A tale of two measurements

Protein-DNA interactions & gene expression in single cells

Corina Maria Markodimitraki

The work described in this thesis was performed at the Hubrecht Institute for Developmental Biology and Stem Cell Research (the Royal Netherlands Academy of Arts and Sciences, KNAW) within the framework of the research school Cancer Stem cells & Developmental biology (CS&D), which is part of the Utrecht Graduate School of Life Sciences (Utrecht University).

Cover: ' Jupiter's Great Red Spot: A Rose By Any Other Name' obtained from the Juno spacecraft.
Image credits: NASA/JPL-Caltech/SwRI/MSSS. Image processing by Mary J. Muprhy.
Cover design: Corina Maria Markodimitraki
Layout: Corina Maria Markodimitraki
Printing: Ridderprint

ISBN: 978-94-6416-075-8

A tale of two measurements

Protein-DNA interactions and gene expression in single cells

Een geschiedenis van twee metingen

Eiwit-DNA interacties en genexpressie in enkele cellen

(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 15 oktober 2020
des middags te 12.45 uur

door

Corina Maria Markodimitraki

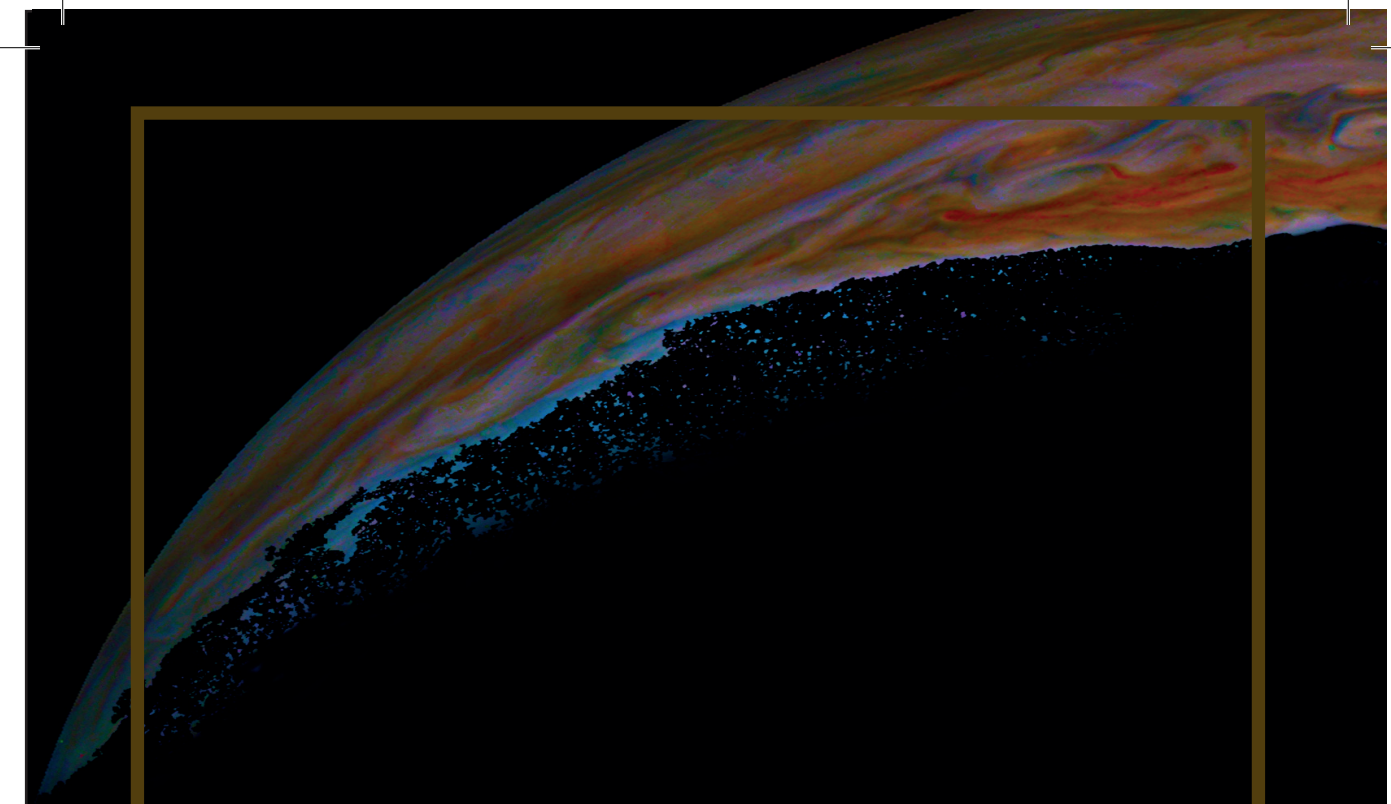
geboren op 4 april 1990
te Herakleion, Griekenland

Promotor:
Prof. dr. ir. A. van Oudenaarden

Copromotor:
Dr. J.H. Kind

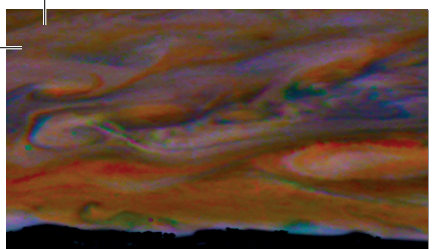
Table of contents

Για το Δημήτρη και τη Λάουρα



Chapter 1

Corina M. Markodimitraki



1

Organisation of the 3D genome

Cells in multicellular organisms all contain the same genetic information, yet vary vastly in morphology and function. Precise gene regulation within each cell is at the base of this great diversity. Tailor-made gene expression is possible through a variety of regulatory pathways that act at different levels, from the direct modification of deoxyribonucleotides in the DNA sequence to the 3D genome organization within the cell nucleus. Packaging the 2-meter-long human DNA fiber into a micron-sized nucleus and keeping the correct transcriptional profile is achieved mainly by the organization of the genome into the chromatin fiber (see below). Chromatin can form large domains that separate transcriptionally active from inactive chromatin regions. This thesis focusses on inactive chromatin regions that are spatially positioned at the periphery of the nucleus and in association with the nuclear lamina (NL). These regions are coined lamina associated domains (LADs) and how they contribute to gene regulation and cell identity has been a major focus of my work.

Genome folding is non-random, occurs at multiple levels (illustrated in Figure **1**), and its basal unit is the chromatin fiber, defined as DNA bound by histones and other chromatin-associated proteins. Histones, the most abundant proteins in the cell, bind the DNA in the form of nucleosomes around which 147 base pairs can be wrapped[1]. A nucleosome typically consists of 2 copies of each of the H2A, H2B, H3 and H4 histones, and these are interchangeable with histone variants that vary according to the state of the cell[1]. Moreover, histone tails protruding from the nucleosome core can be post-translationally modified. The addition of negatively or positively charged groups on the carboxy- and amino-histone tails termed histone post-translational modifications (PTMs) can alter how the nucleosomes interact with each other and the DNA[2]. This leads to locally more compact or relaxed chromatin, which shields or exposes transcription factor binding sites, thus regulating transcriptional output.

Apart from chromatin accessibility, gene expression is regulated by DNA elements such as enhancers and promoters[3]. These are often separated by large genomic distances in *cis*, but can come together in the nucleus, aided by the spatial conformation of the chromatin fiber. This is facilitated by the formation of chromatin loops and allows for the transcriptional machinery to assemble[4]. Enhancer-promoter loops can be cell type-specific, thus contributing to an additional level of gene regulation[4]. Multiple long-range interactions can be found accumulating along stretches of hun-

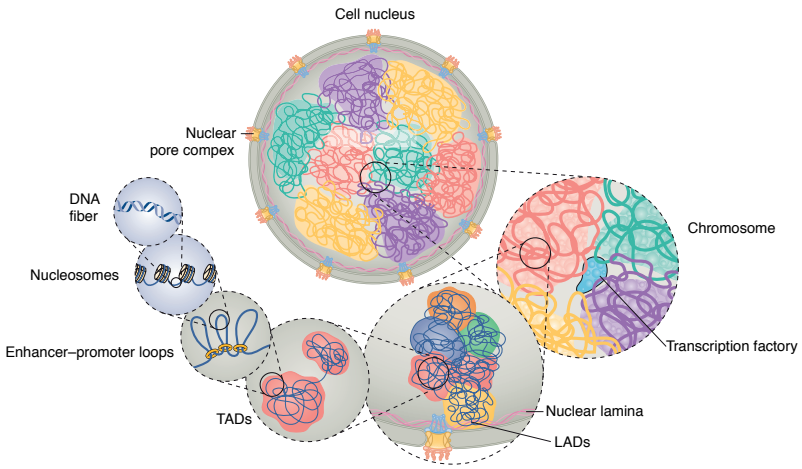


Figure 1 • The genome is organized in structural units. Adapted from [58].

dreds of kilobases in size, thereby forming "hubs" of three-dimensional chromatin conformations, termed topologically-associated domains (TADs)[5]. Architectural proteins like the cohesin complex and the CTCF transcription factor contribute to the organization of TADs, which can differ between cell types, forming another layer of gene regulation[6].

TADs represent a higher level genomic organization at the megabase scale, and can be categorized into active or inactive nuclear territories, termed A and B compartments, respectively[7, 8]. Regions in compartment A are gene-dense and accessible, and are characterized by active histone PTMs, whereas regions in compartment B are inaccessible and are characterized by gene-desserts as well as repressive histone PTMs. The two compartments also occupy different regions in the nucleus: compartment A localizes towards the more transcriptionally-permissive interior, whereas compartment B is positioned more radially and interacts with the nuclear periphery. A key structural component that contributes to this segregation is the NL, consisting of type V filamentous proteins which form a protein meshwork lining the nuclear envelope. The genomic regions interacting with the NL form distinct domains, termed LADs[9, 10].

LADs and their characteristics

LADs were first identified as genomic organizational units in the *Drosophila melanogaster* genome by the DamID technology (Figure 2)[10]. DamID, short for DNA adenine methyltransferase identification, makes use of the bacterial methyltransferase protein Dam which, when in proximity, can stably methylate adenines within a GATC context. When Dam is coupled to a chromatin-interacting protein, the regions bound by the protein will be marked by Dam in the form of the m^6A modification. This mark can be specifically recognized and digested by the restriction enzyme DpnI, leaving blunt genomic free ends. By ligating double-stranded DNA adapters to the free ends, one can amplify these regions by PCR and subsequently identify them with the use of next generation sequencing. For the identification of LADs by sequencing, Dam is typically anchored at the NL through a fusion to the protein LaminB1, and expansions of the DamID technique have been able to identify LADs in single cells by sequencing and microscopy[11, 12].

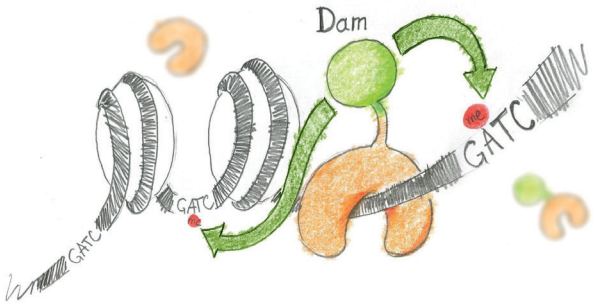


Figure 2 • The DamID technology. Illustration by Guillaume Filion.

The property of chromatin interacting with the NL appears to be a conserved trait. Apart from the fruit fly, LADs have also been mapped in nematode worm, pre-implantation mouse embryos, as well as mouse and human cell lines[13-16]. Across these species, LADs have common characteristics. They occupy about 40% of the genome and are relatively gene-poor compared to regions not associating with the NL, termed inter-LADs (iLADs)[10, 14, 15, 17, 18]. One can find about 1,000 - 1,500 LADs in human and mouse cells and their median size is 0.5 Mb, although they can span across regions of 10 kb to 10 Mb genome-wide[14, 15].

Regions contacting the NL across different cell types are termed constitutive LADs (cLADs) and they are thought to make up the "backbone" structure that facilitates the tethering of chromatin to the NL. In agreement with a possible function as 3D-anchoring regions are the genomic features associated with cLADs. Namely, they are conserved between human and mouse and are especially gene-poor. They are also A/T-rich and contain long interspersed nuclear elements (LINEs)[18]. LADs appear as condensed DNA near the nuclear periphery in early electron microscopy experiments[9], thus it is perhaps not a surprise that they carry typical characteristics of heterochromatin. This is in contrast to cell type-specific LADs or facultative LADs (fLADs), which are more gene-dense and are variable between different cell types. This is in line with the necessity of these genes to be transcribed in a cell type-specific manner[18].

LADs are enriched for the heterochromatin marks H3K9me2/3 and LAD borders show enrichment for the Polycomb group-related heterochromatin mark H3K27me3 in some cell types[11, 12, 14, 19, 20]. They are also late-replicating[14, 15, 21], overlap with B compartments[12] and are mostly devoid of active chromatin marks such as H3K4me2 and H3K36me3[11, 14, 15, 22]. LADs show low RNA polymerase II occupancy and transcription levels of most genes located in NL-binding regions are low if detected at all[14, 15]. The repressive effect of the NL is demonstrated by a genome-wide study integrating thousands of reporters in parallel into the human genome in a random manner[23]. Reporters integrated in LADs generally show 5-6 fold lower activity than those in iLADs[22].

Regulation of gene expression at the nuclear lamina

Gene regulation at the NL varies between species and promoters

The mechanisms of gene regulation at the NL have been elusive because even though as a rule LAD genes are not expressed, a tenth "escapes" the repressive effect of the NL and shows transcriptional activity[14, 22, 24, 25]. Studies have attempted to address whether the NL reinforces transcriptional repression or if regions localize to the NL as a result of their inactivity. Experiments in Drosophila and mammalian cells show that some reporters and endogenous genes become repressed upon movement to the periphery, while others appear almost unaffected by this relocalization[26-29]. Other studies testing loci that naturally interact with the periphery, show that NL association has a repressive effect on transcription upon direction of these loci towards the NL[30, 31]. Together, these results suggest an overall repressive effect of NL association on transcription but outcomes can vary between organisms, promoters or reporter genes.

