attachment regions to lamin polymers involves single-stranded regions and the minor groove. *Mol Cell Biol* **14**, 6297-6305 (1994).

29 Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* **23**, 270-280 (2013).

30 Tardat, M. *et al.* Cbx2 targets PRC1 to constitutive heterochromatin in mouse zygotes in a parent-of-origin-dependent manner. *Mol Cell* **58**, 157-171 (2015).

31 Siegenfeld, A. P. *et al.* Polycomb-lamina antagonism partitions heterochromatin at the nuclear periphery. *Nat Commun* **13**, 4199 (2022).

32 Inoue, A., Jiang, L., Lu, F., Suzuki, T. & Zhang, Y. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature* **547**, 419-424 (2017).

33 Briand, N. & Collas, P. Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol* **21**, 85 (2020).

34 Leemans, C. *et al.* Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**, 852-864.e814 (2019).

35 Markodimitraki, C. M. *et al.* Simultaneous quantification of protein-DNA interactions and transcriptomes in single cells with sc-Dam&T-seq. *Nat Protoc* **15**, 1922-1953 (2020).

36 Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 77 (2016).

37 van Schaik, T., Manzo, S. G. & van Steensel, B. Genome-Wide Mapping and Microscopy Visualization of Protein-DNA Interactions by pA-DamID. *Methods Mol Biol* **2458**, 215-229 (2022).

38 Kubota, T., Nishimura, K., Kanemaki, M. T. & Donaldson, A. D. The Elg1 replication factor C-like complex functions in PCNA unloading during DNA replication. *Molecular cell* **50**, 273-280 (2013).

39 Nora, E. P. *et al.* Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930-944. e922 (2017).

40 Lochs, S. J. & Kind, J. in *Spatial Genome Organization* 215-241 (Springer, 2022).

41 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

42 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019).

43 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

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

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

46 Abdennur, N. *et al.* Cooltools: enabling high-resolution Hi-C analysis in Python. *bioRxiv*, 2022.2010.2031.514564 (2022).

47 Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat Commun* **5**, 4114 (2014).

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

49 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).

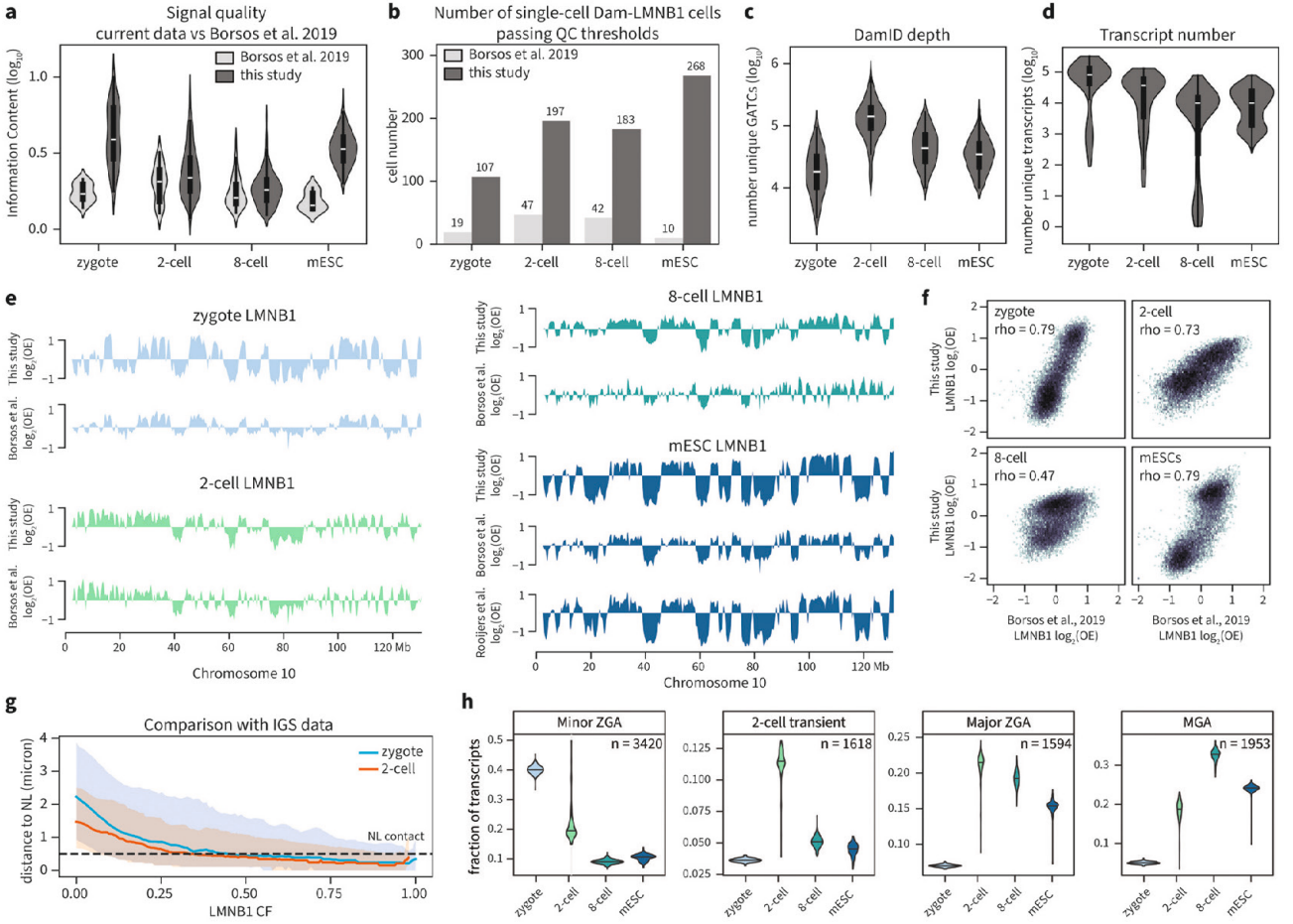
50 Park, S. J. *et al.* Inferring the choreography of parental genomes during fertilization from ultralarge-scale whole-transcriptome analysis. *Genes Dev* **27**, 2736-2748 (2013).

51 Rooijers, K. *et al.* Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. *Nat Biotechnol* **37**, 766-772 (2019).

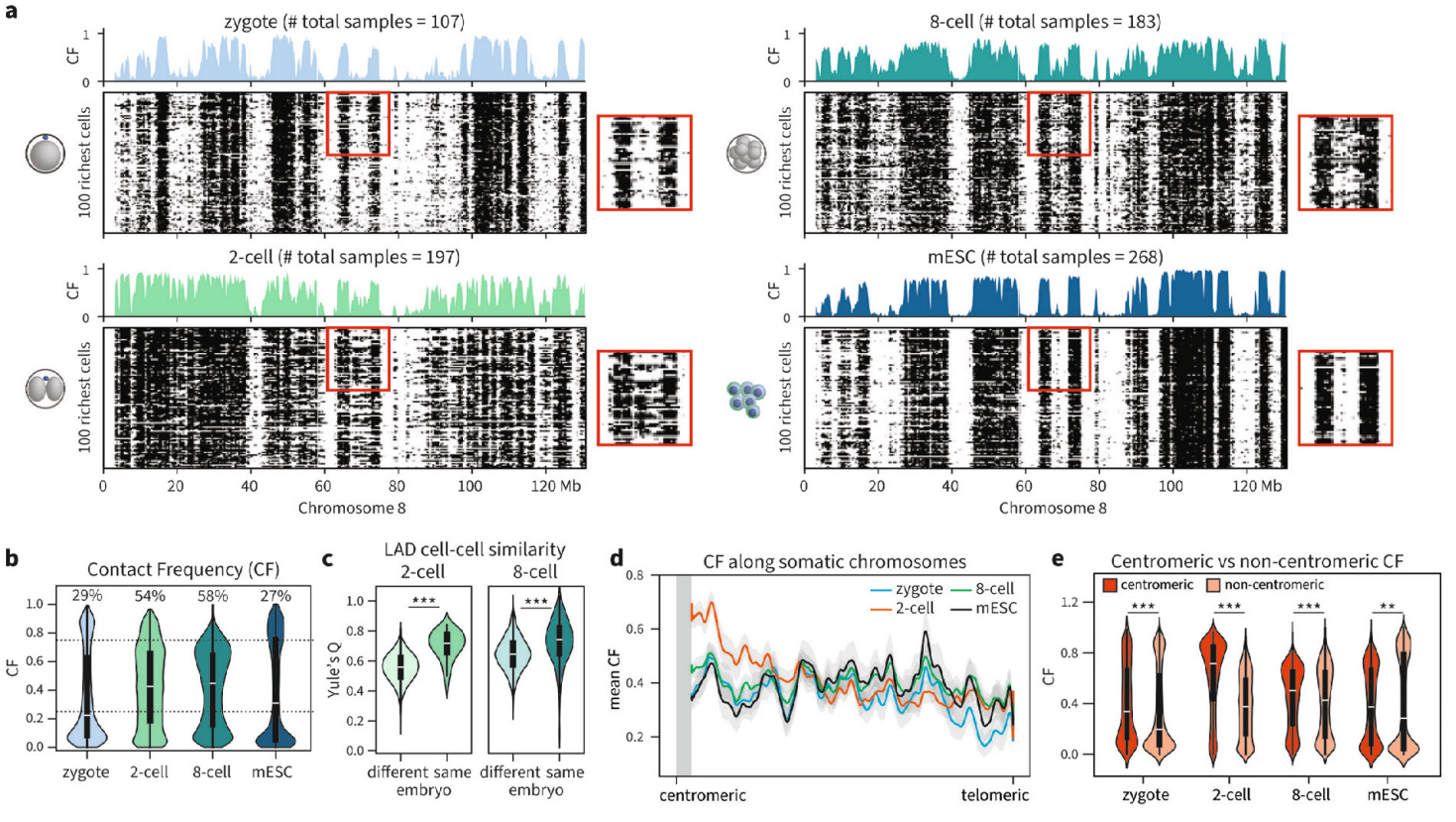
52 Payne, A. C. *et al.* In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371** (2021).

53 Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744-751 (2020).

Supplementary Figures


a
Signal quality
current data vs Borsos et al. 2019
Borsos et al. 2019
this study
Information Content (log10)
zygote
2-cell
8-cell
mESC
b
Number of single-cell Dam-LMNB1 cells
passing QC thresholds
Borsos et al. 2019
this study
cell number
19
107
47
197
42
183
10
268
c
DamID depth
number unique GATCs (log10)
d
Transcript number
number unique transcripts (log10)
e
zygote LMNB1
2-cell LMNB1
8-cell LMNB1
mESC LMNB1
This study log2(OE)
Borsos et al. log2(OE)
Rooijers et al. log2(OE)
Chromosome 10
f
zygote
rho = 0.79
2-cell
rho = 0.73
8-cell
rho = 0.47
mESCs
rho = 0.79
This study LMNB1 log2(OE)
Borsos et al., 2019 LMNB1 log2(OE)
g
Comparison with IGS data
zygote
2-cell
NL contact
distance to NL (micron)
LMNB1 CF
h
Minor ZGA
n = 3420
2-cell transient
n = 1618
Major ZGA
n = 1594
MGA
n = 1953
fraction of transcripts

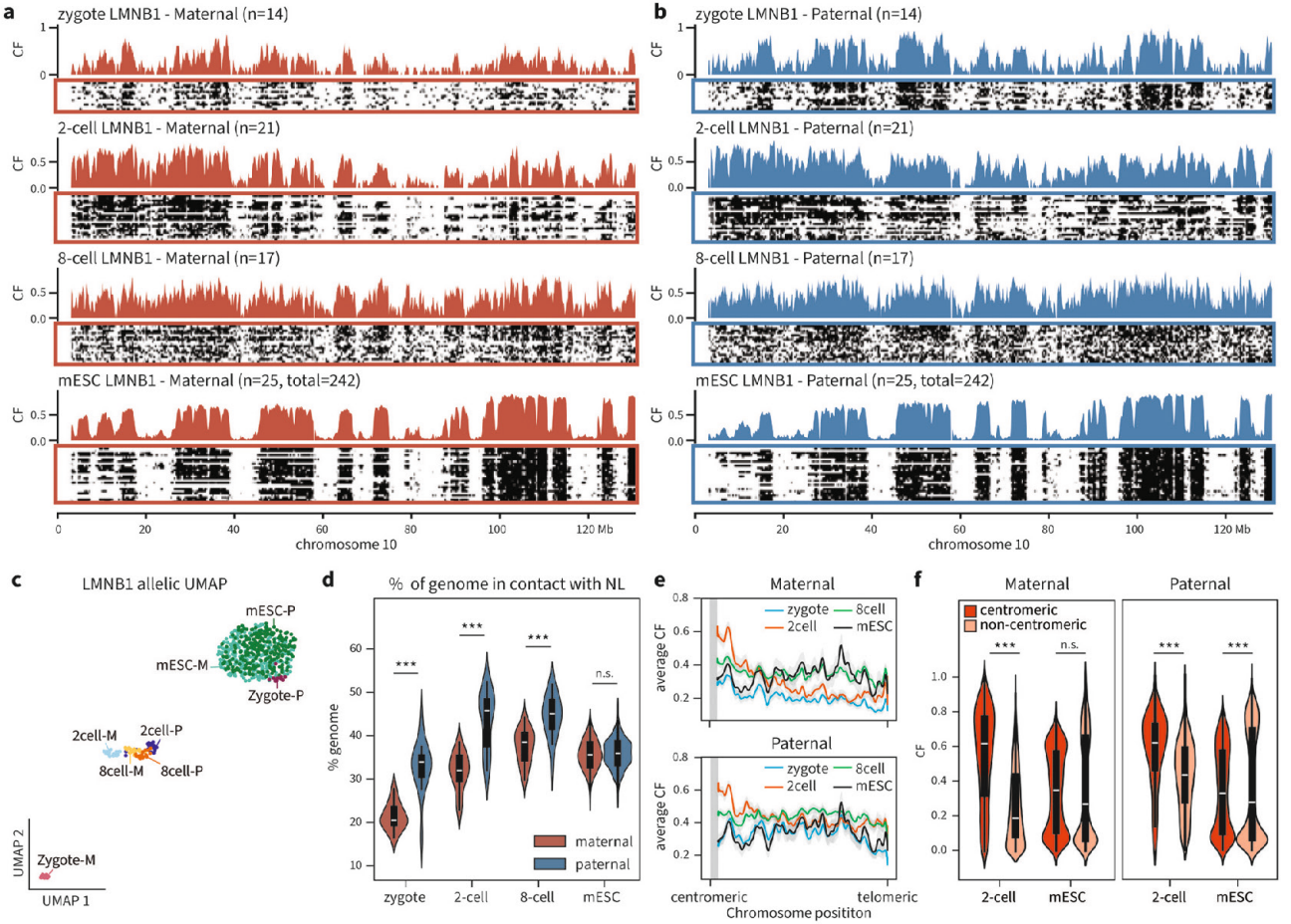
Supplementary Figure 1: Validation of single-cell LAD data during preimplantation stages
a, Information content, a measure of signal quality, plotted for samples from the present study and Borsos et al. (2019)[1]. **b**, Comparison of cell numbers that pass quality thresholds for the present study and Borsos et al. (2019)[1]. **c**, Violin plot depicting the distribution of the number of unique GATCs per stage. **d**, Violin plot depicting the distribution of the number of unique transcripts per stage. **e**, Combined single-cell profiles normalized to mappability (log_2OE) of this work compared to previous studies[1,51] over the entire chromosome 10. **f**, Comparison of single-cell averages from our study and Borsos et al., (2019)[1] for each stage using mappability normalized values (log_2(OE)) in 100-kb bins. Correlations were computed using Spearman's rank-order correlation (all conditions, $p < 1e\text{-}100$). **g**, Comparison of zygote and 2-cell LMNB1 CF values and the distance of genomic regions to the NL as reported in Payne et al. (2020)[52] by in situ genome sequencing (IGS), which generates a combined sequencing and imaging readout. The line shows the median distance of all fragments with a certain CF value, the shaded area shows the inter-quartile range. The dashed line indicates the distance threshold used to designate a genomic region as contacting the NL by Payne et al.[52] **h**, Distribution over all cells per stage showing the fraction of total transcripts that correspond to different gene categories, as defined by their expression dynamics during early development in Park et al. (2013)[50]. ZGA, zygotic genome activation; MGA, mid-preimplantation gene activation. Black line indicates median values. Boxplots included in (a), (c), and (d) indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers)



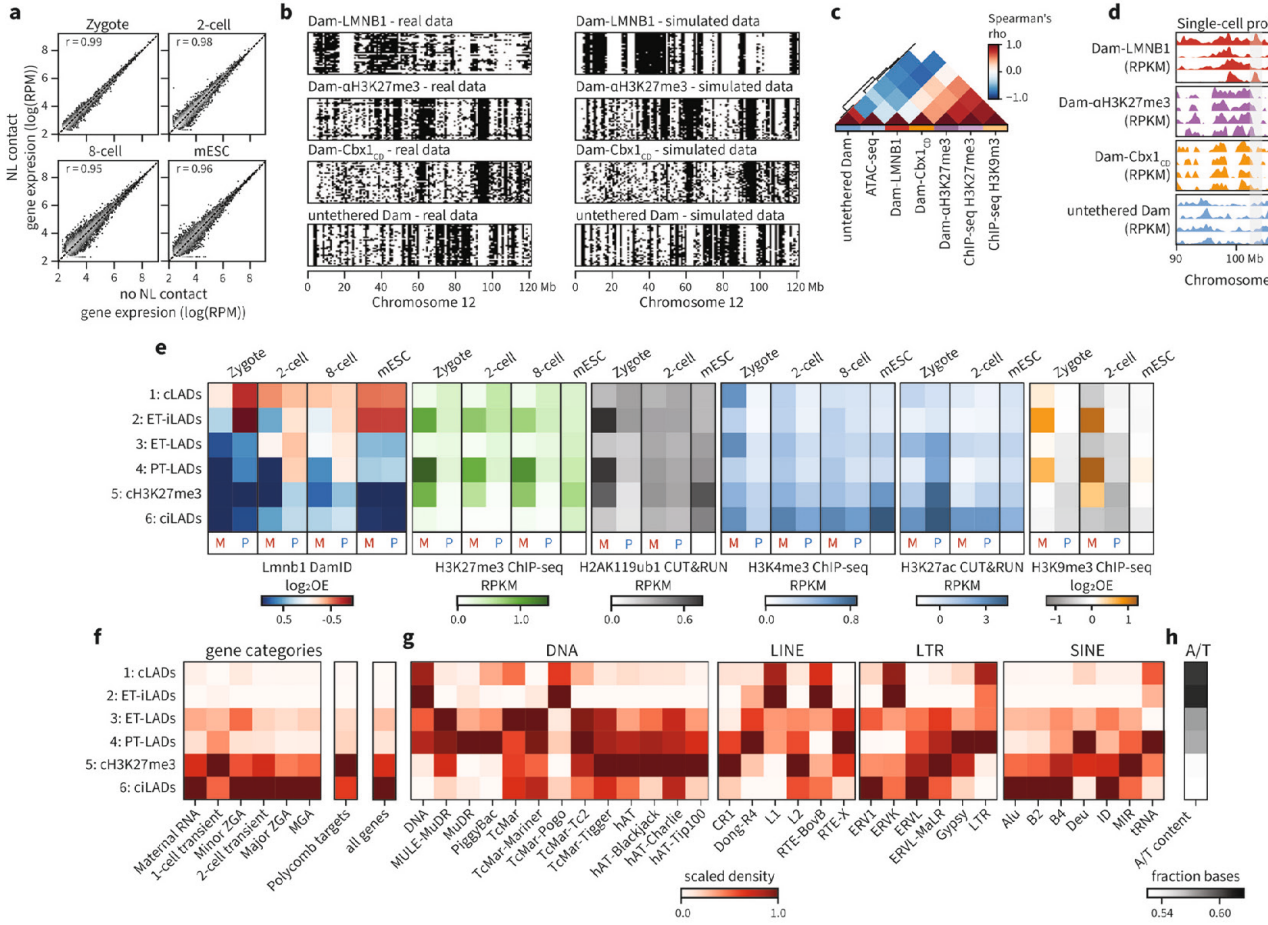
Supplementary Figure 2: Analysis of cell-to-cell LAD variability in preimplantation stages and mESCs

a, Heatmaps show single-cell binarized profiles of 100 example cells that passed quality thresholds along the entire chromosome 8 (left panel) ordered by unique number of unique GATCs (DamID depth) for each stage. Tracks on top of the heatmaps show contact frequency (CF) profiles. On the right side of each heatmap is a

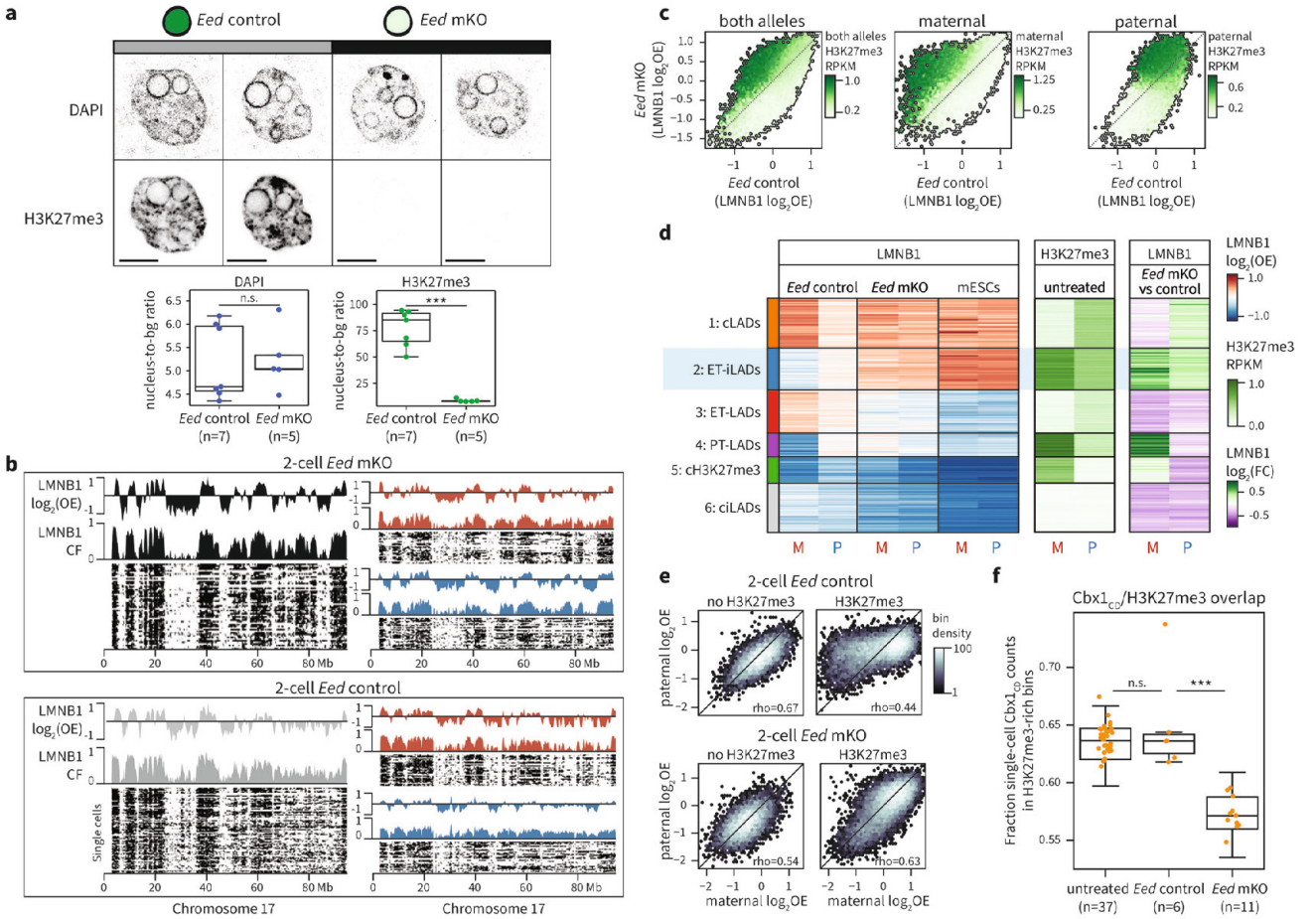
magnification of the region highlighted with a red rectangle. **b**, Distribution of CF values in 100-kb bins per stage. On top, percentages of bins between 0.25 and 0.75 CF values, which are indicative of cell-to-cell variability, are shown. **c**, Violin plot showing cell-to-cell similarity using Yule's Q on cell pairs originating from the same embryo or from a different embryo at the 2-cell stage (left) and at the eight-cell stage (right). Two-sided Mann-Whitney U test was performed (2-cell different embryo, n = 16,944; 2-cell same embryo, n = 76, p = 8.0e-28; 8-cell different embryo, n = 4,744, 8-cell same embryo: n = 206, p = 1.2e-24). **d**, Smoothed mean (1000-Mb Gaussian kernel) of LMNB1 CF values along the length of all autosomal chromosomes scaled to the same size per stage. **e**, Violin plot depicting CF values in the first 30 Mb (n = 4,984 100-kb bins) versus the remainder of the chromosome (n = 18,767 100-kb bins) for the zygote (p = 9.3e-57), 2-cell (p < 1e-100), 8-cell (p = 6.2e-22) stages and mESCs (p = 1.3e-3). Statistical significance was tested using a two-sided Mann-Whitney U test was performed. Boxplots included in (b), (c), and (e) indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).


a
zygote LMNB1 - Maternal (n=14)
2-cell LMNB1 - Maternal (n=21)
8-cell LMNB1 - Maternal (n=17)
mESC LMNB1 - Maternal (n=25, total=242)
CF
chromosome 10
b
zygote LMNB1 - Paternal (n=14)
2-cell LMNB1 - Paternal (n=21)
8-cell LMNB1 - Paternal (n=17)
mESC LMNB1 - Paternal (n=25, total=242)
c
LMNB1 allelic UMAP
mESC-P
mESC-M
Zygote-P
2cell-M
2cell-P
8cell-M
8cell-P
Zygote-M
UMAP 2
UMAP 1
d
% of genome in contact with NL
% genome
maternal
paternal
zygote
2-cell
8-cell
mESC
n.s.
e
Maternal
Paternal
average CF
zygote
2cell
8cell
mESC
centromeric
telomeric
Chromosome posititon
f
Maternal
Paternal
centromeric
non-centromeric
CF
2-cell
mESC

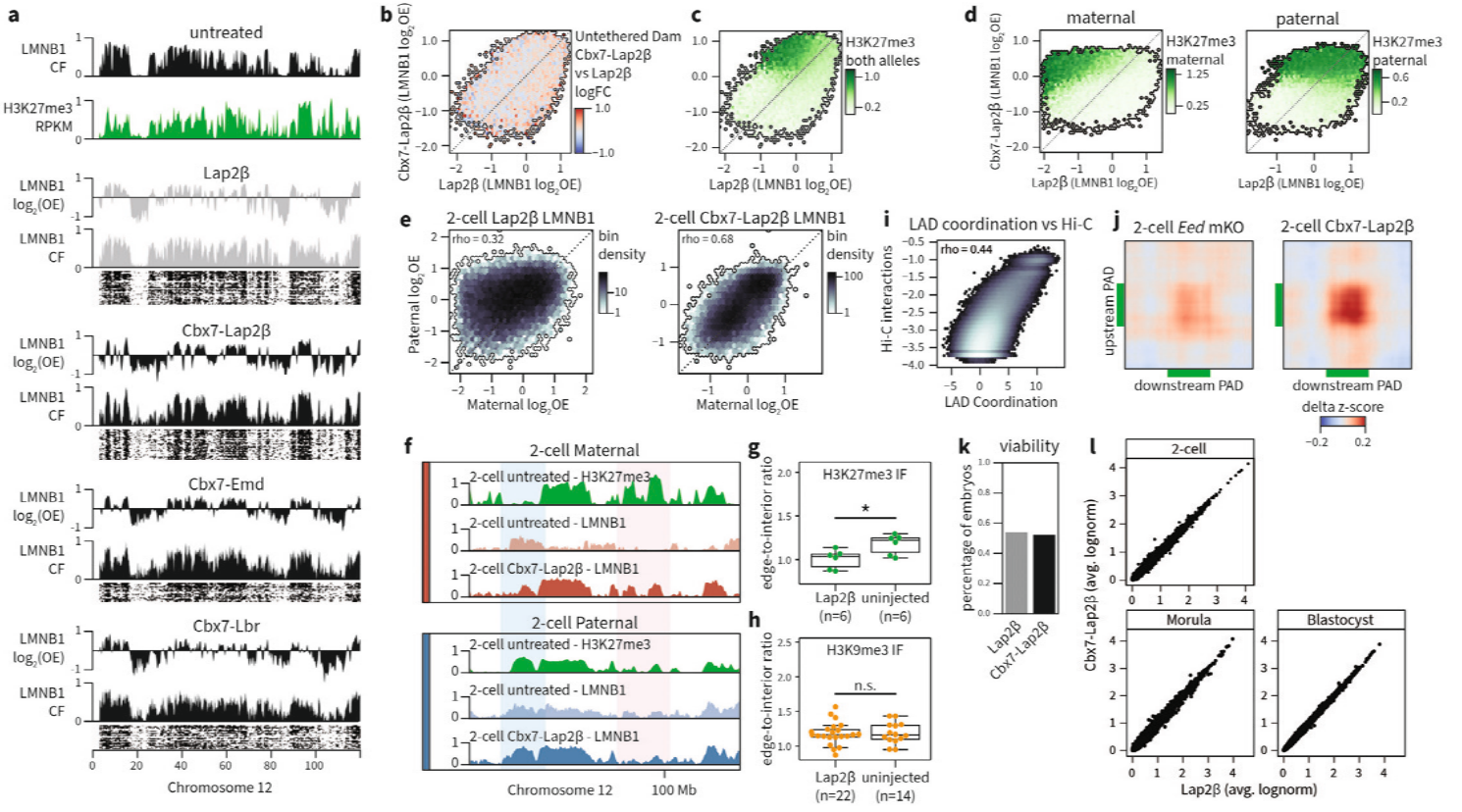
Supplementary Figure 3: Characterization of single-cell LAD profiles split by parental allele
a-b, Contact Frequency (CF) tracks and heatmap of binarized single-cell LAD profiles separated into maternal (**a**) and paternal (**b**) allele for the zygote, 2-cell, 8-cell and mESCs along the entire chromosome 10. Heatmaps are ordered by unique number of GATCs (DamID depth). A subset of the total number of mESC is shown. **c**, UMAP based on allelic Dam-LMNB1 single-cell readout. **d**, Percentage of the genome that is in contact with the NL per allele and per stage. Two-sided Wilcoxon rank-sum test was performed (zygote, n = 14, p = 9.4e-5; 2-cell, n = 26, p = 1.83e-7; 8-cell, n = 21, p = 2.5e-5; mESC, n = 268, p-value=0.36). **e**, Smoothed mean (1000-Mb Gaussian kernel) of LMNB1 CF values along the length of all autosomal chromosomes scaled to the same size per stage and split by allele (maternal - top and paternal – bottom). **f**, Violin plot depicting CF values separated by allele in the first 30 Mb (n = 4865 100-kb bins) versus the remaining of the chromosome (n = 18,330 100-kb bins) for 2-cell embryos and mESCs. Two-sided Mann-Whitney U test was performed (2-cell maternal, p < 1e-100; 2-cell paternal, p < 1e-100; mESC maternal, p = 0.60; mESC paternal, p = 2.3e-4). Boxplots included in (d) and (f) indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).