The varying results of the abovementioned experiments could be attributed to differences in promoter sensitivity towards the repressive environment of the NL. This is suggested by a recent genome-wide study in human cancer cells which grouped LAD promoters according to their sensitivity to the NL environment[22]. Putatively inactive LAD promoters were integrated in episomal plasmids and their activities checked. Whereas one group of promoters stayed silent, another one was activated. This leads to the conclusion that the NL hosts both already repressed loci and loci that are otherwise active, which the NL environment can repress. The study also addressed the case of LAD promoters that are naturally transcribed. These promoters were shown to contain sequence motifs that could be recruiting transcription factors, helping them overcome the repression[22]. Even though this study helps understand differences in regulation between loci, the exact mechanisms that govern transcriptional silencing at the NL remain to be elucidated. Gene silencing is likely the result of multiple phenomena that complement each other such as the embedding of DNA into the NL meshwork, the absence of active transcription marks and the enrichment for heterochromatin marks.

Gene regulation at the NL through histone PTMs

Repressive histone marks enriched at the NL seem to play a role both in transcriptional repression and in LAD anchoring but findings differ between organisms of study. In *C. elegans* for example, H3K9me1/2/3 is enriched at the nuclear periphery, and once all the corresponding histone methyl transferase enzymes are depleted, chromosomal arms detach but are not transcriptionally activated[32]. In mouse embryonic stem cells (mESCs) on the other hand, the conditional knockout of the G9a enzyme catalyzing H3K9me2, results in loss of peripheral H3K9me2 and upregulation of a selection of LAD genes[20]. When the chromatin environment of LADs is disrupted in mouse and human cells by a transcriptional activator peptide, LADs decondense but their genes are not transcriptionally activated[11, 33]. Thus, mechanisms of repression and anchoring at the NL through histone PTMs could differ between species.

Gene regulation at the NL through "locking-in" of genes

The NL could additionally have a more passive way of controlling gene expression by serving as a docking station for "locking in" states of transcriptional repression. The structure of the protein meshwork lining the inner nuclear membrane could facilitate this, as indicated by super-resolution microscopy studies showing LADs embedding into the protein meshwork of the NL[12]. An example indicating such a regulatory role of the NL comes from a study showing that human somatic cells with low LaminA -one of the components of the lamina meshwork- are more efficient in reprogramming than cells with high LaminA levels. This implies that the NL could prevent the promiscuous relocation of pluripotency genes by safely "locking" them in the protein meshwork and keeping them repressed[34]. In Drosophila on the other hand, Lam Dm0, which is the B-type lamin in fruit fly, appears to be crucial in the regulation of testis-specific genes, since its depletion alters the gene cluster's localization, resulting in their upregulation[35]. Depletion of lamin, another component of the NL in Drosophila, results in the extension of neuroblast competence during neuroblast differentiation due to the inability of anchoring and silencing the Hunchback gene (Hb)[36]. These findings allude to a model in which the lamina meshwork serves as a platform for maintaining gene silencing.

Besides the embedding of the genome into the lamin meshwork, proteins that are part of the nuclear lamina could also play a role in the establishment of genome-NL contacts. This is demonstrated in *C. elegans* where LADs are established by the binding of the NL protein CEC-4 to H3K9me3, and this interaction is crucial for proper lineage specification[37]. Similarly, NL component Lamin B receptor (LBR) is crucial for radial recruitment of the Xist transcript, and therefore for the relocation of the X chromosome to the NL where it will be silenced, during differentiation of female mESCs[38]. Thus, it seems that the NL components are required for the maintenance of transcriptional repression, especially during the acquisition of cell fates.

The role of LADs in differentiation and development

Genome organization undergoes dramatic changes upon differentiation

During differentiation, the genome undergoes global rearrangements in order to facilitate cell type-specific transcriptional programs[39]. The openness and plasticity that characterize the stem cell genome make way for more defined and robust compartments and TADs. Microscopy studies show that chromatin becomes more condensed at the nuclear periphery as stem cells differentiate[40-42]. Changes in LAD organization are also apparent from DamID experiments of mESCs differentiating into neural precursor cells (NPCs) where hundreds of iLADs and LADs are found to relocate towards or away from the nuclear periphery, respectively[15]. Similar rearrangements can be observed in other murine differentiation systems such as the induced myogenesis of mouse fibroblasts, car-

diac differentiation of multipotent cardiac progenitor cells and neural or cardiac differentiation of multipotent mouse P9 embryonal carcinoma cells[43-45]. Human ESCs (hESCs) also show distinct nuclear localization of pluripotency genes than differentiated cells. The pluripotency gene NANOG for example, localizes to the more transcriptionally permissive nuclear interior, while in differentiated cells it moves to the periphery[46]. Somatic cells of the fruit fly on the other hand, show a radial positioning and silencing of testis-specific gene clusters, and only upon germ-line differentiation do they relocate to the interior of the nucleus and get activated. Thus, peripheral chromatin rearranges upon differentiation across organisms of study.

Genomic rearrangements and transcription during differentiation

The genomic rearrangements that happen during differentiation are often accompanied by transcriptional changes that contribute to the acquired cell identity. Stemness genes become inactivated and "locked in" at the periphery whereas genes related to the acquired cell fate move towards the nuclear interior and are transcribed. For example, during the mESC to NPC transition, the pluripotency genes Oct4, Nanog and Klf4 increase their interactions with the NL and are downregulated. Inversely, genes important for neural identity show decreased contacts with the periphery and upregulation[15]. These transcriptional changes that accompany shifts in nuclear architecture, indicate that genome organization contributes to lineage specification. Experiments done in Drosophila support this as well. Incapacity to relocate loci to the NL results in failure to commit to a muscle cell fate and similarly, inability to target the *hb* gene to the NL for repression results in the extension of neuroblast competence[36, 37]. Thus, genomic rearrangements at the NL alongside their accompanying transcriptional changes are detrimental for proper lineage acquisition.

Not all changes in genome architecture during differentiation are coupled to changes in transcription, however. Some relocated regions in NPCs for example, only show gene expression changes upon further development into the more terminally-differentiated astrocytes[15]. These poised transcriptional states are also observed during neuroblast differentiation in Drosophila, where silencing of the *hb* gene seems to precede its radial relocation[36]. Lastly, adipocyte-specific genes relocate towards the transcriptionally-permissive nuclear interior upon human adipose cell differentiation, and are either activated immediately or at later stages of differentiation[47]. The above examples indicate that NL embedding is only one of the ways the genome can regulate transcriptional repression, and that additional mechanisms such as chromatin features of LADs are at play.

Changes of chromatin features during lineage acquisition

The changes in genome architecture during differentiation are often facilitated by LAD chromatin. In *C. elegans* and Drosophila, developmental promoters acquire tissue-specific localizations within the nucleus, with the aid of NL protein "anchors" that recognize and bind H3K9me1/2/3[10, 37, 40, 41, 48]. What's more, LAD reorganization during differentiation is often accompanied by a change in LAD chromatin. During differentiation, the nuclear periphery becomes progressively more enriched in H3K9me3. When H3K9 methylation is inhibited, differentiation is delayed and depletion of this mark can enhance reprogramming[41, 49]. Changes in LAD chromatin can be seen also in the nuclei of multipotent mouse P9 embryonal carcinoma cells, in which H3K9me2-rich loci involved in neurogenesis or cardiogenesis lose the H3K9me2 mark as they move from the nuclear periphery to the interior upon differentiation, and in some cases this is accompanied by transcriptional activation[45]. LAD chromatin is thus an important contributor to the faithful acquisition of cell fates.

Rationale behind this thesis

Single-cell measurements of LADs are necessary to understand the changes in genome organization during differentiation.

Since the advent of next generation sequencing, single-cell techniques have become increasingly popular because they can help address biological questions which bulk measurements are not suited for. Single-cell RNA sequencing for example, can reveal transcriptional heterogeneity within the same cell type of an organ, which would otherwise not be apparent from bulk measurements. For example, beta cells of the adult human pancreas can differ in the amounts of insulin secretion and response to external stimuli, which can influence the overall function of the organ[50]. Additionally, during differentiation, transcriptional information at the single-cell level can reveal intermediate transcriptional states, such as the epiblast state during the mESC to NPC transition[51]. Finally, when studying a complex tissue, single-cell transcriptomics can identify new, common and rare cell types[52, 53]. Taken together, important information can be obscured by measurements that are based on thousands or millions of cells and can only be revealed with single-cell approaches.

Following the development of single-cell transcriptomics, the establishment of other single-cell techniques has been on the rise. Genomic techniques such as Hi-C or ATAC-seq have been adjusted to the single-cell level[54, 55]. In a similar fashion, DamID has been recently developed further into a single-cell technique and has revealed details of LAD biology previously unachievable[12]. When applied on cells of a human monoclonal leukemia cell line, single-cell DamID (scDamID) uncovered cell-to-cell variation in the frequency with which genomic regions contact the NL. Regions with high contact frequency (CF) contacted the NL in more than 80 % of measured single cells, and interestingly, these regions seemed to coincide with cLADs[12, 18]. These regions were more strongly repressed than regions of lower CF, and enriched for non-myeloid related genes such as olfactory receptor genes. These genes do not seem to contribute to the expression profile of a leukemia cancer cell as they are repressed throughout the cell population. Cell-to-cell heterogeneity in CF is also confirmed by microscopy experiments in cancer cells which show that only one third of the annotated LADs contact the NL at a given time point in a single cell[11]. This might be explained by the dynamic nature of the interaction of the genome with the NL, as microscopy experiments show that LADs are mobile during interphase and can move as far as 1 μm away from the nuclear periphery[11]. It could also be attributed to the fact that once a cell undergoes mitosis, LADs rearrange in part randomly: some mother LADs re-establish as LADs at the periphery, while others associate with the nucleolus (nucleolus associating domains; NADs), possibly meaning that LADs could be repressed in different subnuclear structures between cells of a homogenous cell population[11, 56]. Taken together, single-cell studies of LAD organization reveal that cancer cells show a cell-to-cell variation in the nuclear positioning of their genome. How this is translated to NL associations in the process of differentiation however, has not been explored yet. It is of interest to understand variability in NL associations in differentiation, especially since failure to position the genome at the NL can result in an inability to faithfully acquire a cell fate[37].

As discussed in previous paragraphs of this introduction, when stem cells differentiate, their open and plastic genome undergoes dramatic rearrangements and chromatin density at the periphery increases[40, 41, 57]. However, based on population DamID studies, the proportion of the genome localizing at the periphery in ESCs and in differentiated cells is similar[15]. A possible explanation is that the dynamic genome-NL interactions in the nuclei of stem cells might change as soon as cells are streamlined towards a cell fate, thereby "locking in" transcriptional states[16]. This would be translated to less dynamic genome-NL interactions in differentiated cells, and therefore less cell-to-cell

differences in NL associations. In favor of this theory comes the finding that mESCs show increasingly robust intra-TAD and B-B-compartment interactions as they progress along the differentiation trajectory[39]. In order to investigate this however, one would need to investigate the changes in genome-NL interactions during a differentiation trajectory from a single-cell perspective.

The need to couple genome architecture to transcriptional heterogeneity in single cells.
Single-cell DamID studies show that frequency of NL association is inversely correlated with mean levels of transcription. This means that high CF LADs are more strongly repressed than low CF LADs. This association however is based on separate DamID and mRNA measurements, and it is not clear how the two are associated and if genome organization at the NL directly influences transcriptional output. In order to directly link the transcriptional profile of a cell to its nuclear architecture, it is necessary to obtain measurements from the same cell simultaneously.

Integrated measurements could not only help linking transcription to NL positioning, but prove an invaluable tool for uncovering chromatin organization in different cell types or states. Within a heterogeneous cell population such as a pool of differentiating cells or a complex tissue such as the brain, transcriptional information is necessary to identify cell states while DamID can subsequently uncover their LAD profiles. In conclusion, in order to understand the role of LADs on transcription and the acquisition of cell fates, there is a need for the development of an integrated technique that allows the measurement of LADs and transcription in single cells of a differentiation trajectory.

REFERENCES

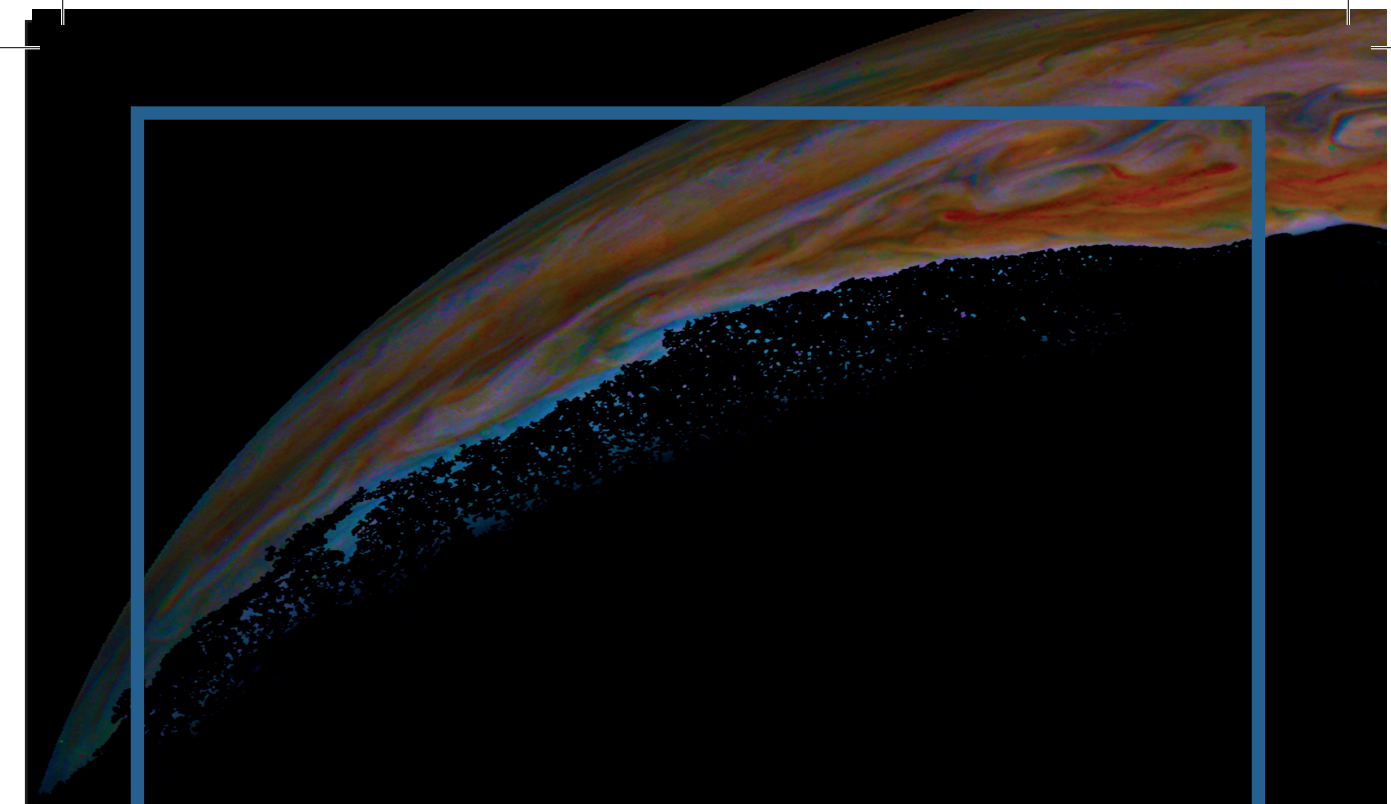
1. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389, 251-260 (1997).
2. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. Cell Res 21, 381-395 (2011).
3. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33, 729-740 (1983).
4. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272-286 (2014).
5. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376-380 (2012).
6. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. Sci Adv 5, eaaw1668 (2019).
7. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289-293 (2009).
8. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38, 1348-1354 (2006).
9. 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).
10. Pickersgill, H. et al. Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat Genet 38, 1005-1014 (2006).
11. Kind, J. et al. Single-cell dynamics of genome-nuclear lamina interactions. Cell 153, 178-192 (2013).
12. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134-147 (2015).
13. Ikegami, K., Egelhofer, T.A., Strome, S. & Lieb, J.D. Caenorhabditis elegans chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. Genome Biol 11, R120 (2010).
14. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948-951 (2008).
15. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell 38, 603-613 (2010).
16. Borsos, M. et al. Genome-lamina interactions are established de novo in the early mouse embryo. Nature (2019).
17. van Bemmel, J.G. et al. The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. PLoS One 5, e15013 (2010).
18. 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).
19. 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).
20. 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).
21. Pope, B.D. et al. Topologically associating domains are stable units of replication-timing regulation. Nature 515, 402-405 (2014).
22. Leemans, C. et al. Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. Cell 177, 852-864 e814 (2019).
23. Akhtar, W. et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell 154, 914-927 (2013).
24. Zheng, X., Kim, Y. & Zheng, Y. Identification of lamin B-regulated chromatin regions based on chromatin landscapes. Mol Biol Cell 26, 2685-2697 (2015).
25. Wu, F. & Yao, J. Identifying Novel Transcriptional and Epigenetic Features of Nuclear Lamina-associated Genes. Sci Rep 7, 100 (2017).
26. Finlan, L.E. et al. Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet 4, e1000039 (2008).
27. Kumaran, R.I. & Spector, D.L. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. J Cell Biol 180, 51-65 (2008).
28. 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).
29. Dialynas, G., Speese, S., Budnik, V., Geyer, P.K. & Wallrath, L.L. The role of Drosophila Lamin C in muscle function and gene expression. Development 137, 3067-3077 (2010).
30. Zullo, J.M. et al. DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell 149, 1474-1487 (2012).
31. Bian, Q., Khanna, N., Alvikas, J. & Belmont, A.S. beta-Globin cis-elements determine differential nuclear targeting through epigenetic modifications. J Cell Biol 203, 767-783 (2013).
32. Towbin, B.D. et al. Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. Cell 150, 934-947 (2012).

33. Therizols, P. et al. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science 346, 1238-1242 (2014).
34. Zuo, B. et al. Influences of lamin A levels on induction of pluripotent stem cells. Biol Open 1, 1118-1127 (2012).
35. Shevelyov, Y.Y. et al. The B-type lamin is required for somatic repression of testis-specific gene clusters. Proc Natl Acad Sci U S A 106, 3282-3287 (2009).
36. Kohwi, M., Lupton, J.R., Lai, S.L., Miller, M.R. & Doe, C.Q. Developmentally regulated subnuclear genome reorganization restricts neural progenitor competence in Drosophila. Cell 152, 97-108 (2013).
37. Gonzalez-Sandoval, A. et al. Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. Cell 163, 1333-1347 (2015).
38. Chen, C.K. et al. Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. Science 354, 468-472 (2016).
39. Bonev, B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557-572 e524 (2017).
40. Ahmed, K. et al. Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. PLoS One 5, e10531 (2010).
41. Ugarte, F. et al. Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells. Stem Cell Reports 5, 728-740 (2015).
42. Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res 20, 155-169 (2010).
43. Robson, M.I. et al. Tissue-Specific Gene Repositioning by Muscle Nuclear Membrane Proteins Enhances Repression of Critical Developmental Genes during Myogenesis. Mol Cell 62, 834-847 (2016).
44. Poleshko, A. et al. Genome-Nuclear Lamina Interactions Regulate Cardiac Stem Cell Lineage Restriction. Cell 171, 573-587 e514 (2017).
45. See, K. et al. Lineage-specific reorganization of nuclear peripheral heterochromatin and H3K9me2 domains. Development 146 (2019).
46. Wiblin, A.E., Cui, W., Clark, A.J. & Bickmore, W.A. Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. J Cell Sci 118, 3861-3868 (2005).
47. Lund, E. et al. Lamin A/C-promoter interactions specify chromatin state-dependent transcription outcomes. Genome Res 23, 1580-1589 (2013).
48. Meister, P., Towbin, B.D., Pike, B.L., Ponti, A. & Gasser, S.M. The spatial dynamics of tissue-specific promoters during C. elegans development. Genes Dev 24, 766-782 (2010).
49. 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).
50. Gutierrez, G.D., Gromada, J. & Sussel, L. Heterogeneity of the Pancreatic Beta Cell. Front Genet 8, 22 (2017).
51. Stumpf, P.S. et al. Stem Cell Differentiation as a Non-Markov Stochastic Process. Cell Syst 5, 268-282 e267 (2017).
52. Grun, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251-255 (2015).
53. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138-1142 (2015).
54. Stevens, T.J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59-64 (2017).
55. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486-490 (2015).
56. van Koningsbruggen, S. et al. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol Biol Cell 21, 3735-3748 (2010).
57. Schlesinger, S. & Meshorer, E. Open Chromatin, Epigenetic Plasticity, and Nuclear Organization in Pluripotency. Dev Cell 48, 135-150 (2019).

Thesis outline

The projects described in this thesis concentrate on the need for technological advancements in the single-cell genomics field to address longstanding questions of the nuclear organization field from a LAD perspective. I present a novel tool to investigate nuclear organization and transcription of the same cell (**Chapters 2&3**) as well as the application of this method on a in vivo developmental system (**Chapter 4**).

In **Chapter 2** I present a novel method that allows for the simultaneous quantification of transcripts and protein-DNA interactions from the same cell and its application for mapping genome architecture, accessibility and protein-DNA interactions in pluripotent and differentiating stem cells. In **Chapter 3** I describe this new tool in detail and in **Chapter 4** I apply it on the cortex of the developing mouse embryo. Finally, in **Chapter 5** I discuss the relationship between genome organization and transcription, possible technical extensions of the method described in **Chapters 2&3** and additional biological questions awaiting to be addressed.



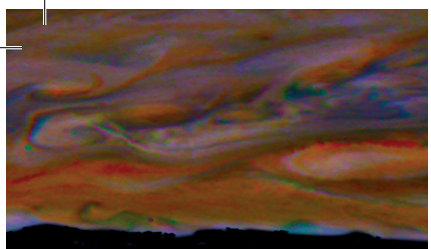
Chapter 2

Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells.

Koos Rooijers*, Corina M. Markodimitraki*, Franka J. Rang, Sandra S. de Vries, Alex Chialastri, Kim L. de Luca, Dylan Mooijman, Siddharth S. Dey# and Jop Kind#

* equal contribution, #co-corresponding

Nature Biotechnology 37, 766–772 (2019)



2

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 mRNA 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 analysis of protein-mediated mechanisms that regulate cell type-specific transcriptional programs in heterogeneous tissues.

INTRODUCTION

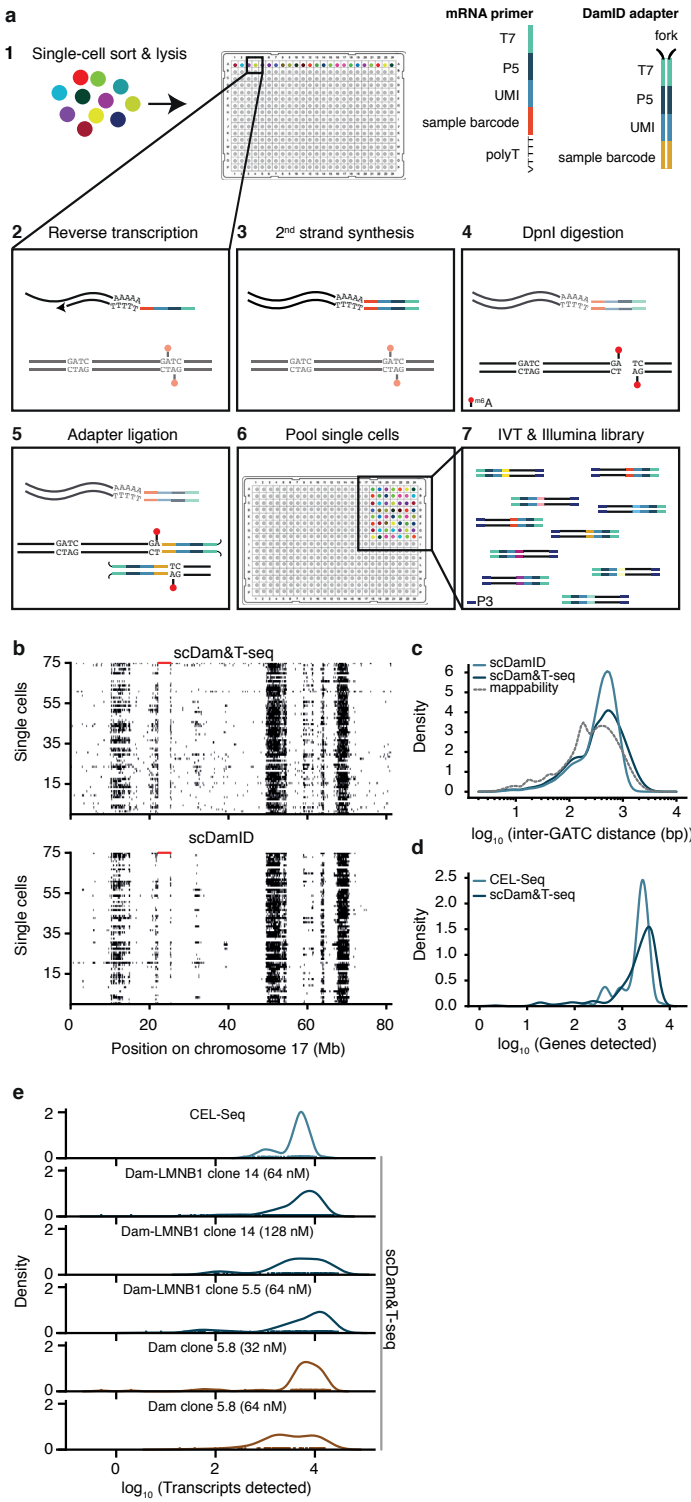
Recent advances in measuring genome architecture (Hi-C and DamID)[1–4], chromatin accessibility (ATAC-seq and DNase-I-seq)[5–7], various DNA modifications[8–13] and histone post-translational modifications (chromatin immunoprecipitation (ChIP)-seq)[14] in single cells have enabled characterization of 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[15–19] 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 har- nesses DamID to map genomic protein localizations together with mRNA sequencing from the same cell.

RESULTS

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').

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 scDamID2 in 75 single cells with highest sequencing depth. 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. **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-Seq2 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.