a
Zygote
2-cell
8-cell
mESC
r = 0.99
r = 0.98
r = 0.95
r = 0.96
NL contact gene expresion (log(RPM))
no NL contact gene expresion (log(RPM))
b
Dam-LMNB1 - real data
Dam-αH3K27me3 - real data
Dam-Cbx1CD - real data
untethered Dam - real data
Dam-LMNB1 - simulated data
Dam-αH3K27me3 - simulated data
Dam-Cbx1CD - simulated data
untethered Dam - simulated data
Chromosome 12
c
Spearman's rho
untethered Dam
ATAC-seq
Dam-LMNB1
Dam-Cbx1CD
Dam-αH3K27me3
ChIP-seq H3K27me3
ChIP-seq H3K9m3
d
Single-cell profiles
Dam-LMNB1 (RPKM)
Dam-αH3K27me3 (RPKM)
Dam-Cbx1CD (RPKM)
untethered Dam (RPKM)
Chromosome 8
e
1: cLADs
2: ET-iLADs
3: ET-LADs
4: PT-LADs
5: cH3K27me3
6: ciLADs
Lmnb1 DamID log2OE
H3K27me3 ChIP-seq RPKM
H2AK119ub1 CUT&RUN RPKM
H3K4me3 ChIP-seq RPKM
H3K27ac CUT&RUN RPKM
H3K9me3 ChIP-seq log2OE
f
gene categories
Maternal RNA
1-cell transient
Minor ZGA
2-cell transient
Major ZGA
MGA
Polycomb targets
all genes
g
DNA
LINE
LTR
SINE
h
A/T
A/T content
scaled density
fraction bases

Supplementary Figure 4: Validation and analysis of single-cell epigenetic marks at the 2-cell stage
a, Gene expression comparison between cells in which the gene contacts the NL and cells for which the same genne does not contact the NL. Correlation was computed using Pearson's correlation (all stages, $p < 1e\text{-}100$). Dashed lines show diagonal. **b**, Left: Heatmap of binarized single-cell profiles for Dam-LMNB1, Dam-aH3K27me3, Dam-$Cbx1_{CD}$ across the entire chromosome 12 ordered by decreasing unique number of GATCs. 30 cells are plotted per construct. Right: Heatmaps showing simulated single-cell samples. This data is used to correct cell-to-cell similarity calculations for construct-specific noise and sparsity levels. **c**, Correlation heatmap relating DamID (present study) and corresponding publicly available ChIP-seq measurements[13] at the 2-cell stage. **d**, Example tracks of four example single cells per construct over a selected region in chromosome 8. A region with high LMNB1 variability is highlighted in grey. **e**, Heatmap with LMNB1 $\log_2$(OE), H3K27me3 RPKM[13] H2Aub RPKM[10], H3K4me3 RPKM[12], H3K27ac RPKM[23] and H3K9me3 $\log_2$(OE)[11], average values per genomic cluster defined in Figure 2e. **f**, Heatmap showing scaled gene density of different categories described in Park et al (2013)[50] , Polycomb targets[53] and coding genes for each cluster. Scaling is done per column to accommodate the vastly different numbers per column. **g**, Heatmap showing scaled repeat density of each genomic cluster split by repeat family and (**h**) A/T content as fraction of bases per bin. Scaling is done per column to accommodate the vastly different numbers per column


a
Eed control
Eed mKO
DAPI
H3K27me3
DAPI
H3K27me3
n.s.
***
nucleus-to-bg ratio
Eed control (n=7)
Eed mKO (n=5)
b
2-cell Eed mKO
2-cell Eed control
LMNB1 log2(OE)
LMNB1 CF
Single cells
Chromosome 17
c
both alleles
maternal
paternal
Eed mKO (LMNB1 log2OE)
Eed control (LMNB1 log2OE)
both alleles H3K27me3 RPKM
maternal H3K27me3 RPKM
paternal H3K27me3 RPKM
d
LMNB1
Eed control
Eed mKO
mESCs
H3K27me3
untreated
LMNB1
Eed mKO vs control
1: cLADs
2: ET-iLADs
3: ET-LADs
4: PT-LADs
5: cH3K27me3
6: ciLADs
LMNB1 log2(OE)
H3K27me3 RPKM
LMNB1 log2(FC)
e
2-cell Eed control
2-cell Eed mKO
no H3K27me3
H3K27me3
paternal log2OE
maternal log2OE
bin density
rho=0.67
rho=0.44
rho=0.54
rho=0.63
f
Cbx1CD/H3K27me3 overlap
Fraction single-cell Cbx1CD counts in H3K27me3-rich bins
n.s.
***
untreated (n=37)
Eed control (n=6)
Eed mKO (n=11)

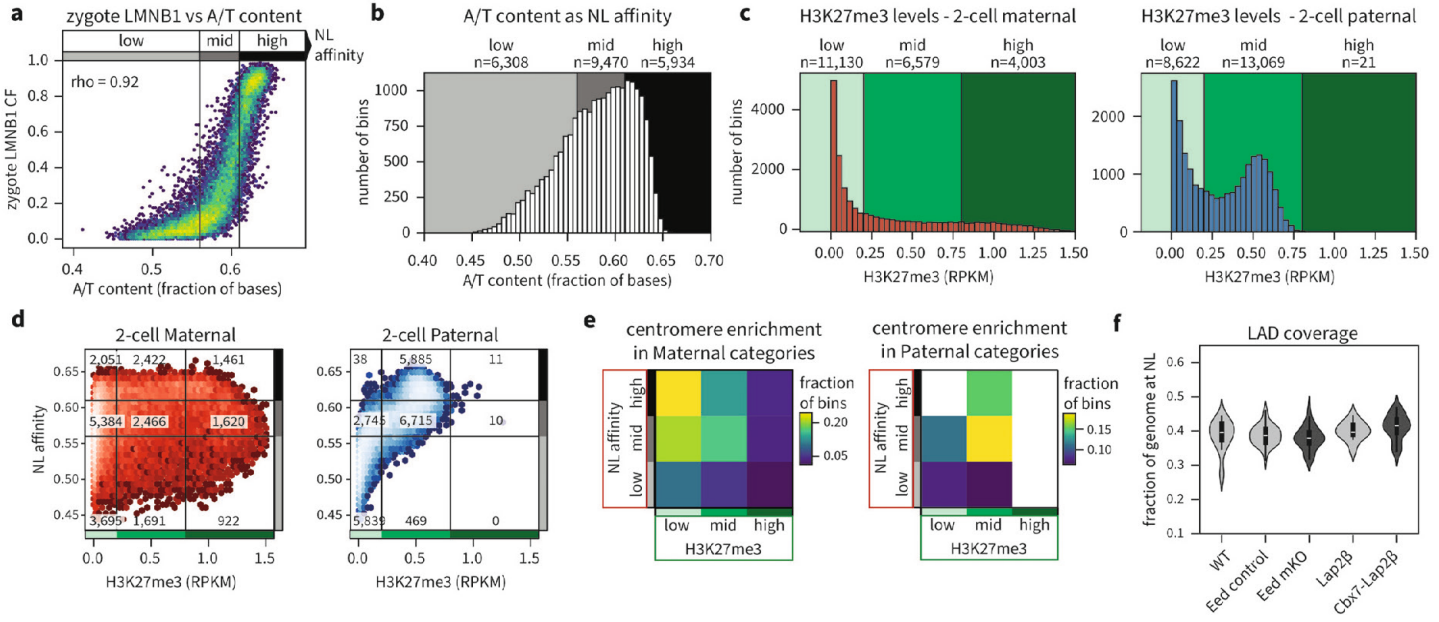
Supplementary Figure 5: Effect of *Eed* mKO on nuclear lamina association at the 2-cell stage
a, Top: DAPI and immunostaining of H3K27me3 in 2-cell *Eed* mKO or *Eed* control embryos (scale bar = 10 mm). Bottom: Quantification of the nucleus-to-background ratio. Two-sided Welch's T-test was performed to test significance (DAPI, p = 0.89; H3K27me3, p = 4.0e-5). **b**, Single-cell heatmaps of binarized LMNB1 profiles of 2-cell *Eed* mKO (top) and control (bottom) embryos with corresponding CF and $\log_2$(OE) values per condition along chromosome 17 of both alleles (left) or separated alleles (right). **c**, Comparison of LMNB1 values in 100-kb bins between *Eed* mKO and control conditions for both alleles (left), or maternal and paternal alleles separately (right). The color refers to average combined or allele-specific H3K27me3 values of 2-cell WT embryos[13]. **d**, Heatmap showing *Eed* mKO and control LMNB1 $\log_2$(OE) values per genomic bin at the 2-cell stage, as well as mESC. 2-cell H3K27me3[13] and differential LAD values between the two conditions for each of the genomic clusters depicted in Figure 2e. **e**, Comparison of maternal and paternal LMNB1 $\log_2$(OE) in 100-kb genomic bins containing H3K27me3 (right) or not (left) in either the control 2-cell condition (top) or in the *Eed* mKO (bottom). Color scale refers to density of genomic bins. **f**, Boxplot showing fraction of Dam-Cbx1$_{CD}$ counts overlap with H3K27me3-rich (RPKM > 0.2) bins in untreated (n = 37), *Eed* control (n = 6) and *Eed* mKO (n = 11) conditions. Two-sided Mann-Whitney U test was performed to test significance (*Eed* control vs untreated, p = 0.88; *Eed* control vs *Eed* mKO, p = 1.6e-4). Boxplots indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).



Supplementary Figure 6: Effect of H3K27me3-tethering to the NL nuclear lamina

a, Single-cell heatmaps of binarized LMNB1 profiles of 2-cell embryos injected with *Lap2b*, *Cbx7-Lap2b*, *Cbx7-Emd*, *Cbx7-Lbr* with corresponding CF and $\log_2$(OE) values

along chromosome 12. On top 2-cell LMNB1 CF values and H3K27me3[13] of untreated embryos are shown for reference. **b**, Comparison of LMNB1 values in 100-kb bins for the Cbx7-Lap2b and the Lap2b conditions. The color indicates the average $\log_2$(FC) of untethered between Cbx7-Lap2b and the Lap2b embryos. **c**, Comparison of LMNB1 values in the *Cbx7-Lap2b* and *Lap2b*-injected hybrid embryos for both alleles or maternal and paternal alleles separately (**d**). The color scale indicates the average allele-specific H3K27me3 values of 2-cell WT embryos[13]. **e**, Comparison of paternal and maternal LMNB1 CF in Lap2b (left, $p < 1e\text{-}100$) and Cbx7-Lap2b conditions (right, $p < 1e\text{-}100$). **f**, Example genomic region with maternal (top) and paternal (bottom) LMNB1 for Cbx7-Lap2b and Lap2b conditions and 2-cell WT H3K27me3. Dashed boxes highlight examples of allelic LAD asymmetry persisting in the Cbx7-Lap2b condition due to allele-specific H3K27me3 enrichment that can result in paternal-specific LADs (blue dashed box) or maternal-specific LADs (red dashed box). **g-h**, Quantification of the radial (1 mm) enrichment of H3K27me3 (**g**) and H3K9me3 (**h**) relative to the rest of the nucleus. Significance for the comparison between Lap2b and uninjected conditions was computed using Welch's two-sided t-test (H3K27me3, $p = 0.024$; H3K9me3, $p = 0.96$). **i**, Correlation between LAD coordination metric and Hi-C interaction values ($p < 1e\text{-}100$, see Methods). **j**, Average LAD coordination between pairs of PADs separated by up to 20 Mb. **k**, Fraction of embryos that reaches the blastocyst stage (E3.5) in Cbx7-Lap2b and Lap2b conditions (Cbx7-Lap2b $n = 67$, Lap2b $n = 65$). **l**, Gene expression comparison between Cbx7-Lap2b and Lap2b embryos at 2-cell, morula and blastocyst stages. Transcriptional data from 2-cell embryos was part of the Dam-LMNB1 scDam&T-seq experiment in hybrid crosses. Data from morula and blastocysts was obtained from whole-embryo CEL-seq2 at the end of the viability experiment. No significant differentially expressed genes were detected. Correlations in (**e**) and (**g**) were computed using Spearman's rank-order correlation.



Supplementary Figure 7: A model based on NL affinity and H3K27me3 to explain atypical LADs of the early embryo

a, Comparison between zygote LMNB1 CF and A/T content in 100-kb bins. Thresholds used for low, mid and high intrinsic NL affinity are depicted. Correlation is computed with Spearman's rank-sum correlation ($p < 1e\text{-}100$). **b**, Histogram showing the distribution of A/T content in 100-kb bins of the mouse genome. Low, mid and high intrinsic NL affinity categories are highlighted by using increasingly darker colors of the grey scale. **c**, Histogram showing 2-cell WT H3K27me3 RPKM distribution in the maternal (left) and paternal (right) alleles. Low, mid and high H3K27me3 categories are highlighted by using increasingly darker colors of green. **d**, Genome-wide comparison of NL affinity (A/T content) and maternal (red, left) or paternal (blue, right) H3K27me3 RPKM values in 100-kb bins. Categories defined by thresholds set in (b-c) are shown with the corresponding color. Number of genomic bins per category is displayed. **e**, Heatmap showing the fraction of bins in each category that is <20 Mb from the centromere. **f**, Distributions of the total fraction of the genome in association with the NL across cells per condition. The total fraction is computed as the average fraction of the maternal and paternal alleles. Boxplots indicate median values (white line), inter-quartile range (IQR, black box) and the range of all data points within 1.5 times the IQR (whiskers).