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 Supplementary Fig. **1a**). 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 (Supplementary Fig. **1b**). In addition, scDam&T-seq and scDamID are similarly enriched on LADs in HT1080 cells[24] (Supplementary Fig. **1c**) and run-length analysis show similar prevalence of long stretches of genome–NL contacts in single cells (Supplementary Fig. **1d**). 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 (Supplementary Fig. **1e**). 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 (Supplementary Fig. **1f**). 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 (Supplementary Fig. **1f**). Besides increased throughput, linear amplification of the DamID-products reduced the loss of reads resulting from incorrect adapter sequences (Supplementary Fig. **1g**) 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=2508.5, scDam&T-seq=2282.5) (Fig.**1d**) and unique transcripts (median: CEL-Seq=4920, scDam&T-seq=4009.5) (Supplementary Fig. **2a**). Transcriptomes measured by scDam&T-seq and CEL-Seq show a high degree of correlation (Supplementary Fig. **2b**, left panel) and display comparable single-cell variations indicated by the fraction of cells with detected genes (Supplementary Fig. **2c**, left panel), as well as by the relationship between mean gene expression and the coefficient of variation (Supplementary Fig. **2d**). These correlations are similar when comparing independent scDam&T-seq libraries (Supplementary Fig. **2b**, **c**, right panels). We observe batch effects between clones, libraries and methods (Supplementary Fig. **2e**). 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 Supplementary Fig. **2f**, **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 condi

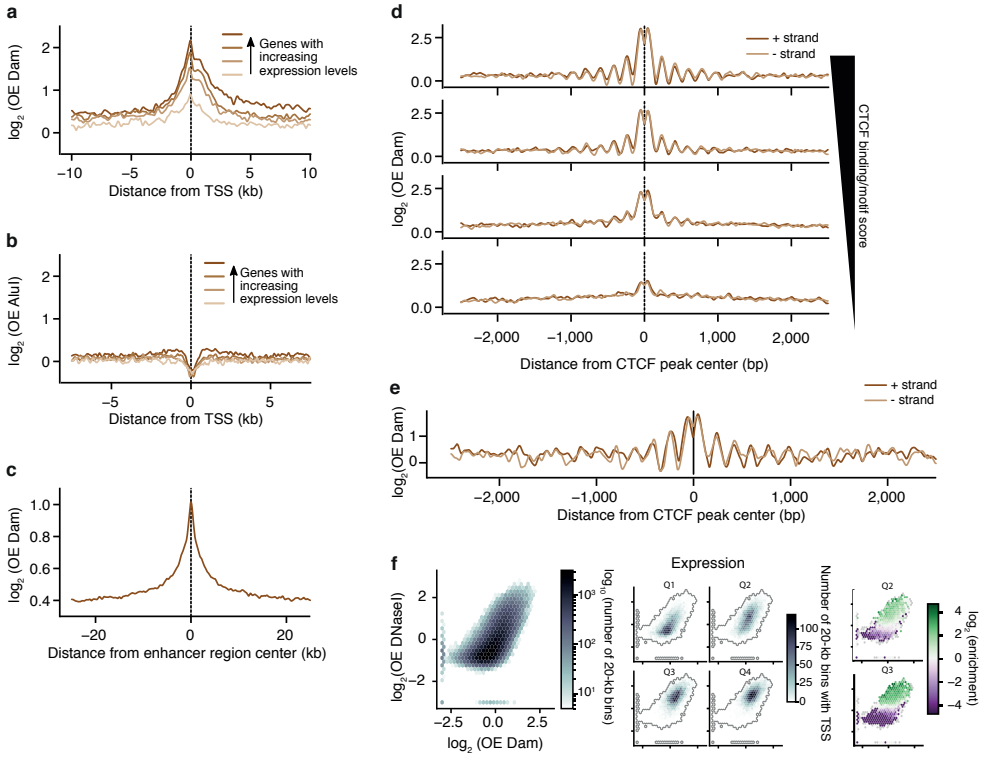


Figure 2 • Untethered Dam marks accessible chromatin in single cells a. Log-transformed OE values (log_2OE) of the Dam signal from an in silico population sample on TSSs of genes grouped into four equal-sized categories with increasing expression levels (ordered light to dark). **b.** log_2OE values 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.

tion (Supplementary Fig. **2h**). Since lowering the double-stranded adapter concentrations does not affect DamID complexity (Supplementary Fig. **1f**), 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] (Supplementary Fig. **3a**). 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 (Supplementary Fig. **3b**).

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] (Supplementary Fig. **4a**). 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.

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) (Supplementary Fig. **4b**), 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 (Supplementary Fig. **4c**). The numbers of detected genes are also very similar between methods at com- parable sequencing depths (Supplementary Fig. **4d**). 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

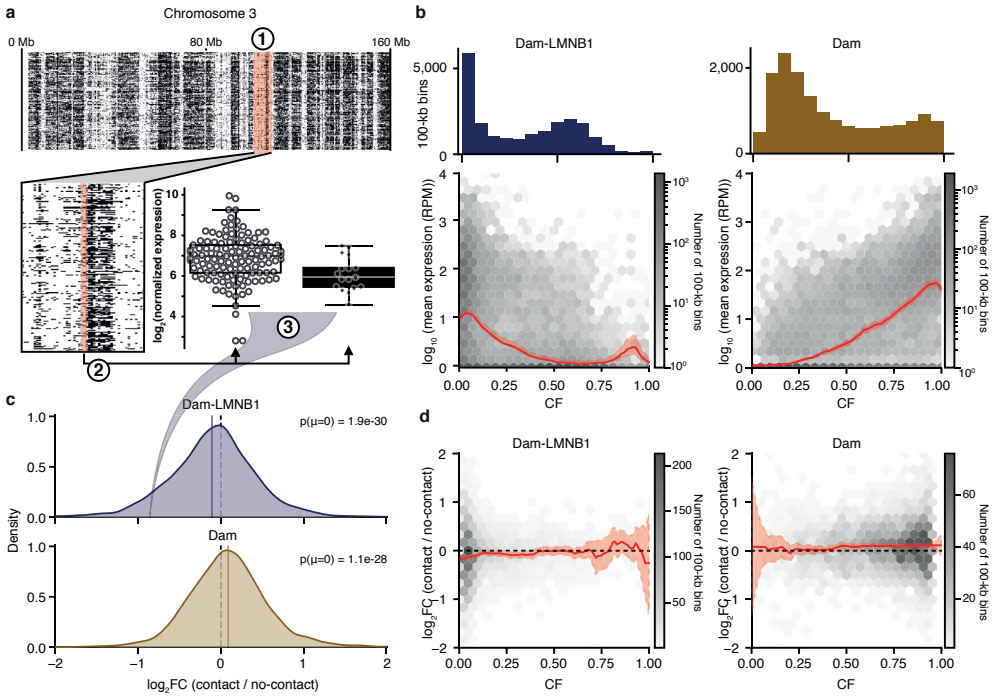


Figure 3 • Parallel transcriptomic and DamiD measurements link transcriptional dependencies to heterogeneity in DamiD contacts. a. Schematic of analysis to determine the log_2FC in transcription between contact and no-contact states. (1) Per genomic bin (of 100 kb), the single-cell samples are binarized 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_2FC 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_2FC) compared to no Dam-LMNB1 contact. The example bin displays NL contacts in 17 out of 143 single cells and log_2FC 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_2FC 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_2FC 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.

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_2FC 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 (Supplementary Fig. **5a–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 (Supplementary Fig. **5e**, 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 (Supplementary Fig. **5f**, top). The opposite patterns can be observed for the Dam contact regions (Supplementary Fig. **5e, 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 (Supplementary Fig. **6a**). Chromosomes 5, 8 and 12 were found frequently (partially) duplicated or lost and were excluded from our analyses (Supplementary Fig. **6a**). 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 (Supplementary Fig. **6b**). 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 (Supplementary Fig. **6c**). 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 (Supplementary Fig. **6d**). Therefore, scDam&T-seq can be employed to investigate single-cell allelic relationships between expression and chromatin states.

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 (Supplementary Fig. **7a**). 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

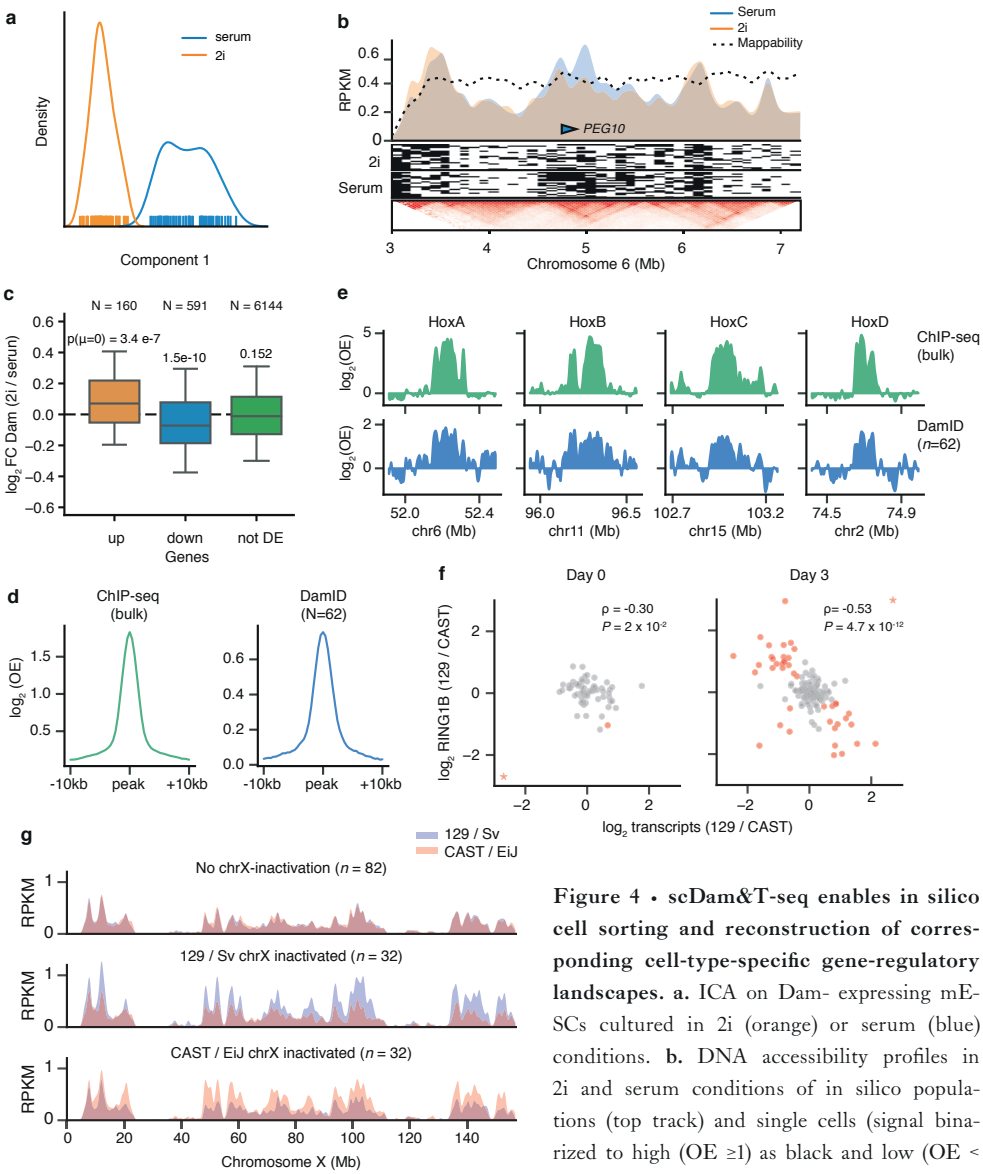


Figure 4 • scDam&T-seq enables in silico cell sorting and reconstruction of corresponding cell-type-specific gene-regulatory landscapes. a. ICA on Dam- expressing mESCs cultured in 2i (orange) or serum (blue) conditions. **b.** DNA accessibility profiles in 2i and serum conditions of in silico populations (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_2OE signal over all RING1B ChIP-seq peaks obtained with ChIP-seq (left) and scDam&T-seq (right) in 2-kb bins. **e.** Signal (log_2OE) 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. *(figure legend continued on next page)*

(figure legend continued) **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.

randomized domains of the same size (Supplementary Fig. 7**b**). 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 Supplementary Fig. 7**c**). Interestingly, the increased accessibility at the TSS extends into the gene body (Supplementary Fig. 7**d**). The same increased accessibility is also observed in single cells for the top five differentially expressed genes between conditions (Supplementary Fig. 7**e**). 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.

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 com- parable (Fig. **4e**) and genome-wide scDam&T-seq and ChIP-seq correlate well (Supplementary Fig. 7**f**). 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 (Supplementary Fig. 7**g**). 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 (Supplementary Fig. 7**h**). 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] (Supplementary Fig. 7**i**). 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.

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 beta-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 beta-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, 00000000 3115836001)) 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 $\mu g\ \mu l^{-1}$. 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 project37 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 and Supplementary Figs. 6 and 7g,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

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

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), TO and TE 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 Supplementary Fig. 2a. 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 (Supplementary Fig. 1c for human data and Supplementary Fig. 3b 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 (Supplementary Fig. 2d). 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

Transcriptomics 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 Supplementary Fig 4d. GpC accessibility data from scNMT-seq were obtained from the processed data of GSE109262.

log_2FC between contact/no-contact groups of samples

log_2FCs 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_2FC 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_2FC 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 Supplementary Fig. 5a) 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). Supplementary Fig. 5b 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 (see Supplementary Fig. 5a), raw VMR values show very similar results. The rolling mean and confidence interval in Supplementary Fig. 5b 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 egdeR45, 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 Supplementary Fig. 7c, 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_2FC 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.

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 Supplementary Fig. 4d 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.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/ s41587-019-0150-y.