Supplementary Tables

Supplementary Table 1: **Cell numbers that pass thresholds (with transcriptomic read-out)**

Dataset	Dam fusion construct	stage	total cells	DamID PASS	CELseq PASS	DamID & CELseq PASS	DamID allelic PASS
homozygous WT	Dam-LMNB1	zygote	246	99	200	88	NA
homozygous WT	Dam-LMNB1	2-cell	210	96	152	69	NA
homozygous WT	Dam-LMNB1	8-cell	501	183	353	129	NA
homozygous WT	Dam-LMNB1	mESC	374	268	220	196	NA
hybrid WT	Dam-LMNB1	zygote	73	14	66	14	13
hybrid WT	Dam-LMNB1	2-cell	182	26	62	24	21
hybrid WT	Dam-LMNB1	8-cell	48	21	26	21	17
other constructs WT	Dam-Cbx1$_{CD}$	2-cell	121	37	99	34	NA
other constructs WT	Dam-aH3K27me3	2-cell	50	24	6	2	NA
other constructs WT	Untethered Dam	2-cell	168	46	166	46	NA
hybrid *Eed* control	Dam-Cbx1$_{CD}$	2-cell	118	6	90	6	0
hybrid *Eed* mKO	Dam-Cbx1$_{CD}$	2-cell	110	11	64	10	0
hybrid *Eed* control	Dam-LMNB1	2-cell	301	87	234	81	51
hybrid *Eed* mKO	Dam-LMNB1	2-cell	225	54	156	51	31
homozygous Cbx7-Emd	Dam-LMNB1	2-cell	42	17	40	15	NA
homozygous Cbx7-Lap2β	Dam-LMNB1	2-cell	118	28	118	28	NA
homozygous Cbx7-Lbr	Dam-LMNB1	2-cell	38	22	34	21	NA
homozygous Lap2β	Dam-LMNB1	2-cell	133	31	125	30	NA
homozygous Cbx7-Lap2β	Untethered Dam	2-cell	98	26	93	22	NA
homozygous Lap2β	Untethered Dam	2-cell	91	15	84	12	NA
hybrid Cbx7-Lap2β	Dam-LMNB1	2-cell	76	39	56	38	39
hybrid Lap2β	Dam-LMNB1	2-cell	38	11	15	11	10

Supplementary Table 2: **Cell numbers that pass thresholds (without transcriptomic read-out)**

Dataset	Dam fusion construct	stage	total cells	DamID PASS
homozygous WT	Dam-LMNB1	zygote	30	8
homozygous WT	Dam-LMNB1	2-cell	188	101
other constructs WT	Dam-aH3K27me3	2-cell	31	6
other constructs WT	Untethered Dam	2-cell	68	31

Supplementary Table 3: **Annotation of genomic clusters per bin**

A table containing genomic bins and their annotation according to the clustering analysis presented in Figure 2e. The table is too large to be presented here.

Supplementary Table 4: **External dataset accession numbers**

Accession	Technique
GEO: GSE71434	ChIP-seq - H3K4me3
GEO: GSE76687	ChIP-seq - H3K27me3
GEO: GSE82185	Hi-C
GEO: GSE97778	ChIP-seq - H3K9me3
GEO: GSE112551	DamID – LMNB1
GEO: GSE153496	CUT&RUN - H2AK119ub1
ENCODE: ENCSR857MYS	ChIP-seq – H3K9me3
GSE207222	ChIP-seq – H3K27ac

Supplementary Table 5: **RNA concentrations of constructs for zygote injections**

construct	dilution	stage collected	induction	collection (hours post hCG)
Dam-LMNB1	5ng/uL	zygote	no	29-31h
Dam-LMNB1	10ng/uL	2-cell	no	52-55h
Dam-LMNB1	100, 150 or 200ng/uL*	8-cell	no	75-78h
Dam-aH3K27me3	5 or 10ng/uL*	2-cell	no	52-55h
untethered Dam-ERT2	20ng/uL	2-cell	tamoxifen 20h	52-55h
Dam-Cbx1$_{CD}$-ERT2	20ng/uL	2-cell	tamoxifen 20h	52-55h
Cbx7-Lap2β	1000ng/uL	2-cell to blastocyst	no	52-55h
Lap2β	1000ng/uL	2-cell	no	52-55h
Cbx7-Lbr	1000ng/uL	2-cell	no	52-55h
Cbx7-Emd	1000ng/uL	2-cell	no	52-55h

*multiple concentrations were the result of optimizations: underlined is the most successful condition.

Supplementary Table 6: **Filtering conditions for single-cell DamID data**

Dam construct	Embryonic stage	Cross	# GATC threshold	IC threshold
Dam-LMNB1	Zygote	all	^{3}3,000	^{3}1.4
Dam-LMNB1	2-cell embryo	all	^{3}10,000	^{3}1.4
Dam-LMNB1	8-cell embryo	all	^{3}10,000	^{3}1.4
Dam-LMNB1	mESC	all	^{3}10,000	^{3}1.4
Untethered Dam	2-cell embryo	homozygous	^{3}5,000	^{3}1.2
Untethered Dam	2-cell embryo	C57BL/6J x JF1/MsJ	^{3}3,000	^{3}1.2
Dam-aH3K27me3	2-cell embryo	all	^{3}10,000	^{3}1.2
Dam-Cbx1$_{CD}$	2-cell embryo	homozygous	^{3}5,000	^{3}1.2
Dam-Cbx1$_{CD}$	2-cell embryo	C57BL/6J x JF1/MsJ	^{3}3,000	^{3}1.2

CHAPTER 7

Discussion

Franka J. Rang[1]

1: Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW), University Medical Center Utrecht, Oncode Institute

Single-cell omics: opportunities and challenges

Over the past 20 years, the field of single-cell sequencing has seen technological development at a breathtaking pace. The first description of a single-cell mRNA sequencing experiment in 2009[1] was quickly followed by a range of improved scRNA-seq protocols[2,3]. By the time I started my PhD research in 2017, labs across the world were performing these types of experiments. Around the same time, the first methods for single-cell epigenomics were being published that enabled the readout of different regulatory features, such as accessibility[4,5], DNA methylation[6,7] and protein-DNA interactions[8,9]. Over the course of my PhD, these technologies have evolved to provide multiple molecular readouts from the same cell, thus giving rise to single-cell multi-omics[10]. A notable example is scDam&T-seq, which I presented in **Chapter 2** of this thesis, a single-cell multi-omics technique that combines the readout of protein-DNA interactions provided by scDamID with the transcriptional readout of CEL-Seq2. Although scDam&T-seq was the first to combine these two readouts, multiple other protocols soon followed[11-14]. Now, by the end of my PhD, a new wave of technological advancements has enabled the readout of genome-wide binding profiles of multiple proteins from the same cell (known as a multifactorial readout)[15-21], even in conjunction with transcriptomics[22]. In addition, techniques recording spatial information have made it possible to relate the transcriptional and/or epigenetic state of a cell to its local environment[10,23]. Besides the increased amount of information extracted per cell, major improvements have been made in the throughput of protocols: Whereas the first single-cell experiments processed tens to hundreds of cells, the ranges now extend from thousands to even millions[10].

Clearly, we can expect technological advancement to continue in the next years, providing an increasing number of readouts for an increasing number of cells. This raises the question of what the ultimate goal of these single-cell protocols should be. Is it necessary, or even beneficial, to include more and more information from the same cell? I believe the answer to this question is yes. Cells are extremely complex systems with a large number of proteins, transcripts and genes interacting with each other in a way that is specific to the type, state and environment of the cell. Obtaining the combined readout of multiple of these factors across a large number of cells provides an unprecedented opportunity to explore these interactions and investigate how they impact cellular function. However, the generation and analysis of big single-cell datasets requires a lot of time and money. As the field matures and multi-modal techniques become more widely implemented, it will thus be important to carefully consider experimental design and set priorities. This should, for example, include the consideration of which cellular features have to be captured from the same cell, and which features could potentially be computationally integrated from existing datasets. While it may be interesting to jointly measure two closely related features in the same cell (e.g. chromatin accessibility and transcription), the global profile of one feature could likely be predicted fairly accurately from the other[24-26], which could be sufficient for many research questions. It may thus be more advantageous to capture a number of cellular features that are not strongly linked and use these as anchor points to computationally incorporate or predict additional features. In

addition, integration of published data could substantially limit the required number of cells that is necessary for the questions of interest. Taking these considerations into account during experimental design will result in more targeted and effective research.

As technological advancements result in more and bigger datasets, computational methods will have to keep up to make optimal use of the available information. Broadly, there are two distinct tasks for these methods: First, processing data to remove technical artefacts and integrate with additional datasets to increase the number of modalities and cells. Second, extracting biological insight. A whole range of tools tackling these challenges is currently available[27], with novel methods continuously being added. While the growing number of specialized single-cell tools is very promising, and even necessary, it complicates choosing the best tool for a given dataset. Moreover, the increasingly complex theoretical foundations of these tools will cause them to work as a black box for most users. Consequently, the assumptions and limitations of these methods will not be clear and can result in overinterpretation or accidental misuse of the processed data. For example, data imputation tools provide seemingly complete expression profiles for all single cells, but can lead to spurious correlations between genes that may be mistaken for actual biology[28-30]. Clear benchmarking, tutorials, and community guidelines (such as recently provided by Heumos et al.[31]) will thus be crucial to make the successful execution and interpretation of single-cell (multi-modal) data possible for all research labs.

The position of scDam&T-seq within the field of single-cell omics

In **Chapters 2-3**, we developed and presented scDam&T-seq to enable the single-cell readout of transcription in combination with a protein-DNA interaction or accessibility profile. In addition, we extended the use of scDam&T-seq to profile histone post-translational modifications (PTMs) with the development of EpiDamID in **Chapter 4**. At the time of development and publishing, these methods offered innovative possibilities in the field of single-cell multi-omics. However, by now, a range of single-cell multi-omic protocols exists as alternatives. So how does scDam&T-seq compare to these other methods?

Advantages and potential of scDam&T-seq

Compared to other techniques with similar readouts, scDam&T-seq offers a number of unique advantages, mainly driven by the unique approach of DamID compared to other genomic techniques. First, it is the only available tool for profiling protein-DNA interactions in single cells that does not rely on an antibody. This makes scDam&T-seq uniquely capable of studying proteins for which no antibodies are available, which is especially relevant for systems for which few reliable antibodies exist, such as the zebrafish. However, in the case of post-translationally modified proteins, DamID does require a targeting domain specific for the PTM (e.g. a single-chain antibody or protein domain) to direct Dam, i.e. EpiDamID. A second advantage of scDam&T-seq is that it can be used to profile single cells from very scarce material. As long as the Dam-fusion protein can be expressed in the system, a small number

of cells can be successfully processed. This in contrast with most other single-cell techniques, where antibody staining is typically performed in bulk to maximize efficiency. In **Chapter 6**, we make use of this property of scDam&T-seq to study lamina-associated domains (LADs) and various histone PTMs in preimplantation embryos.

Besides the advantages of the current implementation of scDam&T-seq, there are a number of interesting opportunities for future adaptation of the protocol that are related to the way protein-DNA interactions are recorded. Specifically, Dam lays down the methylation mark in vivo directly onto the DNA and this marks is stable, only being lost upon DNA replication. The signal remains intact throughout various processing steps such as cell lysis, fixation, or protein digestion. As such, the DamID steps of the scDam&T-seq protocol can potentially be added to a range of existing single-cell genomics techniques. For instance, our lab recently developed Dam&ChIC[21,32], a single-cell method that combines DamID with sortChIC[33] and can thus provide protein-DNA interaction profiles for two distinct proteins of interest (POIs) from the same cell. Furthermore, the implementation of Dam&ChIC capitalized on another powerful feature of the DamID signal: Since the methylation mark laid down by Dam accumulates over time in vivo, it offers a historical record of protein-DNA interactions over the course of one cell cycle. In contrast, the interactions recorded by the ChIC readout represent a snap-shot of the binding profile at the moment of cell harvest. The combination of the two readouts thus provides information on recent changes in protein binding[21]. The ability to record past chromatin states offers the exciting possibility to study chromatin dynamics, the order of epigenetic events, and their relation to subsequent cell state changes.

The fact that the methylation mark is stably laid down in vivo has the additional advantage that the chromatin in contact with the Dam-POI can be visualized with microscopy. The visualization of Dam-methylated DNA is achieved through the use of a tracer protein (^{m6}A-Tracer), which consists of the G^{m6}ATC-binding domain of the restriction enzyme DpnI fused to a fluorescent protein, such as GFP[34]. Recently, the imaging of Dam signal in single cells was combined with sequencing of the same cells in a method called mDamID[35]. Similarly, DamID and ^{m6}A-Tracer imaging could potentially be incorporated in single-cell spatial-omics methods[36], which would provide combined information on the transcription, chromatin organization, and cellular environment.

Finally, the methylation mark laid down by Dam can be directly recorded by the long-read sequencing technology developed by Oxford Nanopore[37]. Notably, this technique is able to sequence nascent DNA and record chemical modifications, thus eliminating the need to enrich the regions of interest via digestion and incorporation of adapters. This principle has been used successfully to map protein-DNA interactions that were recorded by targeting the nonspecific deoxyadenosine methyltransferase Hia5 to sites of antibody binding[38]. Long-read sequencing has the notable advantage that it is capable of charting repetitive regions of the genome, which are largely excluded when using short-read sequencing. As single-cell

protocols for Nanopore sequencing are emerging[39-41], this offers interesting prospects for the implementation of scDam&T-seq.

Limitations of scDam&T-seq

While scDam&T-seq has some notable advantages and potential for future development, it also has a number of clear limitations, which have partially been discussed in **Chapter 3**. First, the Dam-POI needs to be expressed in the system of interest, which for single-cell applications often requires the establishment of stable clones to ensure similar expression levels across cells. This can be a time-consuming step, especially for model organisms. In some cases, it is possible to circumvent this step, as we demonstrated in **Chapters 4** and **6**, where we achieved Dam-POI expression in early zebrafish and mouse embryos by directly injecting mRNA encoding the Dam-POI in the zygote. Second, the Dam enzyme is sensitive to accessibility, resulting in background signal in regions of open chromatin. In our experience, this phenomenon is notably stronger for Dam-POI fusions that freely diffuse throughout the nucleus compared to proteins with a very specific localization, such as components of the nuclear lamina (NL). Most experiments will thus require a negative control in which Dam is untethered to distinguish true signal from background. However, detecting true signal remains difficult for POI that extensively overlap chromatin accessibility, even when using this negative control. Third, Dam only methylates adenines in the GATC motif, which occur every ~263bp in the mouse genome, thus limiting the maximum resolution that can be achieved. In most single-cell applications, this is unlikely to be a limitation, since the achieved practical resolution is much lower due to data sparsity. Still, it may prevent the application of DamID for genomic regions devoid of GATC sequences, such as the centromeres. Fourth, the transcriptional readout of scDam&T-seq is obtained via amplification with primers annealing to the poly-A tail. As a consequence, only polyadenylated RNA can be captured. This excludes other types of transcripts, such as pre-mRNA, many repeat-derived transcripts, and other non-coding transcripts. Finally, the current scDam&T-seq implementation has a throughput of hundreds to thousands of cells. This is comparatively little compared to recent protocols that make use of combinatorial indexing or droplet-based processing to achieve throughputs with a magnitude of 10,000-100,000s of cells[10].