REFERENCES

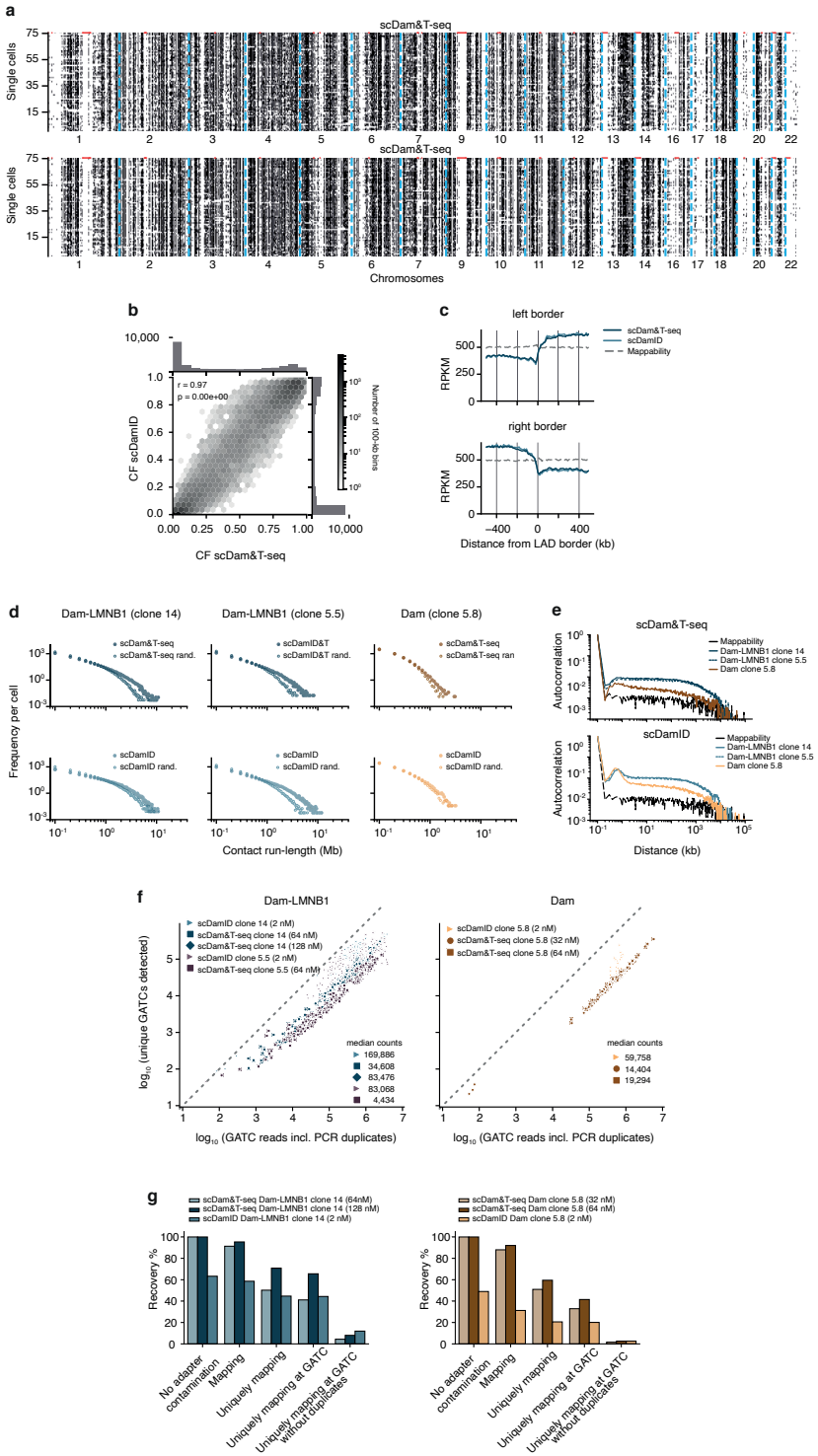
1. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64 (2013).
2. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134–147 (2015).
3. Flyamer, I. M. et al. Single-nucleus Hi-C reveals unique chromatin 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).
36. Sakaue-Sawano, A. et al. Visualizing spatiotemporal dynamics of multicellular cell cycle progression. Cell 132, 487–498 (2008).

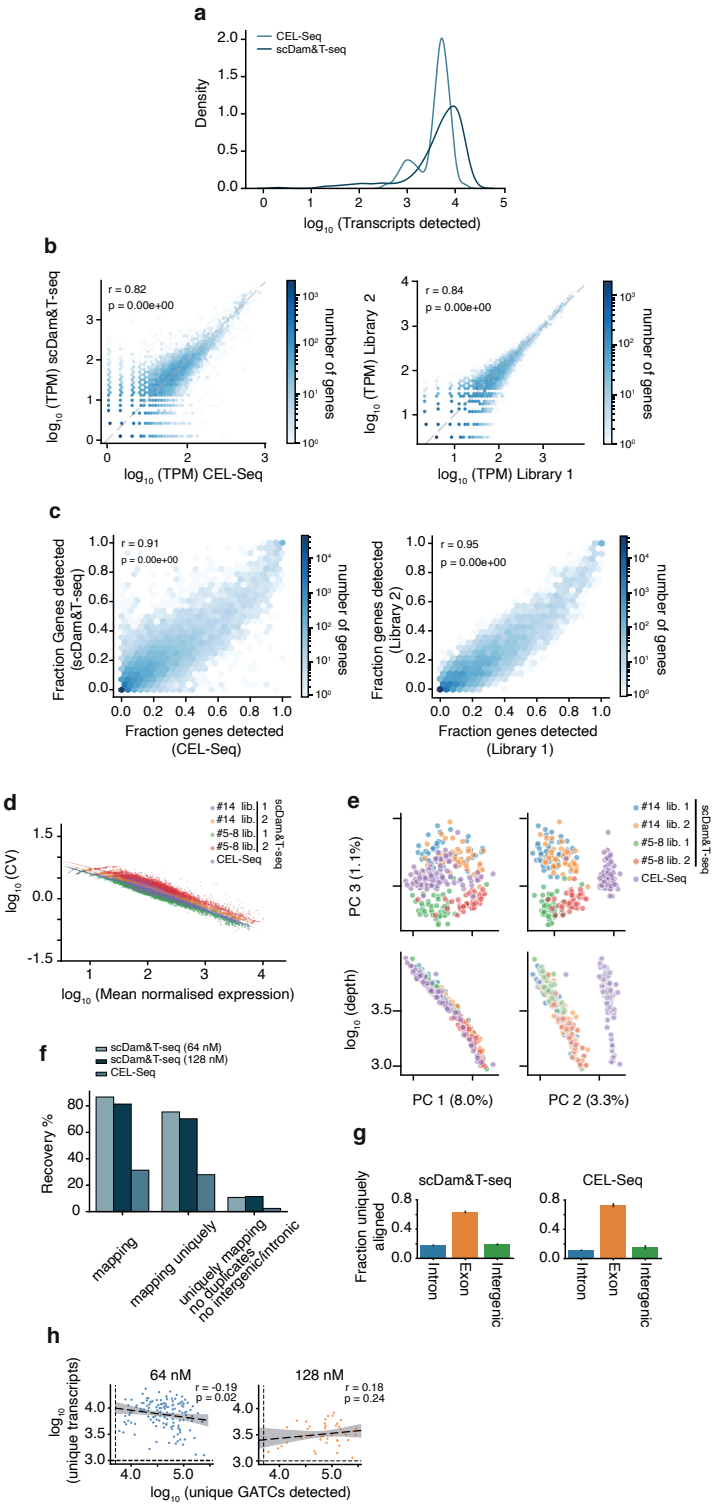
37. Keane, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289–294 (2011).
38. Lun, A. T. et al. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 17, 75 (2016).
39. Lun, A. T. et al. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res 5, 2122 (2016).
40. Boedigheimer, M. J. et al. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics 9, 285 (2008).
41. Johnson, W. E. et al. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 1, 118–127 (2006).
42. Kent, W. J. et al. The human genome browser at UCSC. Genome Res. 12, 996–1006 (2002).
43. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).
44. Knijnenburg, T. A. et al. Multiscale representation of genomic signals. Nat. Methods 11, 689–694 (2014).
45. Robinson, M. D. et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).

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, i.e. 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 (Genome Research 23, 270-281, 2013). **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.

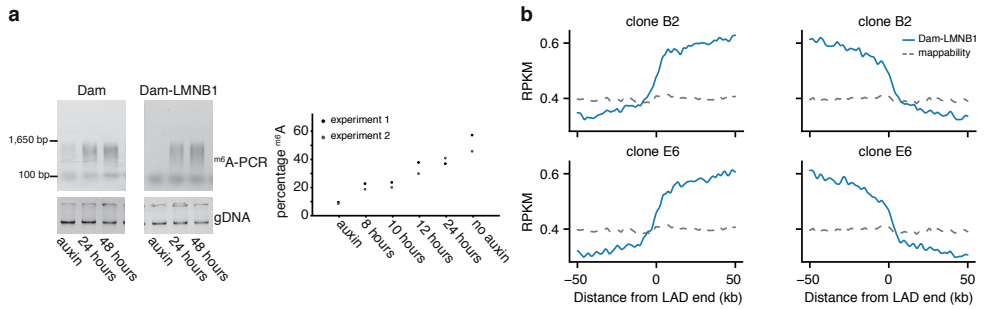
SUPPLEMENTAL FIGURES & TABLES





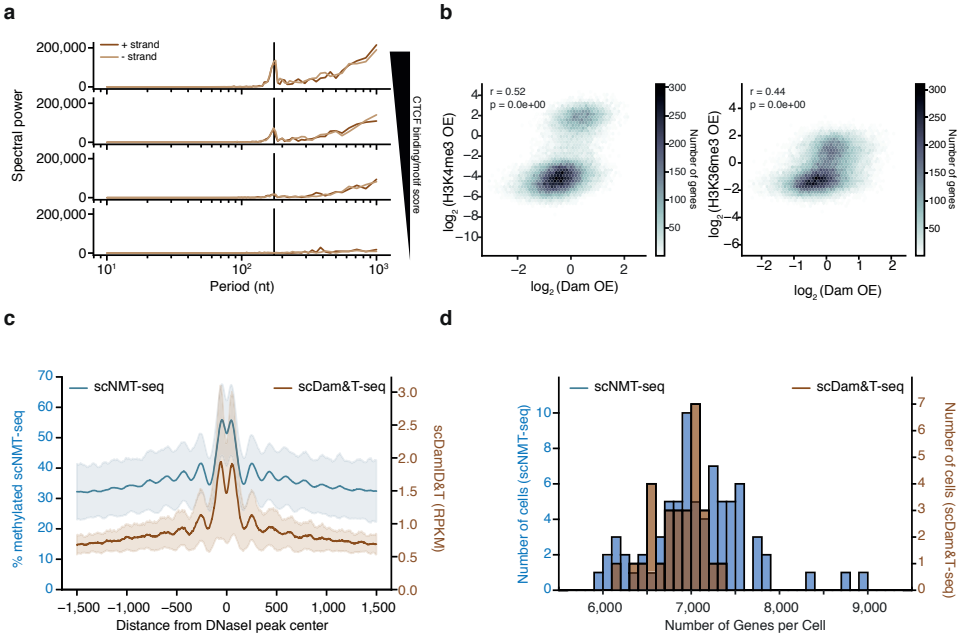
Supplementary figure 2 • Quantitative comparison between CEL-Seq2 and scDam&T-seq

a. Distributions of the number of unique transcripts detected using CEL-Seq (Cell 163, 134-147, 2015) 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).



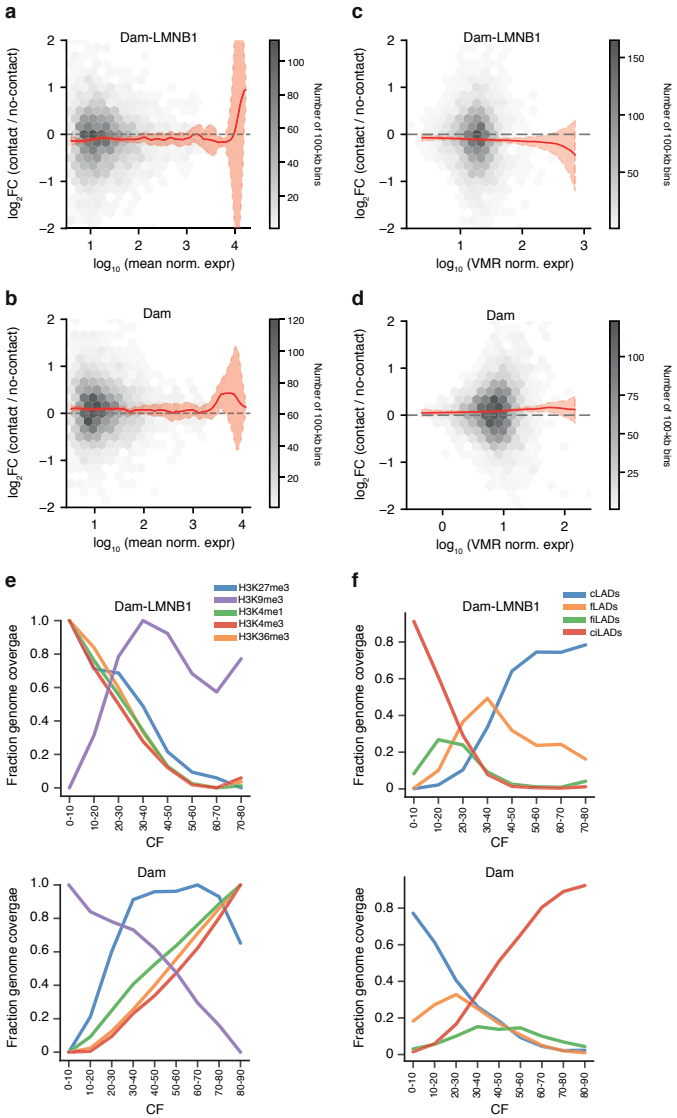
Supplementary figure 3 • scDam&T-seq in hybrid mESCs

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 (Cell 153, 178-192, 2013). 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 (Molecular cell 38, 603-613, 2010), for two different Dam-LMNB1 clones (#b2 and #e6) and a total of 166 single-cell samples.

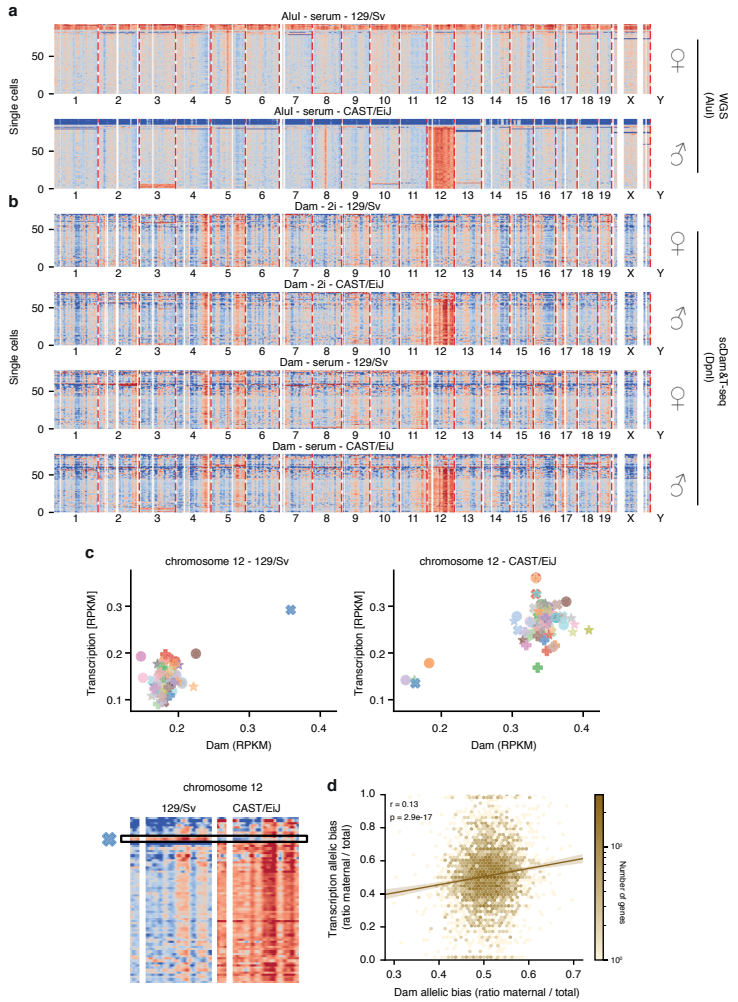


Supplementary figure 4 • Untethered Dam enzyme marks accessible chromatin in single cells

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).

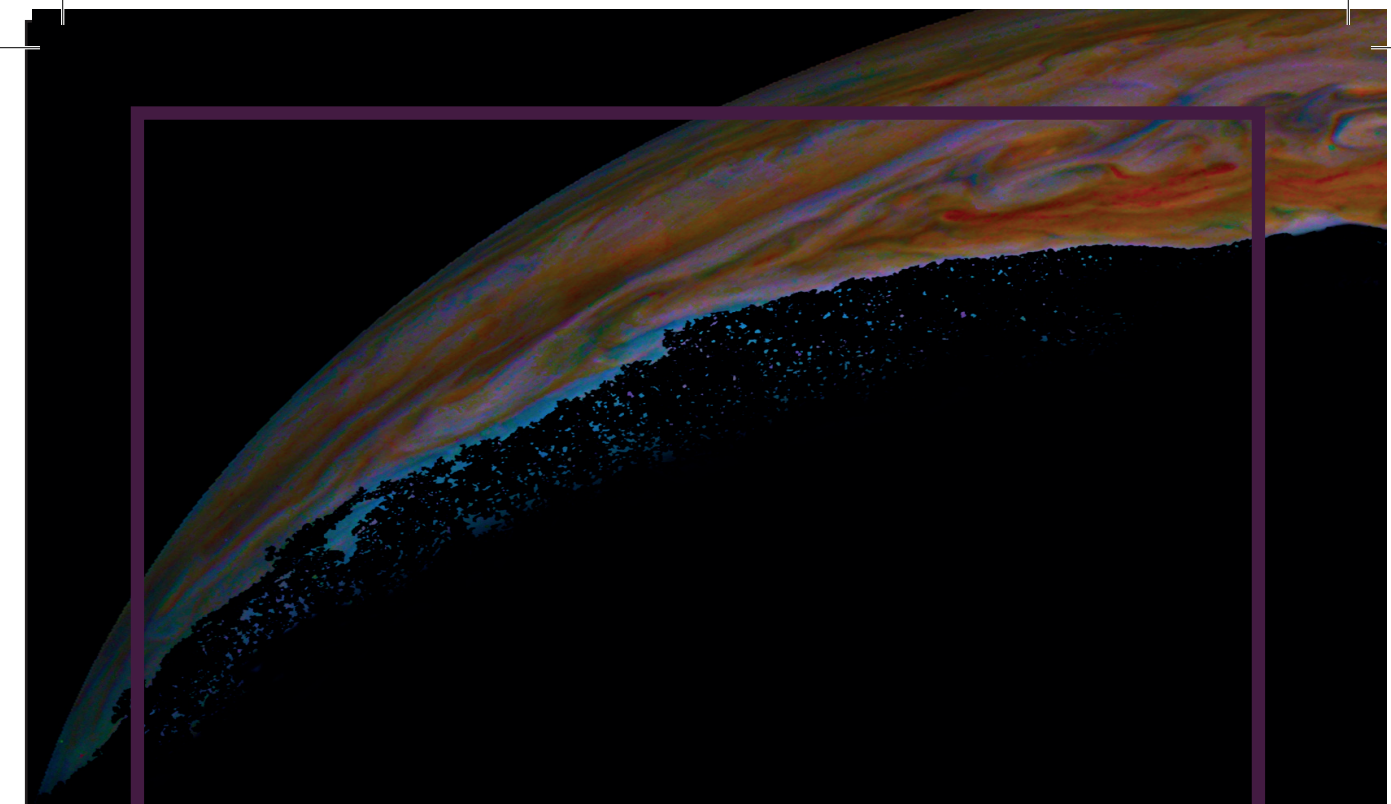
Supplementary Table 1 • Complexity of scDam&T-seq libraries

						Unique DamID events			Unique transcripts			Samples		
KBM7						Q1	median	Q3	Q1	median	Q3	Total	Passing cutoffs	Passing cutoffs (%)
restriction_enzyme	fusion_construct	clone_id	phase	condition	damid2_adapter_concentration_nM									
DpnI	Dam_LmnB1	clone_1A	G1_S	N/A	64	9368,5	34608	67011	4416,5	7979,5	11275,5	192	141	73,4375
				N/A	128	42878	83476	138439	1925	3014,5	4479,5	48	43	89,58333333
		clone_5.5	G1_S	N/A	64	1301,5	4434	24378,5	5005,5	11190	15504	144	56	38,88888889
	Dam_only	clone_5.8	G1_S	N/A	32	6930,5	14404	34133	4033,5	5311	6796,5	48	38	79,16666667
				N/A	64	7659,5	19294,5	39884,5	1111,5	1749,5	2353	48	37	77,08333333
						Unique DamID events			Unique transcripts			Samples		
mESC						Q1	median	Q3	Q1	median	Q3	Total	Passing cutoffs	Passing cutoffs (%)
restriction_enzyme	fusion_construct	clone_id	phase	condition	damid2_adapter_concentration_nM									
DpnI	Dam_LmnB1	clone_B2	G2_M	serum	64	5519,5	24625	77978	2110,5	4473	7545	144	88	61,11111111
			mid5	serum	64	452	4154,5	23030	5778	12823,5	22582,5	48	21	43,75
		clone_E6	G2_M	serum	64	4289	22126	68768	2106	6910,5	10845	144	76	52,77777778
	Dam_Ring1B	clone_1B	Unsorted	differentiation_day3	32	6992,5	16255	33054	10904,5	17230	28475,5	184	145	78,80434783
				serum	32	3264	9633,5	17423,5	6501,5	9520,5	12049,5	92	62	67,39130435
	Dam_only	clone_C8	Unsorted	differentiation_day3	32	1835	14395,5	39982	5357,5	9276,5	12856,5	92	62	67,39130435
			mid5	2i	64	15011	61792	131903	2506,5	4167	5598	96	71	73,95833333
				serum	64	1948	35304,5	93598	7041	9793,5	13968	96	61	63,54166667

Supplementary Table 4 • Statistical tests

Figure	test	test statis	p value	n	df
FigS1b	Pearson correlation	657,01	0	24168	24166
FigS2b (left)	Pearson correlation	201,118	0	20276	20274
FigS2c (left)	Pearson correlation	531,72	0	60028	60026
FigS2b (right)	Pearson correlation	204,334	0	17132	17130
FigS2c (right)	Pearson correlation	721,246	0	60028	60026
FigS2h (top; 64nM)	Pearson correlation	-2,29452	0,02326	141	139
FigS2h (bottom; 128nM)	Pearson correlation	1,20471	0,23522	43	41
FigS4b (left)	Pearson correlation	146,18	0	56621	56619
FigS4b (right)	Pearson correlation	116,207	0	57238	57236
Fig3c (Dam-LMNB1)	One-sample t-test	-11,5798	1,9E-30	3497	3496
Fig3c (Dam)	One-sample t-test	11,2099	1,1E-28	3668	3667
FigS6d	Pearson correlation	8,48884	2,89E-17	4051	4049
Fig4c (DE up)	One-sample t-test	5,3248	3,4E-07	158	157
Fig4c (DE down)	One-sample t-test	-6,52035	1,5E-10	577	576
Fig4c (not DE)	One-sample t-test	-1,43223	0,15213	6056	6055
Fig4f (left, serum)	Spearman correlation	NA	0.0196254	61	59
Fig4f (right, day3)	Spearman correlation	NA	4,7E-12	146	144
FigS7c	Pearson correlation	15,3439	1,3E-52	11221	11219
FigS7f	Pearson correlation	NA	0	22486	22484

Supplementary Table **2** containing the DamID adapter sequences and Supplementary Table **3** containing the CELseq2 primer sequences can be found in the online supplementary materials: www.nature.com/articles/s41587-019-0150-y#Sec35



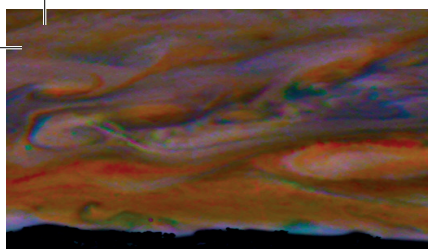
Chapter 3

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

Corina M. Markodimitraki*, Franka J. Rang*, Koos Rooijers, Sandra S. de Vries, Alex Chialastri, Kim L. de Luca, Silke J. A. Lochs, Dylan Mooijman, Siddharth S. Dey# and Jop Kind#

* equal contribution, #co-corresponding

Nature Protocols 15, 1922–1953 (2020)



3

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

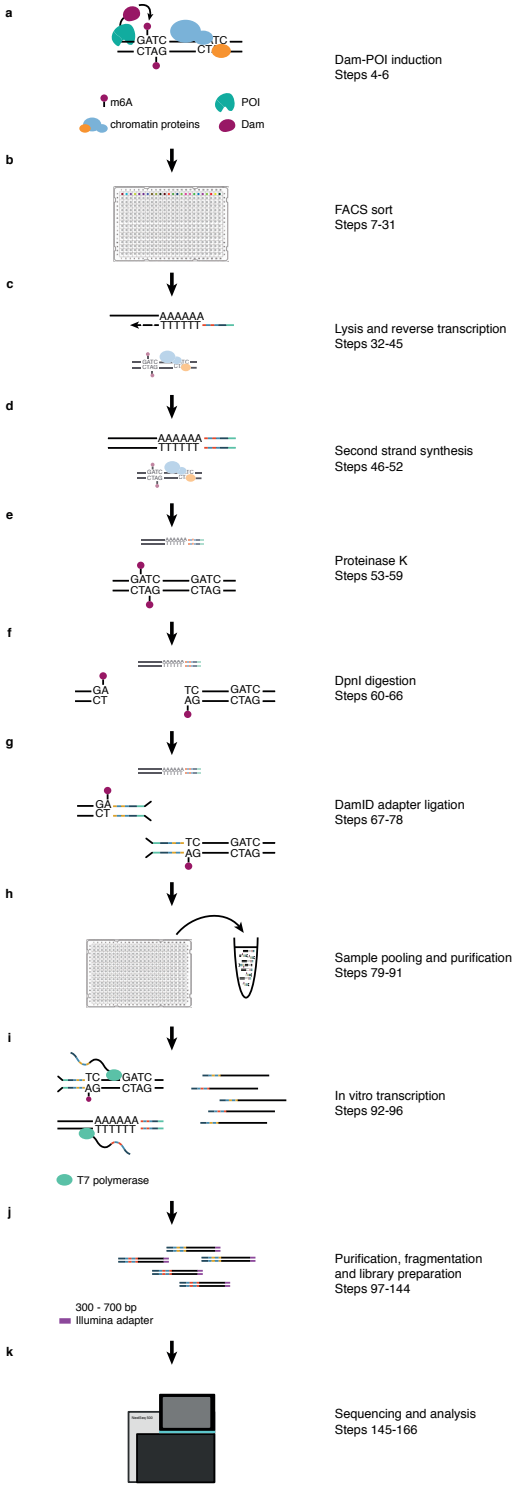
A myriad of proteins cooperate to establish cell type-specific chromatin architecture and gene expression through their contact with DNA. Such proteins range from post-translationally modified histones to transcription factors, from nuclear lamina (NL) constituents to the transcriptional machinery. Methods to measure protein-DNA interactions (ChIP-seq[1] and DamID[2] or their effect on chromatin organisation (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[5-8], chromatin accessibility[9-11], DNA modifications[12-17] and protein-DNA interactions[18-21]. 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[22-24]. 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 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 in order 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].

Figure 1 • Overview of the method

a–k. Panels describe the main parts of the protocol ('Overview of the protocol'). 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.



Overview of the protocol

Prior to 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, 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. 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.

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.

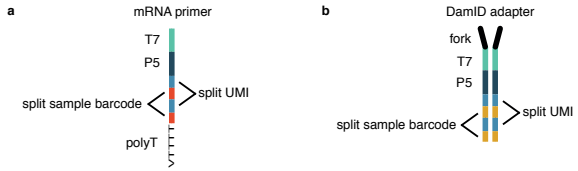


Figure 2 • CEL-Seq2 primer and DamID adapter structure
a. CEL-Seq2 primer. **b.** DamID adapter ('Overview of the protocol' and 'Experimental design: Design and concentration of DamID adapters and CEL-Seq2 primers').

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 only 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[18–21,33,34]. 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 only contain 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 users to 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 system31) 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 clones2. 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 construct28. 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 ≤ 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.