The future of scDam&T-seq

As the work in **Chapters 2**, **4** and **6** demonstrates, scDam&T-seq is a powerful multi-omics technique that can be successfully implemented in a variety of systems. However, the current protocol has a number of practical limitations that reduce its general applicability and may make it less attractive in situations where antibody-based methods are also available. To sustain a competitive position, scDam&T-seq will need to evolve alongside the technological advancements in the field. This includes addressing the current limitations and capitalizing on its unique properties.

Essentially all limitations outlined in the previous section can be addressed by adaptations of the experimental protocol. For example, the necessity of expressing Dam-POI in the system

could be circumvented by a single-cell implementation of pA-DamID[42], in which Dam is targeted to the POI via the antibody-binding properties of protein A. This approach could also mitigate the background signal in regions of open chromatin, as the active Dam enzyme does not freely diffuse throughout the nucleoplasm over a period of time. However, the use of pA-DamID does mitigate some of the advantages of scDam&T-seq mentioned previously, specifically its independence of antibody availability and the possibility to record historic binding events. Another strategy to limit the accessibility bias could be the use of Dam mutants with reduced affinity for DNA, thus increasing their reliance on the POI for DNA binding[43,44]. In **Chapter 4**, we found that the Dam N126A mutant markedly reduces off-target signal in population samples, but its use will have to be further optimized for single-cell implementations. Finally, the shortcomings of the current transcriptional readout and the throughput are not related to the fundamental principle of DamID and can thus be remedied by incorporation of existing methodologies.

In addition to experimental modifications, there is a role for computational strategies to mitigate some of the challenges of single-cell DamID data and maximize the information that can be gleaned from it. For example, the development or implementation of multi-omic data imputation strategies could help to limit the problem of sparsity[45-47]. In the case of scDam&T-seq experiments, there is an additional motivation to develop imputation tools that can share information between cells with signal from different Dam constructs, as they would enable a cell-specific prediction of the Dam background signal that can be subtracted from the Dam-POI signal. Alternatively, it would be possible to employ a statistical approach that uses data generated with untethered Dam to estimate the distribution of accessibility values per cell type (as determined by the transcriptome). These distributions could then be used to detect regions with significant Dam-POI binding, including a confidence interval. As mentioned at the beginning of this chapter, computational strategies keep evolving to enable the optimal use of experimental innovations. It is thus also important to keep up with these developments, ideally via the development of tools that make use of the specific characteristics of DamID data.

In conclusion, scDam&T-seq fulfills a unique position within the field of single-cell epigenomics due to its implementation of DamID. While the protocol currently enables high-quality research, it will eventually have to evolve to stay up-to-date with technological developments and new standards in the field of single-cell omics. The exact modifications of the protocol will depend on the experimental interest. I believe that the biggest promise lies in the integration of the DamID readout into multi-modal protocols, especially spatial-omics techniques, and its capacity to track historic chromatin states. Part of this potential has already been realized by the development of Dam&ChIC and ongoing research in our group and others indicates that more developments will soon follow. As such, scDam&T-seq may have been the first DamID-based single-cell multi-omic approach, but it certainly will not be the last.

Are stochastic NL contacts inhibitory for transcription?

As discussed in **Chapter 1**, the extent to which NL association inhibits gene expression and the mechanisms by which inhibition is achieved have not been completely elucidated. Studies integrating reporters at various loci in the genome have demonstrated that LADs indeed form a more repressive environment[48-50], but that repression also strongly depends on chromatin context and promoter sequence[51]. However, from these results it is not clear to which extent NL contact by itself directly impacts gene expression, or whether it mainly serves to reinforce the repressive state created by epigenetic modifications. The potential influence of NL on gene expression is especially interesting given the presence of stochastic LAD variability between cells of the same cell type[8,34]. If NL contact indeed directly inhibits transcription, such variability in LAD contacts could result in considerable transcriptional and thus cellular heterogeneity.

The development of scDam&T-seq has enabled the simultaneous readout of NL contacts and transcription. In **Chapter 2**, we applied scDam&T-seq for the NL component Lamin B1 in mouse embryonic stem cells (mESC) and indeed found that some genes tend to be more lowly expressed when they are in contact with the NL. Surprisingly, the genes that are most strongly affected by NL association have low contact frequencies (i.e. infrequently associate with the NL) and are thus more likely to be located in euchromatin. These results suggest that euchromatic genes are susceptible to repression at the NL, potentially through interactions with repressive heterochromatin proteins such as Hdac3[52]. In **Chapter 6**, we revisited the relationship between variability in NL association and transcription, this time in the context of preimplantation development. We observed that LADs are much more heterogenous in cleavage-stage embryos compared to mESC, but that this variability did not influence the expression of the underlying genes. Notably, we did not evaluate the relationship between NL contact and expression separately for genes with different contact frequencies , as the focus was on identifying a potential transcriptome-wide effect related to the high level of LAD variability. Stratifying genes on their contact frequencies did reveal a similar influence of NL association on low contact frequency genes, as previously seen for mESCs (data not shown). From these results, it seems that in both cases active regions of the genome experience transcriptional repression at the nuclear periphery, while heterochromatic regions may be less sensitive to stochastic changes in their radial positioning.

The lack of an effect for genes with mid to high contact frequencies could be explained by both biological and technical factors. In both mESC and early embryos, we found that such regions are enriched for heterochromatic histone marks, including H3K27me3 and H3K9me3. These chromatin states likely ensure robust repression and may thus be insensitive to occasional dissociation from the NL. A technical explanation could lie in the cumulative nature of DamID signal, which would obscure a loss of NL contact if previous association had already resulted in methylation of the DNA. This is more likely to happen for mid to high contact frequency regions, which spend a considerable time at the NL. In addition, the expression detected in

scDam&T-seq reflects the pool of mRNA in the cell, which also represents a historical view of transcriptional activity rather than ongoing transcription. Other factors diminishing the overall effect size for all genes include the limited resolution of 100-kb bins, noise in the single-cell data, and the presence of two (or more) alleles.

Recently, Su et al. developed DNA multiplexed error-robust FISH (DNA-MERFISH), a microscopy technique that can detect the nuclear position of >1,000 genomic loci and the expression of >1,000 genes in the same cells[53]. By spacing ~1,000 genomic probes homogenously across the genome, the researchers applied this technique to reconstruct single-cell chromosome conformation alongside information on nascent transcription of the ~1,100 genes at the imaged loci. Since both spatial and transcriptional information were available for these genes, the researchers were able to investigate the effect of peripheral positioning on transcription. They found that genes had a median 25% lower transcriptional firing rate when positioned at the NL compared to the nuclear interior, thus confirming that NL contact correlates with a repressive state. Whereas scDam&T-seq only detected an effect for low contact frequency genes, nearly all genes assayed in DNA-MERFISH had a reduced firing rate at the NL. The increased effect size observed in the microscopy data could be related to the higher temporal resolution (i.e. a snapshot view of both radial positioning and transcription) and the allele-specific detection. The researchers did not evaluate whether this effect was further influenced by chromatin state, so it cannot be excluded that the extent of silencing at the NL depends on other factors as well.

The coincidence of NL association and reduced transcription strongly suggests that genes are repressed upon contact with the NL. However, the results only provide evidence of correlation and not causation. A different explanation could be that genes are more likely to associate with the NL in moments when they are not being transcribed, for example due to the stochastic acquisition of heterochromatic marks or the absence of interactions with the transcriptional machinery. In support of a role for transcription in regulating NL association, Su et al. showed that a 6-hour treatment with the transcriptional inhibitor alpha-amanitin resulted in a median increase of 50% in NL association[53]. Moreover, targeted activation or repression of individual genes can also influence their radial positioning[54]. Alternatively, genes may occasionally become more peripherally located, which coincides with a reduced expression due to a lower concentration of transcriptional activators. These different scenarios are not mutually exclusive, but are difficult to disentangle especially in the context of stochastic and potentially more dynamic variability in NL contacts.

Does H3K27me3 inhibit NL association?

In **Chapter 6**, we demonstrated that H3K27me3 mediated by maternal Polycomb Repressive Complex 2 (matPRC2) strongly reduces the association of canonical LADs with the NL in mouse preimplantation embryos. These results are in line with recent work by Siegenfeld et al., who evaluated the relationship between H3K27me3 and NL association in a human cell

line via inhibition of Ezh2, the catalytic subunit of PRC2[55]. Inhibition of Ezh2 led to the loss of H3K27me3 and a concomitant gain in NL contacts. Earlier research, including the work presented in **Chapter 2**, had already demonstrated a genome-wide negative correlation between H3K27me3 and NL association in both human and mouse cell lines[8]. Together, these results demonstrate a direct inhibitory effect of PRC2/H3K27me3 on NL association.

While the results presented above seem clear, there are a number of studies that report evidence that suggests H3K27me3 promotes NL association. In the first characterization of human LADs, an enrichment of H3K27me3 on the inside border of LADs was observed[56]. The same phenomenon was later reported in mouse fibroblasts[57]. While the presence of H3K27me3 at the borders of LADs seems to imply a positive effect on NL association, this is not necessarily the case. Closer inspection of Lamin B1 DamID and H3K27me3 ChIP-seq profiles used in one of these studies[57] indeed shows a general overlap, but a local negative association (Fig. 1a). Speculating based on these preliminary observations, it is possible that H3K27me3 weakens LAD boundaries, rather than strengthening them. Indeed, Siegenfeld et al. also observed the presence of H3K27me3 at the boundaries of facultative LADs, which tended to gain NL association upon loss of H3K27me3[55]. However, the functional role of H3K27me3 at LAD borders remains unclear. Potentially, it could prevent the spread of constitutive chromatin features, such as H3K9me3, into neighboring regions[55].

While the presence of H3K27me3 at LAD boundaries does not necessarily contradict an antagonistic relationship between this mark and NL association, other evidence is harder to reconcile. Specifically, two studies have reported that NL association of several endogenous loci was lost upon inhibition of Ezh2, suggesting a direct dependency on H3K27me3 for peripheral localization[57,58]. Similarly, Siegenfeld et al. observed that a subset of H3K27me3-marked regions lost NL association upon Ezh2 inhibition, even though a majority gained NL association[55]. How can H3K27me3 seemingly promote contact with the NL for some loci, while hampering it in others? The answer could potentially lie in the interplay between NL affinity, H3K27me3 enrichment, and 3D contacts between H3K27me3 domains. In **Chapter 6**, we demonstrated that A/T content is a good predictor of intrinsic NL affinity, which can be counteracted by H3K27me3. The observed NL association thus depends on the balance between NL affinity and H3K27me3 level, at least in early mouse embryos. Notably, the studies reporting a loss of NL association upon Ezh2 inhibition specifically tested loci that contained genes[57,58], which are typically G/C rich. By choosing genic regions, the selection of loci may thus be skewed towards regions with low intrinsic NL affinity. In control conditions, these loci may be recruited to the NL via homotypic interactions with other H3K27me3-marked regions, some of which could have sufficiently high intrinsic NL affinity to drive the peripheral localization. Upon loss of H3K27me3, such homotypic interactions vanish and the tested loci would thus lose their peripheral anchor (Fig. 1b). This hypothesis could be investigated by further exploration of the data presented by Siegenfeld et al.[55]: Stratifying regions based on H3K27me3 level and A/T content could reveal whether the response to treatment is indeed dependent on a combination of these two factors.

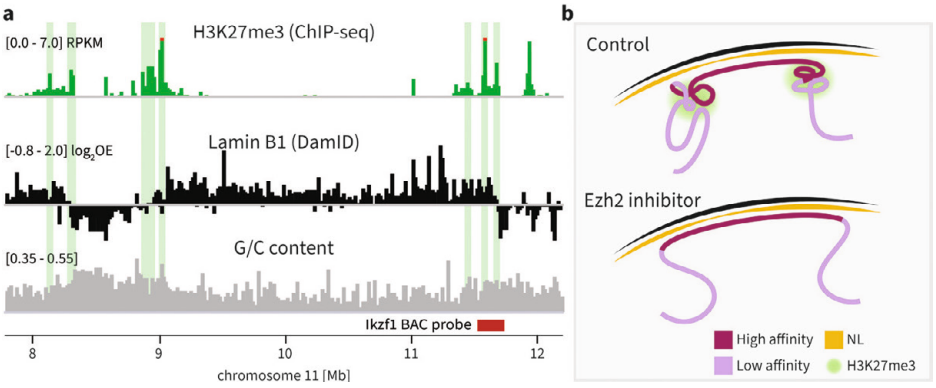


Figure 1: Speculative consolidation of conflicting reports on the role of H3K27me3 in nuclear lamina (NL) association
a, IGV browser tracks showing H3K27me3 signal around LAD borders in mouse fibroblasts. Top: H3K27me3 ChIP-seq signal (GSE48649); Middle: Lamin B1 DamID signal normalized for Dam control (GSE56990); Bottom: G/C content as a fraction of G/C nucleotides per bin. The selected region was chosen to match an example in one of the publications reporting reduced NL association of reporter loci upon Ezh2 inhibition[57] and the same data was used. The dark red box indicates the location of the BAC probe that was used for DNA FISH to determine NL association in treated and untreated cells. **b**, Hypothetical model to explain how Ezh2 inhibition results in increased NL association of some H3K27me3-marked regions, while resulting in decreased NL association of others.

Altogether, it thus seems likely that PRC2 activity hampers NL association. However, the mechanisms that cause this antagonistic relationship are entirely unknown. Moreover, it is currently unclear whether H3K27me3 or PRC2 itself is causing this effect. One possibility is that PRC2 and/or H3K27me3 impede the binding of one or more proteins promoting NL association, although little evidence is currently available to suggest which proteins these could be. A second explanation could be the capacity of H3K27me3-marked chromatin to form long-range homotypic interactions[59]. Such interactions could effectively pull chromatin away from the NL, especially if they occur with Polycomb domains in the more interiorly located compartment A. Indeed, H3K27me3 domains form extensive 3D contact domains in the early mouse embryo, which are absent in embryos lacking matPRC2 activity[60]. At the moment, little evidence is available to draw a concrete conclusion on whether either or both of these mechanisms contribute to the antagonistic effect of PRC2/H3K27me3 on NL association. A better understanding of LAD establishment in general may be required to conclusively answer this question.

What is the role of non-canonical NL association in the early embryo?

In mouse preimplantation embryos, H3K27me3 forms broad non-canonical domains with a high degree of allelic asymmetry[61], as discussed in **Chapter 5** and **6**. In **Chapter 6**, we have shown that these domains strongly overlap regions of canonical LADs, resulting in their dissociation from the NL. The atypical H3K27me3 profile may be the result of the A/T-binding domain of PRC1 subunit Cbx1, which has been shown to recruit PRC1 and PRC2 to paternal

pericentromeric regions in mouse zygotes[62]. Since canonical LADs are A/T rich[63], a similar mechanism could therefore give rise to the high levels of H3K27me3 in these regions. Due to extensive enrichment of H3K27me3 in canonical LADs, the antagonism between this mark and NL association directly affects ~50% of the genome and thus extensively impacts the 3D genome organization during early mouse development. This raises the question what the function of non-canonical H3K27me3 is in the early embryo and whether the effect of H3K27me3 on LADs plays a role in performing this function.

Several studies have investigated the role of matPRC2 in embryonic development by performing a maternal knock-out (mKO) of Eed, a core subunit of PRC2[60,64-66]. In these experiments, the conditional KO of maternal Eed occurs during oocyte maturation, which largely prevents the establishment of non-canonical H3K27me3 domains. As a result, H3K27me3 is dramatically reduced in fully grown oocytes, although low levels of H3K27me3 remain in patterns similar to wildtype oocytes[60,64]. Upon zygotic genome activation (ZGA), *Eed* starts to be transcribed from the paternal allele, resulting in H3K27me3 levels comparable to wildtype by the morula stage, although the non-canonical domains that are normally present at this stage are not re-established[64]. Prior to implantation, only a small number of genes is differentially expressed in Eed mKO embryos (115 and 40 differentially expressed genes in 4-cell and morula embryos, respectively)[64,65]. The effect of Eed loss on gene expression can be partially attributed to the role of H3K27me3 in establishing non-canonical imprinting[64]. In wild-type embryos, non-canonical imprinting is achieved by maternal-specific H3K27me3 domains that repress a number of genes (76 candidates) maternal allele, resulting in paternally biased expression[67]. Among these imprinted genes is Xist[64,66,68], the master regulator of X chromosome inactivation (see **Chapter 1**). Prior to implantation, female mouse embryos typically undergo inactivation of the paternal X allele, followed by reactivation and random X inactivation in the epiblast[69]. In the absence of H3K27me3, Xist is also activated on the maternal allele in the early embryo, causing all X chromosomes to be inactivated in both female and male embryos[64,66]. However, post-implantation the inactivated X chromosomes are reactivated and random X inactivation occurs[64,66]. In line with the relatively small changes in gene expression, Eed mKO embryos show no decrease in viability up to the blastocyst stage[64]. Post implantation, approximately half of Eed mKO embryos dies off with a bias in lethality towards male embryos, potentially due to lingering effects of the inactivation of their X chromosome[64,66]. The embryos that survive show relatively mild phenotypes[64].

From the results described above, it is clear that matPRC2 and H3K27me3 play a role in regulating gene expression and imprinted X inactivation during the early stages of embryogenesis. However, it is not clear whether these functions are in part mediated by the effect of H3K27me3 on NL association. In **Chapter 6**, we artificially tethered H3K27me3-marked chromatin to the NL with a high efficiency in early embryos, thus mimicking the effects of Eed mKO on LAD organization without removing PRC2 or H3K27me3. With this system, we tested the effects of altered NL association on transcription and viability up to the blastocyst stage, but found no differences compared to the control condition. In addition, we did not observe

any indication in the transcriptional data that imprinted X inactivation was affected (data not shown). As both H3K27me3 and NL association are typically associated with gene silencing, we have also considered the possibility that the relocalization to the nuclear periphery works as a redundant mechanism of gene repression, potentially explaining why relatively few genes are differentially expressed in Eed mKO embryos. However, genes that are de-repressed in Eed mKO show a similar gain in NL association as genes that do not become differentially expressed (data not shown). This argues against a redundant role of NL in gene repression, although it is possible that the small set of upregulated genes represent "escaper genes" that can override silencing by the NL[51]. Finally, we wondered whether the broad H3K27me3 domains and consequent reduction in NL affinity may play a role in protecting the early embryo from chromosomal segregation errors, as it was recently shown that chromosomes with stronger NL interactions have a higher chance of being missegregated[70]. To test this, we performed single-cell Karyo-seq (developed in **Chapter 2**) in embryos with peripheral tethering of H3K27me3 and control embryos, but observed no differences is large copy number variations (data not included in **Chapter 6**, but presented here in Figure 2).

Based on these results, it seems that exclusion from the NL is not required for the gene regulatory role of matPRC2-mediated H3K27me3 in the early embryo, nor does it protect the embryo from chromosome segregation errors. This lack of an observed developmental effect could be explained by several scenarios. First, it is possible that the NL does not fulfill a large role in regulating gene expression in the early stages of embryogenesis, potentially due to the absence of heterochromatic effector proteins.

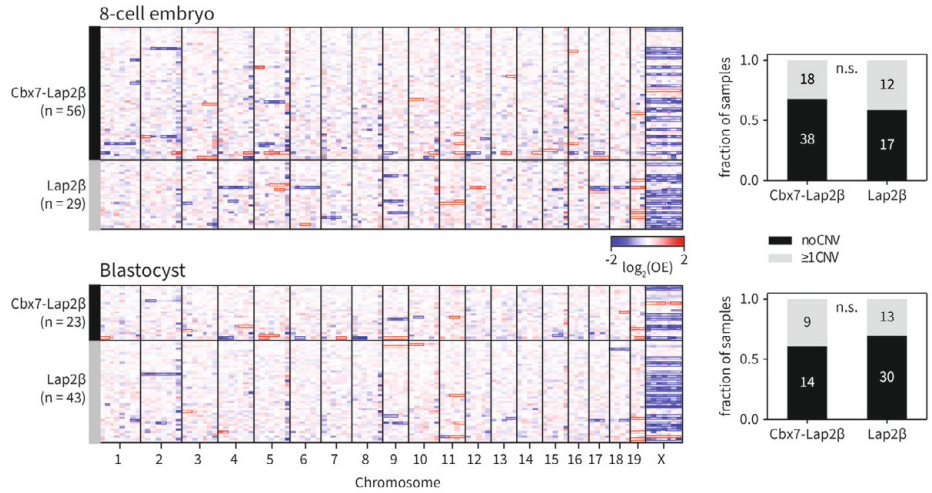


Figure 2: Tethering H3K27me3-marked chromatin to the nuclear lamina (NL) does not result in an increase of large copy number variations (CNVs)
Results from a scKaryo-seq experiment on embryos injected with mRNA encoding either Cbx7-Lapb (tether) or Lap2b (control). Embryo injection was performed as described in **Chapter 6**. Left: The heatmap shows the observed over expected genomic coverage of all somatic chromosomes and chromosome X. Putative CNVs are outlined in red (gain) and blue (loss) boxes. Right: The fraction of samples with somatic CNVs is indicated. The significance of the observed effect was tested with a Chi-squared test.

In this scenario, the atypical contact profiles could be a mere side effect of the non-canonical distribution of H3K27me3. This could also explain why the extensive variability in LAD profiles between embryos does not lead to a measurable effect on gene expression. If this is indeed the case, it would be interesting to determine at which point in development transcriptional control is commenced at the NL. This moment could very well lie around implantation, when several heterochromatic marks adopt their canonical form[61,71] (see **Chapter 5**). Second, the functional role of altered NL contacts could take place during oocyte maturation. Previous work by our lab has shown that oocytes have a near-complete lack of LADs[72], which could very well be explained by the non-canonical H3K27me3 domains that are established during oogenesis[61]. Similar to our hypothesis for the early embryo, diminished NL contacts may reduce the chance of chromosome segregation errors during meiosis. It is important to note here that Eed mKO mice generate an equal number of oocytes compared to wildtype, so lack of H3K27me3 and altered NL contacts is certainly not detrimental during oocyte maturation[64]. Third, non-canonical NL contacts could be important in processes that are not routinely observed in the lab, where mice are protected from environmental hazards (besides scientists and their experiments). A notable example is the capacity of some mammalian species to pause embryonic development at the blastocyst stage right before implantation, a process called embryonic diapause, which in mice happens in response to simultaneous lactation and pregnancy[73]. Interestingly, cells in diapaused embryos contain a substantial amount of condensed heterochromatin, which partially resides at the NL and disappears again in reactivated embryos[74]. Finally, it is possible that the atypical NL contacts do fulfill a functional role in normal development, but we failed to observe it. For example, several classes of retroviral elements are actively transcribed in the early embryo, including LINE L1 elements[75,76], which are strongly enriched in regions with high H3K27me3 and reduced NL association (**Chapter 6**). While we attempted to evaluate the effect of our different conditions on repeat expression, these analyses were severely hampered by the fact that scDam&T-seq only captures poly-adenylated transcripts and the difficulty of uniquely mapping repetitive sequences with short reads. Clearly, many potential avenues of investigation remain that could be explored to determine the role of NL association in early development. Such research could help us better understand the unusual heterochromatic state that is unique to the preimplantation embryo and the chromatin factors that contribute to totipotency.

Concluding remarks

Embryonic development forms the fundamental process of renewal that allows a species to persevere and evolve, while its individuals age and perish. Epigenetic control plays a crucial part in this process by enabling the emergence of a huge variety of phenotypes from a single genotype. As a consequence of this cellular expansion and diversification, the embryo contains a large pool of different cell types and intermediate states. It is impossible to understand the epigenetic control in this highly complex system with genomic assays that pool together thousands or millions of cells. Instead, single-cell multi-omics techniques have the capacity to chart this heterogenous system and start to draw mechanistic connections between

different layers of epigenetic regulation. In my PhD research, I have had the opportunity to contribute to the cycle of innovation and discovery that is fueling this new era of chromatin research. This has included the development and expansion of novel single-cell multi-omics techniques (**Chapter 2-4**) and their implementation to study heterochromatin in the early embryo (**Chapter 6**). Nevertheless, it is evident from the discussion above that numerous open questions remain. New advances in single-cell omics techniques will undoubtedly contribute to their further exploration, thereby advancing the goal of understanding the fundamental process of embryonic development.

References

1 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).

2 Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat Commun* **11**, 4307 (2020).

3 Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610-620 (2015).

4 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

5 Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).

6 Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **23**, 2126-2135 (2013).

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

8 Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134-147 (2015).

9 Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33**, 1165-1172 (2015).

10 Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* **24**, 494-515 (2023).

11 Pan, L., Ku, W. L., Tang, Q., Cao, Y. & Zhao, K. scPCOR-seq enables co-profiling of chromatin occupancy and RNAs in single cells. *Commun Biol* **5**, 678 (2022).

12 Xiong, H., Luo, Y., Wang, Q., Yu, X. & He, A. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat Methods* **18**, 652-660 (2021).

13 Zhu, C. *et al.* Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat Methods* **18**, 283-292 (2021).

14 Sun, Z. *et al.* Joint single-cell multiomic analysis in Wnt3a induced asymmetric stem cell division. *Nat Commun* **12**, 5941 (2021).

15 Bartosovic, M. & Castelo-Branco, G. Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. *Nat Biotechnol* **41**, 794-805 (2023).

16 Gopalan, S., Wang, Y., Harper, N. W., Garber, M. & Fazzio, T. G. Simultaneous profiling of multiple chromatin proteins in the same cells. *Mol Cell* **81**, 4736-4746 e4735 (2021).

17 Handa, T. *et al.* Chromatin integration labeling for mapping DNA-binding proteins and modifications with low input. *Nat Protoc* **15**, 3334-3360 (2020).

18 Meers, M. P., Llagas, G., Janssens, D. H., Codomo, C. A. & Henikoff, S. Multifactorial profiling of epigenetic landscapes at single-cell resolution using MulTI-Tag. *Nat Biotechnol* **41**, 708-716 (2023).

19 Stuart, T. *et al.* Nanobody-tethered transposition enables multifactorial chromatin profiling at single-cell resolution. *Nat Biotechnol* **41**, 806-812 (2023).

20 Lochs, S. J. A. *et al.* Combinatorial single-cell profiling of major chromatin types with MAbID. *Nat Methods* **21**, 72-82 (2024).

21 Kefalopoulou, S. *et al.* Time-resolved and multifactorial profiling in single cells resolves the order of heterochromatin formation events during X-chromosome inactivation. *bioRxiv*, 2023.2012. 2015.571749 (2023).

22 Xiong, H., Wang, Q., Li, C. C. & He, A. Single-cell joint profiling of multiple epigenetic proteins and gene transcription. *Sci Adv* **10**, eadi3664 (2024).

23 Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023).

24 Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403-411 (2021).

25 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

26 Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817 (2019).

27 Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020).

28 Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Res* **7**, 1740 (2018).

29 Ly, L. H. & Vingron, M. Effect of imputation on gene network reconstruction from single-cell RNA-seq data. *Patterns (N Y)* **3**, 100414 (2022).

30 Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity-current challenges and future perspectives. *Mol Syst Biol* **17**, e10282 (2021).

31 Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550-572 (2023).

32 de Luca, K. L., Rullens, P. M., Legube, G. & Kind, J. Genome-wide profiling of DNA repair identifies higher-order coordination in single cells. *bioRxiv*, 2023.2005. 2010.540169 (2023).

33 Zeller, P. *et al.* Single-cell sortChIC identifies hierarchical chromatin dynamics during hematopoiesis. *Nat Genet* **55**, 333-345 (2023).

34 Kind, J. *et al.* Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178-192 (2013).

35 Altemose, N. *et al.* muDamID: A Microfluidic Approach for Joint Imaging and Sequencing of Protein-DNA Interactions in Single Cells. *Cell Syst* **11**, 354-366 e359 (2020).

36 Zhang, D. *et al.* Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113-122 (2023).

37 Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**, 239 (2016).

38 Altemose, N. *et al.* DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. *Nat Methods* **19**, 711-723 (2022).

39 Lin, J. *et al.* scNanoCOOL-seq: a long-read single-cell sequencing method for multi-omics profiling within individual cells. *Cell Res* **33**, 879-882 (2023).

40 Philpott, M. *et al.* Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat Biotechnol* **39**, 1517-1520 (2021).

41 Shiau, C. K. *et al.* High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat Commun* **14**, 4124 (2023).

42 van Schaik, T., Vos, M., Peric-Hupkes, D., Hn Celie, P. & van Steensel, B. Cell cycle dynamics of lamina-associated DNA. *EMBO Rep* **21**, e50636 (2020).

43 Park, M., Patel, N., Keung, A. J. & Khalil, A. S. Engineering Epigenetic Regulation Using Synthetic Read-Write Modules. *Cell* **176**, 227-238 e220 (2019).

44 Szczesnik, T., Ho, J. W. K. & Sherwood, R. Dam mutants provide improved sensitivity and spatial resolution for profiling transcription factor binding. *Epigenetics Chromatin* **12**, 36 (2019).

45 Lotfollahi, M., Litinetskaya, A. & Theis, F. J. Multigrate: single-cell multi-omic data integration. *BioRxiv*, 2022.2003. 2016.484643 (2022).

46 Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* **18**, 272-282 (2021).

47 Du, J. H., Cai, Z. & Roeder, K. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT. *Proc Natl Acad Sci U S A* **119**, e2214414119 (2022).

48 Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914-927 (2013).

49 Finlan, L. E. *et al.* Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet* **4**, e1000039 (2008).

50 Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243-247 (2008).

51 Leemans, C. *et al.* Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**, 852-864 e814 (2019).

52 Demmerle, J., Koch, A. J. & Holaska, J. M. The nuclear envelope protein emerin binds directly to histone deacetylase 3 (HDAC3) and activates HDAC3 activity. *J Biol Chem* **287**, 22080-22088 (2012).