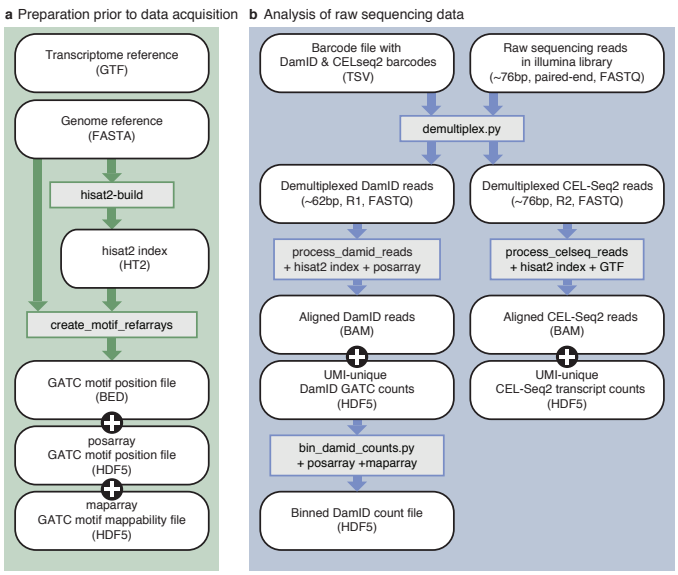


Figure 3 • Bioinformatics workflow

a. Preparation of reference files, which needs to be performed only 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; rectangular boxes show programs and necessary reference files. Arrows indicate which files are used as input for subsequent programs. GTF, Gene Transfer Format; FASTA, Fast-All; HISAT2, Hierarchical Indexing for Spliced Alignment of Transcripts 2; BED, Browser Extensible Data; HDF5, Hierarchical Data Format 5; TSV, tab-separated values; FASTQ, Fast-Quality; BAM, Binary Alignment Map.

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); an 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; an 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. Fur- thermore, 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.

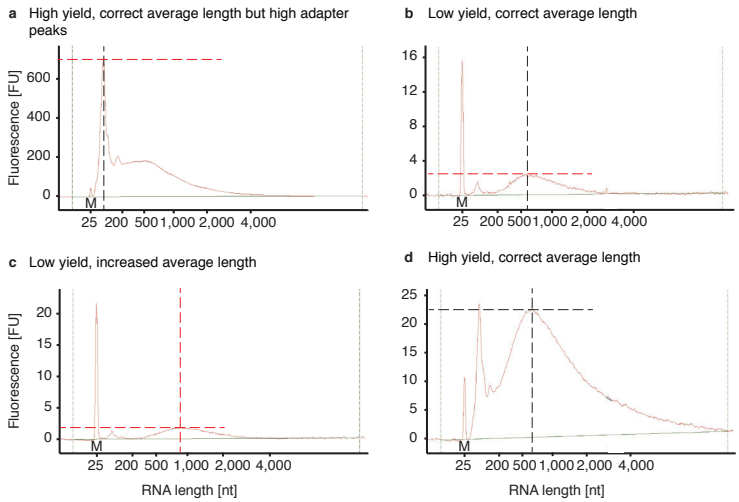


Figure 4 • Examples of aRNA bioanalyzer plots. Bioanalyzer results after IVT, bead purification, aRNA fragmentation and another bead purification.

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.

MATERIALS

Biological materials

Cell lines.

For Figures **3-6** 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 DNAZap is chronically toxic for 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 (100x; Gibco, cat. no. 11140035)
- Penicillin/streptomycin (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)
- beta-mercaptoethanol (1M; Sigma, cat. no. M3148) **!CAUTION** B-mercaptoethanol is 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** Trypan Blue solution 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-Aldrich, 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** Igepal is a skin, mouth and eye irritant. Use with care and wear gloves and mouth mask in case of insufficient ventilation. It is also chronically toxic for aquatic systems. Do not pour it into the 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)
- Ribonuclease H (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** Ethanol is 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)

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'):

  `GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNGATGNNNT-`

  `CATTTTTTTTTTTTTTTTTTTTTTTTV`

  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 UMI that labels a transcript molecule uniquely (here: `NNNNNN`, where '`N`' is `G`, `C`, `A` or `T`). The barcode is split in 2 × 4 nt and alternates in the primer sequence with the UMI, which is split in 2 × 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 500μM 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 the container until room temperature (20–22 °C) is reached. Vortex the plate for 5 s to resuspend adapters and pulse-spin at room temperature for 10 s. Store adapters at −20 °C indefinitely. Use this plate for adapter dispensing.

- Example of one DamID adapter (5' -> 3'):
  Top oligo:

  `GGTGATCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATCNNNTGCANNNTATGGA`

Bottom oligo:

`/5Phos/TCCATANNNTGCANNNGATCGTCGGACTGTAGAACTCTGAACCCCTATAGTGGTCGTATTACCGGGAGCTT`

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 UMI that labels a cut GATC site uniquely (here: `NNNNNN`, where '`N`' is `G`, `C`, `A` or `T`). The barcode is split in 2 × 4 nt and alternates in the primer sequence with the UMI, which is split in 2 × 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

- randomhexRT 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'):

  `CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCC*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'):

  `AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCG*A`

  Critical: The "*" indicates a phosphorothioate bond which protects the DNA from endo- and exonuclease activity therefore increasing the stability of the oligo.

Box 1 • mosquito robot handling

Procedure

Primer plate preparation

Timing: 20 min for one plate (30 min for max four plates)

1. Turn on the computer workstation and the robot and initialize.
2. Use the 'Humidify' function to humidify the plate deck chamber to 80% (wt/vol).
3. Remove the 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.
4. Remove seals and insert the destination plate(s) containing 5 μl of mineral oil in position(s) 2, 3, 4 and 5 with corner A1 facing the upper left corner of the magnetic holder.
5. 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 max four plates)

1. Turn on the computer workstation and the robot and initialize.
2. Use the 'Humidify' function to humidify the plate deck chamber to 80% (wt/vol).
3. Remove the 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.

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% CO2 (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

- **Auxin[42] 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 beta-mercaptoethanol and ESGROmLIF:** In 430 ml of Glasgow's MEM, add 50 ml of FBS, 5 ml of Pen/Strep, 5 ml of GlutaMAX 100x, 5 ml of non-essential amino acids 100x and 5 ml of sodium pyruvate 100 mM. The final concentrations in the solution are 10% (vol/vol) FBS, 100 U/ml Pen/Strep, GlutaMAX 1x, non-essential amino acids 1x, sodium pyruvate 1 mM. Store at 4 °C for ≤ 1 month.
- **mESC complete culture media with beta-mercaptoethanol and ESGROmLIF:** In 50 ml of mESC complete culture media without beta-mercaptoethanol and ESGROmLIF add 5 µl beta-mercaptoethanol 1M and 5 µl ESGROmLIF 10,000,000 U/ml. The final concentrations in the solution are beta-mercaptoethanol 0.1 mM and ESGRO mLIF 1,000 U/ml. Store at 4 °C for ≤ 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 ≤ 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 for ≤ 1 year.
- **2 % (vol/vol) cleaning solution for Nanodrop II robot:** To 49 ml of 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 9,990 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 expiration date indicated and do not freeze-thaw the ERCC RNA Spike-Ins more than eight 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 5 x:** Add 89 ml of nuclease-free water to a sterile glass bottle and add 5 ml of Tris 1 M pH 7.5, 5 ml of 5 M NaCl and 1 ml of 0.5 M EDTA. Final concentrations in the solution are 10 mM Tris, 50 mM NaCl and 1 mM EDTA. Store at room temperature. Prepare new buffer when precipitates start to form.
- **Igepal 1 % (vol/vol):** Add 49.5 ml of nuclease-free water to a 50-ml tube. With a cut tip, take 0.5 ml of Igepal and slowly add it to the water. Resuspend until the tip is empty. Close the tube, invert it a few times and leave it on the rotor until homogenous. Store at room temperature for ≤1 year.
- **dNTP mix (10 mM):** Add 60 μl of nuclease-free water to a tube. Add 10 μl of 10 mM dCTP, 10 μl of 10 mM dGTP, 10 μl of 10 mM dATP and 10 μl of 10 mM dTTP. Make aliquots of 20 μl and store at −20 °C for ≤1 year. Avoid multiple freeze-thaw cycles.
- **Bead buffer:** Dissolve 20 g of PEG 8000 with 48.75 ml of nuclease-free water. Add 1 ml of 1 M Tris-HCl, 0.2 ml of 0.5 M EDTA, 50 ml of 5 M NaCl and 50 μl of Tween 20 pH 8.0. Final concentrations in the solution are 20% (wt/vol) PEG 8000, 10 mM Tris-HCL, 1 mM EDTA, 2.5 M NaCl and 0.05% (vol/vol) Tween 20. Store at room temperature for ≤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 a 15-ml tube and add 8 ml of 100% (vol/vol, absolute) ethanol. Store at room temperature for ≤1 d. 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 for ≤1 year.
- **Fragmentation buffer:** Add 5 ml of nuclease-free water in a tube and add 2 ml of 5 M potassium acetate, 3 ml of 1 M magnesium acetate and 10 ml of 1 M Tris-acetate. Final concentrations in the solution are 500 mM potassium acetate, 150 mM magnesium acetate and 200 mM Tris-acetate. Keep at room temperature for ≤1 year.
- **Stop buffer:** Use 0.5 M EDTA pH 8. Keep at room temperature for ≤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 for ≤1 year.

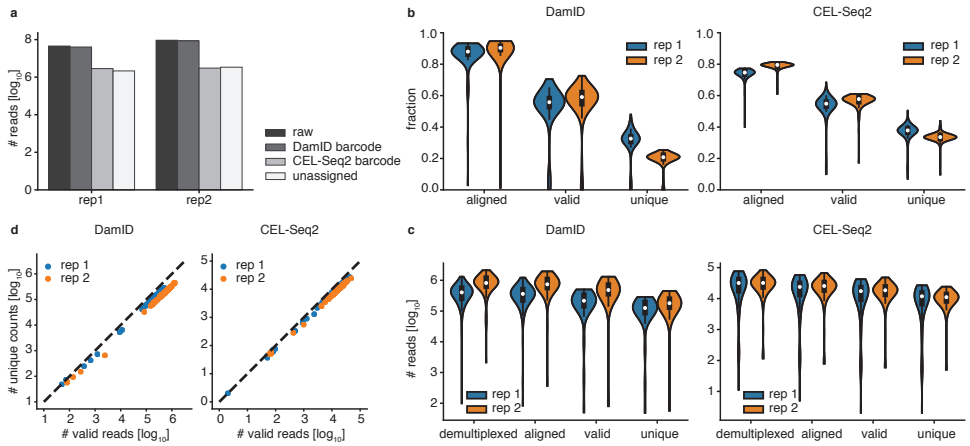


Figure 5 • Technical statistics of a scDam&T-seq run

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** and **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.

Box 2 • Nanodrop II robot handling

Procedure

Master mix dispension

Timing: 20 min for one plate (30 min for max four plates)

1. Perform a 'Daily clean' program with 2% (vol/vol) cleaning solution.
2. Prepare a 'mock' eight-well PCR strip containing nuclear-free water at the same volumes 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 the 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 the 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 54-60) perform a 'Daily clean' program with 2% (vol/vol) cleaning solution to remove excess Proteinase K from the tubing systems to avoid contamination in the next dispensing steps.

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, 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 s to resuspend primers and pulse-spin at 4 °C for 10 s.

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 Fig. **5**, mESCs expressing Dam-LmnB1 were cultured for 2 d on feeder cells till 80% colony confluency, and the Dam-LmnB1 construct was expressed for 6 h before collection.

4 • Pre-warm cell culture medium and PBS at 37 °C for 30 min.

5 • Remove IAA^{42}-containing medium from cells and wash 3 times 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 x h after induction and prepare a single-cell suspension. For the mESCs presented in Fig. **5** induction was 6 h before harvest.

8 • Clean the 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 three to four times.

11 • Pipette 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 10x 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. Pipette 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 Fig. **5**, induction cells were stained with 30 μg/ml Hoechst solution.

Critical step: The final concentration of the Hoechst solution 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 > 3 h.

FACS sorting

Timing: 1 h for three plates

21 • Thaw the primer plate from Step 3 on ice.

22 • Centrifuge plate at 2,000g 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 the plate and the tube with cell suspension onto the 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 alive and in the 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 2,000g 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.

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.

33 • 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 one well (nl)	Amount for one 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	-

34 • In an eight-well PCR strip, prepare aliquots of 11.1 μl of mix per well of the strip. Spin for 3-5 s on a 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.

Remove the sorted plate (Step 31) from -80 ˚C, remove the 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 2,000g 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 2,000g for 1 min at 4 ˚C.

Reverse transcription

Timing: 1 h 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 one well (nl)	Amount for one plate (μl)	Final concentration in mix
Nuclease-free water	10	8.6	-
5x first-strand buffer	70	60.5	2.3x
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 eight-well PCR strip, prepare aliquots of 15.5 μl of mix per well of the strip. Spin for 3-5

s on a 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 2,000g for 1 min at 4 °C.

44 • Put the plate in a thermocycler at 42 °C for 1 h, 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 2,000g 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 one well (nl)	Amount for one plate (µl)	Final concentration in mix
Nuclease-free water	1,347.5	569.3	-
5x second-strand buffer	437.5	184.8	1.1x
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 eight-well PCR strip, prepare aliquots of 101 µl of mix per well of the strip. Spin for 3-5 s on a 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 2,270 nl.

50 • Seal the plate and centrifuge at 2,000g for 1 min at 4 °C.

51 • Put the plate in a thermocycler at 16 °C for 2 h with the lid at 100 °C and let the thermocycler go to 4 °C at the end of the program.

52 • Centrifuge at 2,000g for 1 min at 4 °C.

Proteinase K treatment

Timing: 11 h (overnight reaction)

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

55 • In an eight-well PCR strip, prepare aliquots of 30.3 µl of mix per well of the strip. Spin for 3-5 s on a 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 2,770 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 2,000g for 1 min at 4 °C.

58 • Put the plate in a thermocycler at 50 °C for 10 h, 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 2,000g 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 one well (nl)	Amount for one plate (µl)	Final concentration in mix
Nuclease-free water	177	108.5	-
10x CutSmart buffer	23	14.1	1x
DpnI (20 U/µl)	30	18.4	0.4 U/µl
Total volume	230	141	-

62 • In an eight-well PCR strip, prepare aliquots of 17.4 µl of mix per well of the strip. Spin for 3-5 s on a 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 3,000 nl.

64 • Seal the plate and centrifuge at 2,000g for 1 min at 4 °C

65 • Put the plate in a thermocycler at 37 °C for 6 h, 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 2,000g 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 3,050 nl.