53 Su, J. H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659 e1626 (2020).

54 Brueckner, L. *et al.* Local rewiring of genome-nuclear lamina interactions by transcription. *EMBO J* **39**, e103159 (2020).

55 Siegenfeld, A. P. *et al.* Polycomb-lamina antagonism partitions heterochromatin at the nuclear periphery. *Nature Communications* **13**, 4199 (2022).

56 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).

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

58 Vertii, A. *et al.* Two contrasting classes of nucleolus-associated domains in mouse fibroblast heterochromatin. *Genome Res* **29**, 1235-1249 (2019).

59 Cheutin, T. & Cavalli, G. The multiscale effects of polycomb mechanisms on 3D chromatin folding. *Crit Rev Biochem Mol Biol* **54**, 399-417 (2019).

60 Du, Z. *et al.* Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Mol Cell* **77**, 825-839 e827 (2020).

61 Zheng, H. *et al.* Resetting Epigenetic Memory by Reprogramming of Histone Modifications in Mammals. *Mol Cell* **63**, 1066-1079 (2016).

62 Tardat, M. *et al.* Cbx2 targets PRC1 to constitutive heterochromatin in mouse zygotes in a parent-of-origin-dependent manner. *Mol Cell* **58**, 157-171 (2015).

63 Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* **23**, 270-280 (2013).

64 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).

65 Chen, Z., Djekidel, M. N. & Zhang, Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. *Nat Genet* **53**, 551-563 (2021).

66 Harris, C. *et al.* Conversion of random X-inactivation to imprinted X-inactivation by maternal PRC2. *Elife* **8** (2019).

67 Inoue, A., Jiang, L., Lu, F., Suzuki, T. & Zhang, Y. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature* **547**, 419-424 (2017).

68 Inoue, A., Jiang, L., Lu, F. & Zhang, Y. Genomic imprinting of Xist by maternal H3K27me3. *Genes Dev* **31**, 1927-1932 (2017).

69 Loda, A., Collombet, S. & Heard, E. Gene regulation in time and space during X-chromosome inactivation. *Nat Rev Mol Cell Biol* **23**, 231-249 (2022).

70 Klaasen, S. J. *et al.* Nuclear chromosome locations dictate segregation error frequencies. *Nature* **607**, 604-609 (2022).

71 Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**, 620-631 (2018).

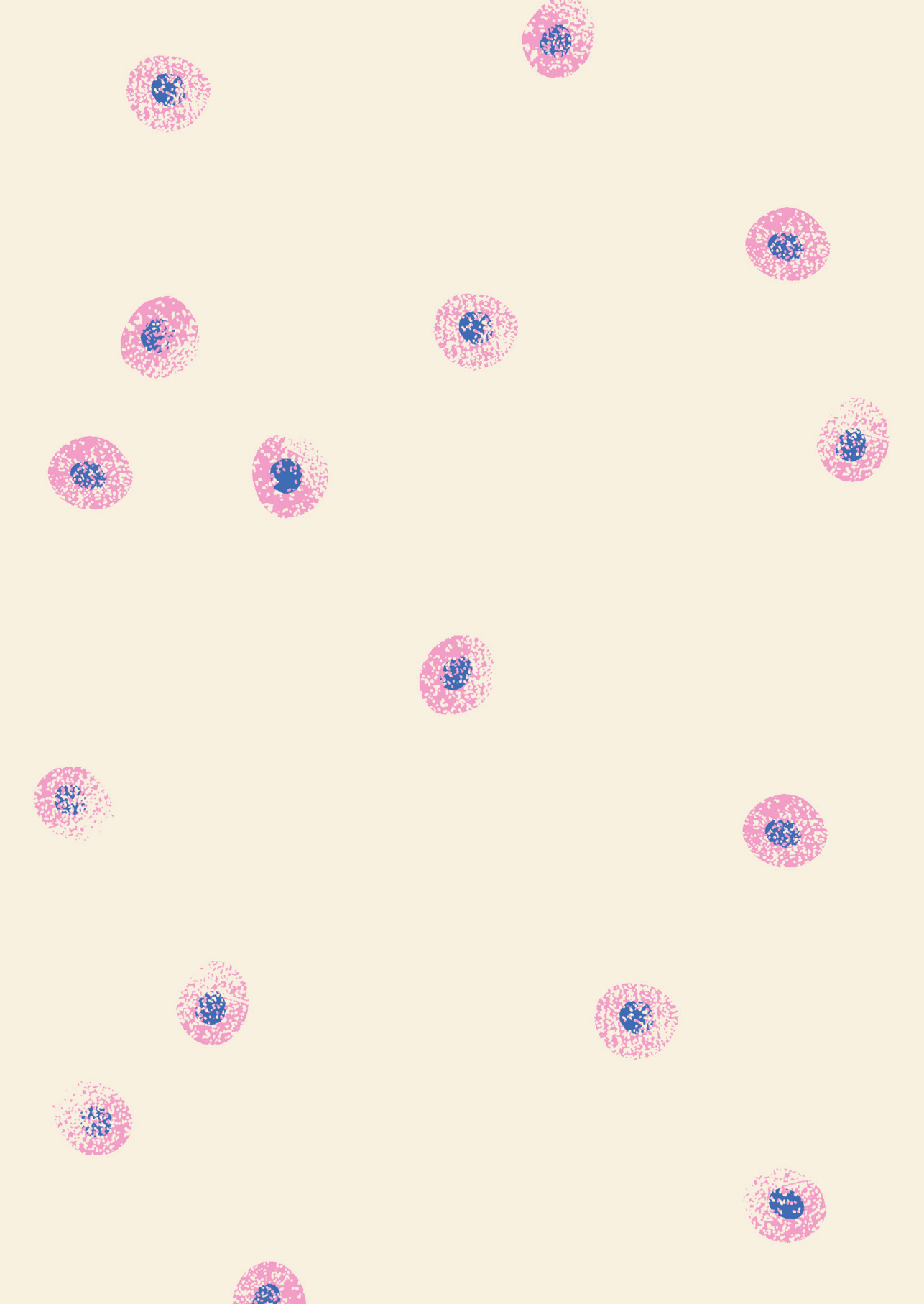
72 Borsos, M. *et al.* Genome-lamina interactions are established de novo in the early mouse embryo. *Nature* **569**, 729-733 (2019).

73 van der Weijden, V. A. & Bulut-Karslioglu, A. Molecular Regulation of Paused Pluripotency in Early Mammalian Embryos and Stem Cells. *Front Cell Dev Biol* **9**, 708318 (2021).

74 Fu, Z. *et al.* Integral proteomic analysis of blastocysts reveals key molecular machinery governing embryonic diapause and reactivation for implantation in mice. *Biol Reprod* **90**, 52 (2014).

75 Peaston, A. E. *et al.* Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7**, 597-606 (2004).

76 Fadloun, A. *et al.* Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat Struct Mol Biol* **20**, 332-338 (2013).



ADDENDUM

Summary

Every cell in our body contains the same genetic information, which is encoded in the DNA. This information is used to make the proteins that are responsible for all cellular functions. For this reason, the DNA is often referred to as the blueprint of the cell. Remarkably, all cells contain the same DNA, even though there are many different cell types with various forms and functions. To achieve this diversity, another layer of information is imposed upon the genetic code that determines which parts of the DNA are switched on or off. This extra layer of information is referred to as the epigenome (from the old Greek word “epi”, meaning “on top of”) and can be encoded in diverse forms, including small chemical modifications of the DNA and the composition of protein complexes around which the DNA is wrapped. All these forms of epigenetic information are regulated by specialized proteins. Disruption of the epigenome is associated with several diseases, including cancer. To study and understand the epigenome, it is important to measure both the interactions of proteins with DNA and which regions of the DNA are active. Ideally, these measurements are taken in single cells so they can be directly related.

In **Chapter 1**, I provide a general introduction on the epigenome, techniques that are used to study it on a single-cell level, and the principles of single-cell data analysis.

In **Chapter 2**, my colleagues and I develop scDam&T-seq, the first technique to simultaneously measure protein-DNA interactions and gene expression in single cells. Gene expression is the active use of DNA by copying specific pieces (genes) to use as blueprints for proteins. In **Chapter 3**, we provide an extensive protocol for scDam&T-seq and instructions on raw data processing.

In **Chapter 4**, we develop an extension of scDam&T-seq to enable its use for proteins that have been (temporarily) chemically modified. This is very relevant because the modification of certain proteins is an important form of epigenetic information. This technological extension is called EpiDamID.

In **Chapter 5**, I provide a literature review on epigenetic regulation during the very early development of mouse embryos, directly after fertilization of the oocyte by the sperm. This is an interesting moment in development, because the epigenetic information of the parents needs to be reset to an embryonic state and the single-cell embryo starts to divide to eventually give rise to all the cells of the body. I specifically discuss several forms of epigenetic modifications that are typically associated with gene repression and how these relate to the three-dimensional organization of the genome.

In **Chapter 6**, we apply scDam&T-seq and EpiDamID to further study the epigenetic state and spatial organization of the genome in early mouse embryos. We discover that the spatial organization is extremely variable in these early embryos. In particular, we find that there are large differences in which regions of the DNA are located at the periphery of the cell

nucleus. We relate this extreme variability to the unusual expansion of a particular epigenetic modification, called H3K27me3. Through further experiments we establish that the variable organization is the result of a thug-of-war between the affinity for the nuclear periphery as encoded in the DNA itself and repulsion from the periphery due to the presence of H3K27me3.

In **Chapter 7**, I discuss the current state of technological development in the field and how scDam&T-seq fits in. Since our publication, more techniques have been developed to measure protein-DNA interactions and gene expression in single cells. All of these are based on a fundamentally different strategy and are generally easier to apply. However, due to its unique design, scDam&T-seq can be applied in many ways and can also be incorporated into other techniques. Due to this versatility, scDam&T-seq will likely remain in use and be further adapted for different applications. In addition, I discuss our biological findings in more detail and compare them with the results of other researchers. For example, there are contradictory reports on the relationship between H3K27me3 and association of the DNA with the nuclear periphery. While one large study supports our findings, two others report that H3K27me3 *increases* association with the periphery. By carefully comparing the experimental designs and data, I propose a hypothesis that could explain these seemingly contradictory results.

Samenvatting

Iedere cel in ons lichaam bevat genetische informatie in de vorm van DNA. Deze informatie wordt gebruikt om eiwitten te maken die verantwoordelijk zijn voor het functioneren van de cellen. Daarom wordt DNA ook wel de blauwdruk van de cel genoemd. Opmerkelijk genoeg bevatten alle cellen hetzelfde DNA, ondanks het feit dat er veel verschillende soorten cellen zijn met compleet verschillende vormen en functies. Om deze diversiteit te bereiken is er nog een laag aan informatie bovenop de ruwe genetische code, die bepaalt welke delen van het DNA “aan” of “uit” staan. Deze extra laag wordt het epigenoom genoemd (“epi” betekent in oud Grieks o.a. “bovenop”) en neemt veel verschillende vormen aan, zoals kleine chemische modificaties op het DNA en de compositie van de eiwitcomplexen waar het DNA omheen gewikkeld is. Al deze epigenetische aanpassingen worden bewerkstelligd door gespecialiseerde eiwitten. Verstoring van het epigenoom is geassocieerd met verschillende ziektes zoals kanker. Om het epigenoom te begrijpen is het belangrijk om de interacties van eiwitten met het DNA te meten én vast te stellen welke delen van het DNA actief zijn. Idealiter vinden deze metingen plaats in individuele cellen, zodat ze direct aan elkaar gerelateerd kunnen worden.

In **Hoofdstuk 1** geef ik een algemene introductie op het epigenoom, technieken die gebruikt worden om het epigenoom te bestuderen op cellulaire resolutie, en de principes van de bijbehorende data-analyse.

In **Hoofdstuk 2** ontwikkelen mijn collega’s en ik scDam&T-seq, de eerste techniek ter wereld die in staat is om zowel eiwitbinding aan het DNA als genexpressie te meten in individuele cellen. Genexpressie refereert aan het actief gebruik van het DNA waarbij specifieke stukjes (genen) gekopieerd worden om als blauwdruk te gebruiken voor eiwitten.

In **Hoofdstuk 3** geven we een uitgebreid protocol voor het gebruik van scDam&T-seq en het verwerken van de ruwe data.

In **Hoofdstuk 4** ontwikkelen we een uitbreiding op scDam&T-seq, zodat deze techniek ook toegepast kan worden op eiwitten die kleine (soms tijdelijke) chemische modificaties hebben. Dit is heel relevant, omdat de chemische modificatie van verschillende eiwitten ook een belangrijke vorm van epigenetische informatie is. Deze uitbreiding heet EpiDamID.

In **Hoofdstuk 5** geef ik een literatuuroverzicht van epigenetische regulatie tijdens de zeer vroege ontwikkeling van muisembryo’s, vlak na de bevruchting van de eicel door de spermacel. Dit is een interessant moment in de ontwikkeling, omdat de epigenetische informatie van de ouders moet worden gereset naar een embryonale staat en het eencellige embryo begint met delen om uiteindelijk alle cellen van het lichaam te vormen. Ik ga specifiek in op verschillende epigenetische modificaties die (meestal) verantwoordelijk zijn voor het onderdrukken van genexpressie en hoe deze zich relateren tot de driedimensionale organisatie van het DNA.

In **Hoofdstuk 6** passen we scDam&T-seq en EpiDamID toe om de epigenetische staat en organisatie van het genoom in vroege muisembryo's verder te bestuderen. We komen erachter dat de spatiële organisatie van het genoom extreem variabel is in deze vroege embryo's. In het bijzonder zijn er grote verschillen in welke delen van het DNA zich aan de rand van de celkern vinden. We relateren deze extreme variabiliteit aan de ongebruikelijke uitbreiding van een specifieke epigenetische modificatie, genaamd H3K27me3. Door verdere experimenten stellen we vast dat de variabele organisatie van het DNA een gevolg is van een soort touwtrekken tussen de affiniteit voor de celkernrand die is gecodeerd in het DNA zelf en een afstoting daarvan door de aanwezigheid van H3K27me3.

In **Hoofdstuk 7** bespreek ik de huidige staat van technologische ontwikkeling in het veld en hoe scDam&T-seq daarin past. Ondertussen zijn er meer technieken verschenen die eiwitbinding en genexpressie kunnen meten in individuele cellen. Al deze technieken zijn gebaseerd op een fundamenteel andere strategie en zijn in veel gevallen makkelijker te implementeren. Door het unieke ontwerp van scDam&T-seq kan deze echter op veel verschillende manieren gebruikt worden en ook geïncorporeerd worden in andere technieken. Door deze veelzijdigheid zal scDam&T-seq dus waarschijnlijk in gebruik blijven en verder ontwikkeld worden. Verder bespreek ik onze biologische bevindingen in meer detail en vergelijk deze met resultaten van andere onderzoekers. Zo zijn er bijvoorbeeld tegenstrijdige bevindingen over de relatie tussen H3K27me3 en de associatie van DNA met de rand van de celkern. Terwijl een groot onderzoek onze bevindingen steunt, beschrijven twee andere onderzoeken juist dat H3K27me3 de associatie met de rand versterkt. Door kritisch naar de experimenten en resultaten te kijken stel ik een hypothese voor die deze ogenschijnlijke tegenstrijdigheid kan verklaren.

Curriculum Vitae

Franka Rang was born on December 20th, 1992, in Maastricht as the daughter of Marijn Rang and Nelleke Kleijn and grew up in the village of Eijsden. She is the oldest of four siblings, which further include her sister Kitty and brothers Felix and Koert. Franka attended high school at the Sint-Maartenscollege in Maastricht, from which she graduated cum laude in 2011. Later that year she started her Bachelor's degree at University College Utrecht, where she developed her interest for molecular biology and, specifically, epigenetics. She wrote her thesis under supervision of Prof. Dr. Mark Timmers and an honor's thesis under supervision of Prof. Dr. Johannes Boonstra on these topics. After graduating cum laude, she took a gap year in which she performed internships at the São Paulo State University in Botucatu, Brasil, and the Institute of Environmental Science and Research in Auckland, New Zealand. In 2015, Franka started the Master program "Cancer, Stem Cells and Developmental Biology" at the Graduate School of Life Sciences of Utrecht University. As part of this program, she performed research internships in the groups of Prof. Dr. Jop Kind (Hubrecht Institute, Utrecht), Prof. Dr. Ana Pombo (BIMSB MDC, Berlin), and Prof. Dr. Jeroen de Ridder (UMC Utrecht). In addition, she was chair of the CSND Student Committee, which organized various scientific and social activities for the students. Over the course of her Master's studies, Franka shifted her attention from laboratory work to data analysis and specialized in bioinformatics, graduating cum laude in 2017. Directly after finishing her studies, Franka returned to the group of Jop Kind at the Hubrecht Institute to start a PhD. She remained in the Kind group for six years and worked on various scientific projects, which were all eventually published and make up the main chapters of this thesis. In addition, she was part of the work council for three years, representing the interests of employees at the institute. At the end of 2023, Franka left the Kind group and shortly after, in February 2024, started as a postdoctoral researcher in the group of Prof. Dr. Jeroen de Ridder at the UMC Utrecht. In her new position, Franka will develop AI tools for the analysis of spatial transcriptomics data in the context of therapy resistance in melanoma. Currently, Franka lives in Zaltbommel together with her partner Alexander and her daughter Frida.

List of publications

Chapter 2

Rooijers, K.*, Markodimitraki, C.M.*, Rang, F.J., de Vries, S.S., Chialastri, A., de Luca, K.L., Mooijman, D., Dey, S.S.# and Kind, J.#, 2019. Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nature biotechnology*, *37*(7), pp.766-772.

Chapter 3

Markodimitraki, C.M.*, Rang, F.J.*, Rooijers, K., de Vries, S.S., Chialastri, A., de Luca, K.L., Lochs, S.J., Mooijman, D., Dey, S.S.# and Kind, J.#, 2020. Simultaneous quantification of protein–DNA interactions and transcriptomes in single cells with scDam&T-seq. *Nature protocols*, *15*(6), pp.1922-1953.

Chapter 4

Rang, F.J.*, de Luca, K.L.*, de Vries, S.S., Valdes-Quezada, C., Boele, E., Nguyen, P.D., Guerreiro, I., Sato, Y., Kimura, H., Bakkers, J. and Kind, J.#, 2022. Single-cell profiling of transcriptome and histone modifications with EpiDamID. *Molecular Cell*, *82*(10), pp.1956-1970.

Chapter 5

Rang, F.J., Kind, J.# and Guerreiro, I.#, 2023. The role of heterochromatin in 3D genome organization during preimplantation development. *Cell Reports*, *42*(4).

Chapter 6

Guerreiro, I.*#, Rang, F.J.*, Kawamura, Y.K., Kroon-Veenboer, C., Korving, J., Groenveld, F.C., van Beek, R.E., Lochs, S.J.A., Boele, E., Peters, A.H.M.F., Kind, J.#, 2024. Antagonism between H3K27me3 and genome lamina-association drives atypical spatial genome organization in the totipotent embryo. In press at *Nature genetics*.

* These authors contributed equally
# Corresponding author

Acknowledgements – Dankwoord

It feels quite unreal to sit down and write the acknowledgements to my dissertation, almost seven years after starting this journey. Seven years in which I have explored the world of epigenomics through data analysis. Seven years of paper reading, crossing my fingers as sequencing runs came in, trying to make sense of our results, feeling so smart, feeling so ignorant. Seven years of brainstorms, disappointment, encouragement, coffee, and chats with my wonderful colleagues. But also seven years of life with its own ups and downs. And, at the end of these seven years, this little book. While my name is on the cover, I was only a part of the tremendous effort required to make all this research possible. So, with pleasure, I would like to acknowledge and thank all the other people that contributed to the work.

First, let me thank you, **Jop**. I remember when I first came by in 2015 to discuss the possibility of performing my first master internship in your brand-new lab. Little did I know that we would be spending the better part of nine years working together. While the internship was not as fruitful as either of us would have liked, I did enjoy my time in your group and I am happy that you were willing to welcome me back as a computational PhD student, despite my limited expertise in that area. I have always enjoyed working together, discussing the various projects and bringing them to a successful end. Thank you for trusting me and giving me the space to work in my own way. The lab has evolved tremendously since I started, and I am curious to see where you will take it next.

Laat me dan de twee vrouwen bedanken met wie ik de eerste vier jaar van mijn PhD zij-aan-zij heb gewerkt: **Kim** en **Sandra**. Hoewel het aan het begin een betrekkelijk makkelijk project leek heeft EpiDamID (of EpiID/EpicID) ons doorzettingsvermogen ernstig op de proef gesteld. Zo'n beetje alles wat mis *kon* gaan is op enig moment ook echt misgegaan, vaak op manieren die we nooit hadden kunnen bedenken. Het is uitsluitend aan jullie te danken dat ik mijn toetsenbord niet uit het raam heb gesmeten en ben weggelopen. Maar aan het einde van de lange rit hebben we een paper neergezet waar we echt heel trots op kunnen zijn. Sandra, jouw pragmatische optimisme en eindeloos geduld waren allebei onmisbaar tijdens dit project, kudos voor de honderdduizend RPE1-experimenten die je hebt gedaan. Ik wil je ook bedanken dat je mij als beginnend masterstudentje al zo serieus nam, ook in de celkweek waar jouw expertise toch met hoofd en schouders boven mijn gepruts uitstak. Dat vertrouwen heeft een groot verschil gemaakt voor mijn wetenschappelijke zelfverzekerdheid. Kim, bedankt dat je bereid was samen met mij een onrealistisch hoge standaard na te streven, ook al maakten we het er onszelf niet makkelijker op. Ik heb ongelofelijk veel respect voor jouw kennis (op inhoudelijk, beleidsmatig, en politiek/sociaal vlak), je visie, en doorzettingsvermogen. Ik benieuwd naar wat je allemaal gaat bereiken de komende jaren! Als laatste bedankt dat je ook aan mijn zijde staat als paranimf tijdens de afsluiting van mijn PhD, het is een eer!

Isabel, you were the other important co-author in my life! I am writing this two days after our preimplantation paper FINALLY got accepted, and I am so incredibly pleased to bring

this chapter to a successful end before my defense. After working on the technical EpiDamID project, it was a tremendous joy to sink my teeth into the complex chromatin biology of early development. Thank you for accepting me so fully into the project and entrusting me with the analyses. I think we made a great team and really managed to uncover some interesting biology. I have been very impressed by your vision for this project and your incredible work ethos that was pushed to its limit during nine months of revision. You deserve this beautiful paper! I am sure there will be many more exciting projects and papers coming from you in the future, I will be keeping an eye out!

Dan zijn we aangekomen bij mijn andere fantastische paranimf: **Silke**! We kwamen er laatst achter dat we ons allebei niet kunnen herinneren wanneer we nou precies zo goed bevriend zijn geraakt. Waarschijnlijk ergens in de loop van 2020 door al het digitaal werken en een verandering van werkplekken. Hoe het ook zij, ik ben erg dankbaar dat het is gebeurd. Je bent uiterst intelligent, grappig en pragmatisch, in andere woorden de ideale vriendin en collega. Wat een feest was het dan ook dat we in september 2022 samen naar New York konden gaan voor de CSHL Epigenetics & Chromatin conferentie. Daarnaast ben je de enige persoon die ik ken die zo goed kan lullen én poetsen! Mijn enige gemis is dat wij nooit samen aan een project hebben kunnen werken. Misschien moeten we binnenkort maar eens onze koppen bij elkaar steken om te zien of we als PostDocs een samenwerking kunnen beginnen. Bedankt dat ik jouw paranimf mocht zijn en dat jij nu mijn paranimf bent, ik kijk er naar uit om mijn verdediging samen met jou te doen.

Then I would like to thank the rest of my colleagues. **Samy**, together with Silke and Kim, we formed the long-PhD-syndrome club. Thanks to the excellent company, this club turned out to be a largely joyful experience! I am very grateful for all the support we have been able to give each other throughout the years. I wish you all the best in the last part of your own PhD journey. You know I'm just a stone's throw away if ever you need a little distraction and/or support! **Pim**, **Robin** and **Marta**, it has been very nice to have you as my bioinformatic co-nerds and I am sure you will continue to bring the computational level in the Kind group to ever greater heights! **Moritz**, **Chris**, **Izz**, **Carlotta**, **Hidde**, **Lisa**, **Carla**, thank you for being such wonderful colleagues. Sitting in group meeting and hearing all the high-level discussions always impressed me with the amount of expertise and creativity in one room. Thank you also for your recent efforts for Silke's PhD movie, it really warmed my heart so see everyone so involved. Best of luck to the new colleagues, **Fieke, Michelle, Taleen**, enjoy your scientific adventures in the Kind group!

Besides these recent colleagues, I also want to thank all my former colleagues that played such important roles during my first years. **Corina**, bedankt voor de positieve en chaotische energie die je iedere dag mee naar kantoor nam; geen saaie dag met jou als collega! Het was heel fijn om samen te werken aan het Nature Biotechnology paper en daarna samen het Nature Protocols paper te schrijven. Ik ben heel blij dat we allebei deze mooie hoofdstukken aan onze proefschriften hebben kunnen toevoegen! **Koos**, bedankt voor alle hulp die je mij tijdens de

eerste twee jaar hebt geboden. Dankzij jouw pipeline en input kon ik echt een vliegende start maken, en ik heb ook zeker jou te danken voor deze twee hoofdstukken. Also many thanks to **Sara**, **Tess**, **Ramada**, **Ellen**, **Leila** and **Femke** for all the fun times!

Alexander, bedankt dat je mijn promoter bent geweest deze jaren. Ik stelde je input tijdens mijn committee meetings zeer op prijs. Hetzelfde geldt voor jou, **Jeroen** (de Ridder), ik waardeer het erg dat je mijn ontwikkeling en belang zo voorop zette tijdens deze discussies. Daarnaast ben ik ontzettend blij dat ik dit jaar als PostDoc bij jou aan de slag kon, ik kijk ernaar uit om de komende jaren samen te werken! Verder wil ik **Jeroen** (Bakkers) bedanken voor onze samenwerking aan het EpiDamID project en nu ook voor het voorzitten van de leescommissie. I would also like to thank all the members of my assessment committee for taking the time to read my dissertation: **Jeroen**, **Sanne**, **Tuncay**, **Bas** and **Gert Jan**. I am looking forward to your thoughts and questions during the defense.

Before leaving the world of science, I would still like to thank the external collaborators whose contribution was crucial in various papers: **Phong Nguyen**, **Yumiko Kawamura**, and **Antoine Peters**!

Then, let me thank my friends who have supported me from the sidelines. Als eerste natuurlijk jou, **Rosine**, mijn beste vriendin sinds ik me kan herinneren! Bedankt dat je er altijd voor me bent. Ik kan niet geloven dat we na twee compleet verschillende academische carrières één dag na elkaar promoveren! Ik ben zo trots op jou en ik verheug me op de rest van onze levenslange vriendschap! **Liselotte**, **Laura**, **Anna**, **Sarah**, jullie ook bedankt voor jullie vriendschap en steun sinds de middelbare school! Dankzij jullie kan ik nog steeds met zo veel plezier nadenken over die tijd en ik weet zeker dat ons hoge streberniveau de basis heeft gelegd voor mijn verdere academische ontwikkeling. Het is enorm leuk om te zien wat voor een verschillende levens jullie opbouwen en ik hoop daar nog héél lang getuige van te mogen zijn. **Valeria**, thank you for lending me your crazy creative mind at times, and regaling me with your unbelievable adventures! You are a superhero, and it is only because of your generous personality that I never got jealous of all your scientific achievements. Keep rocking! En dan moet ik natuurlijk mijn lieve vrienden van FriDi bedanken: **Annerijn**, **Marleen**, **Joost**, **Mara**, **Floortje**, **Sander**, **Lieke**, **Robert**, **Rens**, **Kavish** en **Diederik** (en jullie fantastische partners)! Jullie zijn mijn lijntje naar de rest van de wereld, een constante bron aan gezellige uitjes en goede verhalen. Dankjewel voor alle goede tijden die we de afgelopen 10+ jaar hebben gehad! Ook veel dank aan **Tico** en **Lotte**, de beste buren, met wie ik ongeremd mijn meest nerdy zelf kan zijn en zonder wie de coronatijd een stuk somberder zou zijn geweest!

Pap, **mam**, jullie zijn het fundament waarop ik mijn leven heb kunnen bouwen. Dankzij jullie steun in allerlei verschillende vormen heb ik me zorgeloos kunnen ontwikkelen tot de persoon en wetenschapper die ik vandaag ben. Een paar zinnen in een dankwoord doen er geen recht aan, maar ik wil toch van deze gelegenheid gebruik maken om te zeggen dat ik jullie extreem dankbaar ben. Daarnaast wil ik natuurlijk ook mijn zus en broers bedanken. **Kitty**, **Felix** en

Koert, ik wil niet klef doen, maar jullie zijn uitstekende Geschwister! Het is een genot om jullie te zien ontwikkelen van nutteloze ukkepukkies tot volwaardige mensen. Kitty, het was heel fijn om jou zo dicht bij te hebben in Gouda, vooral tijdens corona. Ik mis je, verhuis je snel naar Zaltbommel?

Alex, my love, you have been by my side during my whole PhD and it is hard to express in words how much that has meant to me. While you probably still don't exactly know what my research is about, you have greatly contributed to it. You inspire me with your stoic endurance, your strong intrinsic drive, and your superhuman ability to not be offended and take things as they come. The fact that I have rarely been stressed can be largely attributed to your presence and influence. I love our life together, our many shared hobbies, our silliness, and now also our shared parenthood. Thank you for making my life so good, I love you! Dan, **Frida**, misschien lees je dit over 10 of 20 jaar wanneer mijn PhD alweer prehistorisch lijkt. Ik zou niet durven beweren dat je mijn leven makkelijker hebt gemaakt, maar wel oneindig veel interessanter en mooier. Iedere ochtend ben ik blij om je uit bed te halen en iedere werkdag kijk ik er naar uit om je 's avonds weer te zien. Je komst was ook een goede stok achter de deur om alles toch maar eens een keertje af te maken. Bedankt dat je er bent, moppie, ik hou van je!