Critical step: The final concentration of DamID adapter needs to be optimized for the expression of Dam-POI construct of choice. We recommend that the final concentration fall within the 1.25 – 100 nM range.

70 • Remove the plate and the adapter plate from the robot and seal.

71 • Centrifuge at 2,000g 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 one well (nl)	Amount for one 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 eight-well PCR strip, prepare aliquots of 28 μl of mix per well of the strip. Spin for 3-5 s on a 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 3,500 nl.

76 • Seal the plate and centrifuge at 2,000g for 1 min at 4 °C.

77 • Put the plate in a thermocycler at 16 °C for 12 h, 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 2,000g for 1 min at 4 °C.

Pause point: The processed plate can be kept at -20 °C for ≤ 1 month.

(Day 4) Pool cells

Timing: 1 h

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 h for one 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 200g 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, which is the bottom phase, containing the barcoded material. Keep on ice.

82 • Repeat Step 81 and transfer aqueous phase to a new tube to remove mineral oil completely.

Purification of barcoded material

Timing: 1 h

Critical: Depending on the number of pools, bead purifications can be a bottleneck. We therefore do not recommend cleaning more than eight 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 >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 ~30 µl of AMPure XP beads in the final mix, to enable the water elution in Step 91.

83 • Equilibrate diluted AMPure XP beads to room temperature for 30 min. Vortex untill the bead-buffer mix is homogenous.

84 • Add 0.8 volumes 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 the AMPure XP beads.

87 • Add 500 µl of fresh 80% (vol/vol) ethanol and leave for 30 s.

88 • Remove the ethanol carefully without disturbing the AMPure XP beads.

89 • Repeat steps 87 and 88. Pulse-spin the tube and place it in a magnetic rack to remove excess ethanol.

?Troubleshooting

90 • Let the AMPure XP beads air-dry for 5 min or until they appear 'matte'.

Critical step: Do not let the 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 the beads and water form a homogenous mix. Place the 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. The 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 the thermocycler go to 4 °C after the program is finished.

Critical step: Do not let the reaction stay in the thermocycler for more than a few hours after the programs is finished, to maintain RNA integrity.

Pause point: The reaction can be kept at -20 °C for ≤ 1 d.

(Day 5) Purification of aRNA

Timing: 1 h

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 until the bead-buffer mix is homogenous.

100 • Add 0.8 volumes of undiluted AMPure XP beads to the reaction and allow the material to bind to the AMPure XP beads for 10 min. the 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 the AMPure XP beads to accumulate. Keep samples on magnetic rack until step 106.

102 • Remove the liquid carefully without disturbing the AMPure XP beads.

103 • Add 500 µl of fresh 80% (vol/vol) ethanol and leave for 30 s.

104 • Remove the ethanol carefully without disturbing AMPure XP beads.

105 • Repeat Steps 103 and 104. Pulse-spin the tube and place it in a magnetic rack to remove excess ethanol.

106 • Let the 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 the magnetic rack and allow the material to elute for 5 min.

108 • Place the tube in a magnetic rack, and without disturbing the AMPure XP beads, carefully transfer 22 μl of 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 from Step 108 while on ice and resuspend. The cumulative volume is 26.4 μl.

111 • Quickly transfer the tube to 94 °C for 2 min.

112 • Remove the tube and quickly put it on ice.

113 • Add 0.1 volumes of fragmentation STOP buffer to the tube as fast as possible and resuspend. Keep on ice. The cumulative volume is 29.04 μl.

?Troubleshooting

Purification and quantification of fragmented aRNA

Timing: 1 h 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 purifirations can increase the library prep efficiency because of adapter depletion. For 384 pooled cells, we recommend at least two rounds of bead purification at this stage of the protocol. For 96 pooled cells, we recommend one round of bead purification, as stated in Steps 114-124.

114 • Equilibrate undiluted AMPure XP beads to room temperature for 30 min. Vortex until the bead-buffer mix is homogenous.

115 • Add 0.8 volumes of undiluted AMPure XP beads to the reaction and allow the material to bind to the AMPure XP beads for 10 min. the 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 the AMPure XP beads to accumulate. Keep samples on magnetic rack until Step 122.

117 • Remove the aqueous phase carefully without disturbing the AMPure XP beads.

118 • Add 500 μl of fresh 80% (vol/vol) ethanol and leave for 30 s.

119 • Remove the ethanol carefully without disturbing the AMPure XP beads.

120 • Repeat Steps 118-119.

121 • Pulse-spin the tube and place in a magnetic rack to remove excess ethanol.

122 • Let the AMPure XP beads air-dry for 5 min or until they appear 'matte'.

Critical step: Do not let the 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 the beads and water form a homogenous mix. Remove the tube from the magnetic rack and allow the material to elute for 5 min.

124 • Place the tube in a magnetic rack, and without disturbing the AMPure XP beads, carefully

transfer 12 μl of solution to a clean tube and place on ice.

Pause point: The sample can be stored at -80 °C for ≤ 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 fragmentation and bead cleanup for library preparation are shown in Fig. 4.

?Troubleshooting

Reverse transcription

Timing: 1 h 30 min

126 • Prepare the randomhexRT mix as indicated in the table below and heat in a thermocycler at 65 °C for 5 min with the 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 h 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	2.5x
DTT (0.1 M)	1	25 mM
RNase OUT (40 U/μl)	0.5	5 U/μl
Superscript II (200 U/μl)	0.5	25 U/μl
Total volume of reaction	10.5	-

PCR indexing

Timing: 30 min

Critical: Do not overamplify the material. We always recommend a minimum of eight cycles of PCR for the generation of enough Illumina-indexed molecules. We recommend ≥ 10 PCR cycles for an aRNA product between 1 and 10 fluorescent units (FU) and 8-9 cycles for aRNA product >10 FU (Fig. 4). Prepare the indexing mix as indicated in the table below. Keep the mix on ice. Index each sample with a unique RNA PCR index 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 primer 1 (10 μM)	2	0.5 μM
RNA PCR index primer (10 μM)	2	0.5 μM
Total volume of reaction	50	-

128 • Run the PCR program in a 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 h 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 until the bead-buffer mix is homogenous.

130 • Add 0.8 volumes of undiluted AMPure XP beads to the reaction and allow the material to bind to the AMPure XP beads for 10 min. The 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 the AMPure XP beads to accumulate. Keep samples on the magnetic rack until Step 138.

132 • Remove the aqueous phase carefully without disturbing the AMPure XP beads.

133 • Add 500 μl of fresh 80 % (vol/vol) ethanol and leave for 30 s.

134 • Remove the ethanol carefully without disturbing the AMPure XP beads.

135 • Repeat Steps 133-134. Pulse-spin the tube and place it in a magnetic rack to remove excess ethanol.

136 • Let the AMPure XP beads air-dry for 5 min or until they appear 'matte'.

Critical step: Do not let the 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 the tube from the magnetic rack and allow the material to elute for 5 min.

138 • Place the tube in a magnetic rack, and without disturbing the AMPure XP beads, carefully transfer 25 μl of the 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 the beads and water form a homogenous mix. Remove the tube from the magnetic rack and allow the material to elute for 5 min.

141 • Place the tube in a 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.

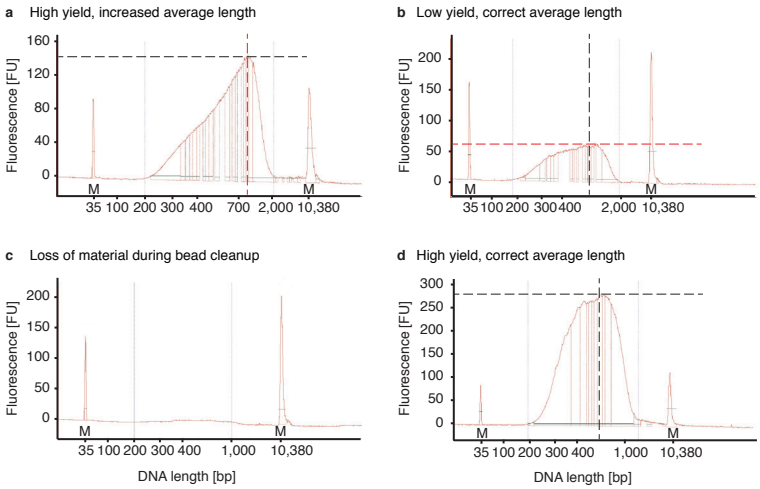


Figure 6 • Examples of DNA bioanalyzer plots.

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 the 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. Peaks marked with an 'M' indicate the reference markers; black and red dashed lines indicate the relevant optimal and suboptimal features, respectively.

Library quantification and sequencing

Timing: 19 h

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. 6.

?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 the 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 eight 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.

Downloading genome reference files and generating HISAT2 index

Timing: 1 h 30 min

Critical: For the purpose of this protocol, the FASTA (Fast-All) file, GTF (Gene Transfer Format) 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 (e.g., 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 (e.g., 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 for future alignment of both DamID- and CEL-Seq2-derived reads. For the HISAT2 index to optimally align spliced transcripts, it is necessary to provide information on exon and splice site locations. These can be extracted from the GTF file with scripts provided by HISAT2. Incorporating splice site and exon information requires a great deal of working memory—approximately 200 Gb RAM for the human genome. If this memory is not available, HISAT2 indices for the most commonly used genomes can also be directly downloaded from the HISAT2 website (http://daehwankimlab.github.io/hisat2/download/). To build the index yourself:

```
GTF="./references/Mus_musculus.GRCm38.98.with_ERCC.gtf"
FASTAFN="./references/Mus_musculus.GRCm38.dna.primary_assembly.with_ERCC.fa"
HISAT2_INDEX="./references/Mus_musculus.GRCm38.dna.primary_assembly.with_ERCC"
SSFN="./references/Mus_musculus.GRCm38.dna.primary_assembly.with_ERCC. \
splice_sites.txt"
EXONFN="./references/Mus_musculus.GRCm38.dna.primary_assembly.with_ \
ERCC.exons.txt"

hisat2_extract_splice_sites.py $GTFFN > $SSFN
hisat2_extract_exons.py $GTFFN > $EXONFN
hisat2-build --ss $SSFN --exon $EXONFN $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 represent five single cells from an 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 the scDam&T-seq package:

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

Generate GATC reference arrays

Timing: 1 h 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_assembly"
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 (Browser Extensible Data) format), '.posarray.hdf5' (all occurrences of the GATC motif as a HDF5 (Hierarchical Data Format 5) array), and '.maparray.hdf5' (the mappability of all GATC motifs as a HDF5 array), respectively.

Demultiplex raw data

Timing: 1 h

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 (e.g., from the different sequencing lanes). These files should be concatenated before the processing of DamID and CEL-Seq2 reads (Steps 157 and 159, respectively).

156 • For each Illumina library, demultiplex the reads based on the used adapters. Make sure that 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 (Fast-Quality) 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 h

Critical: The subsequent steps (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 h

Critical: The subsequent step (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.

TROUBLESHOOTING

Table 2: Troubleshooting

Step	Problem	Possible reason	Solution
28	Unexpected cell-cycle profile	Hoechst solution is too old. Hoechst solution has 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 that the 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 preparation with leftover aRNA Increase PCR cycles
	Low product (Fig. **3b**)	High adapter/product ratio inhibits library preparation	Bead purify the aRNA one to two times extra Repeat library preparation
After 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
	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 before induction.
	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
	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 that single-cell transcription data can be obtained from the material (e.g., wild-type control). Renew CEL-Seq2 primer plate. Optimize CEL-Seq2 primer and DamID adapter concentrations (e.g., reduce DamID adapter concentration)

After 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

TIMING

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

ANTICIPATED RESULTS

Figure **5** 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. **5a**). Most of the demultiplexed reads successfully align (Fig. **5b**), 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. **5b**).

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. **5c**) 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.

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. **5b**). To establish the complexity of the samples, we compare the number of valid reads to the

final number of unique counts (Fig. **5d**). 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).

Acknowledgements

We thank the members of the Kind laboratory 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. S.S.D. was also supported by the NIH grant R01HG011013. This work was funded by a European Research Council Starting grant (ERC-StG 678423-EpiID), a Nederlandse Orga- nisatie 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.

Additional information

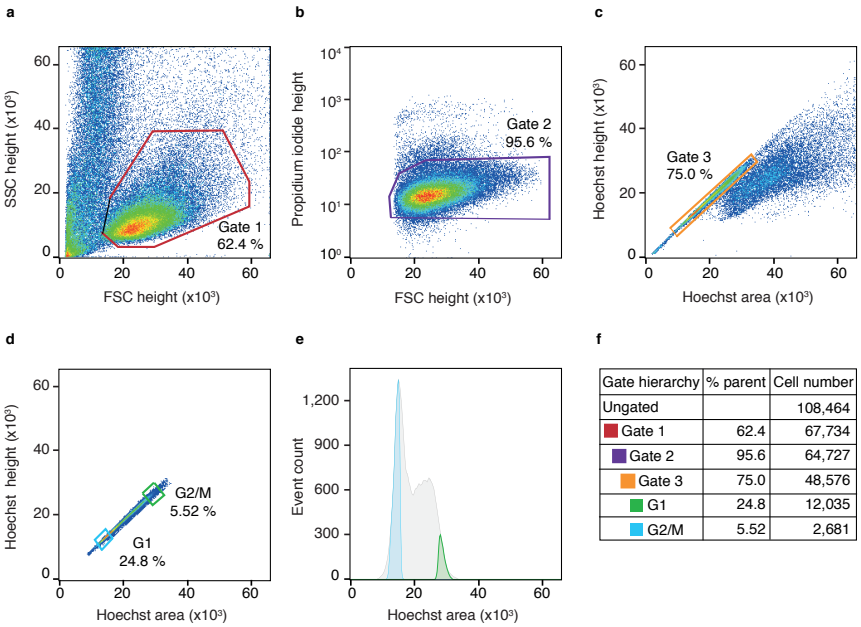
Supplementary information is available for this paper at https://doi.org/10.1038/s41596-020-0314-8.

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).

SUPPLEMENTAL FIGURE



f

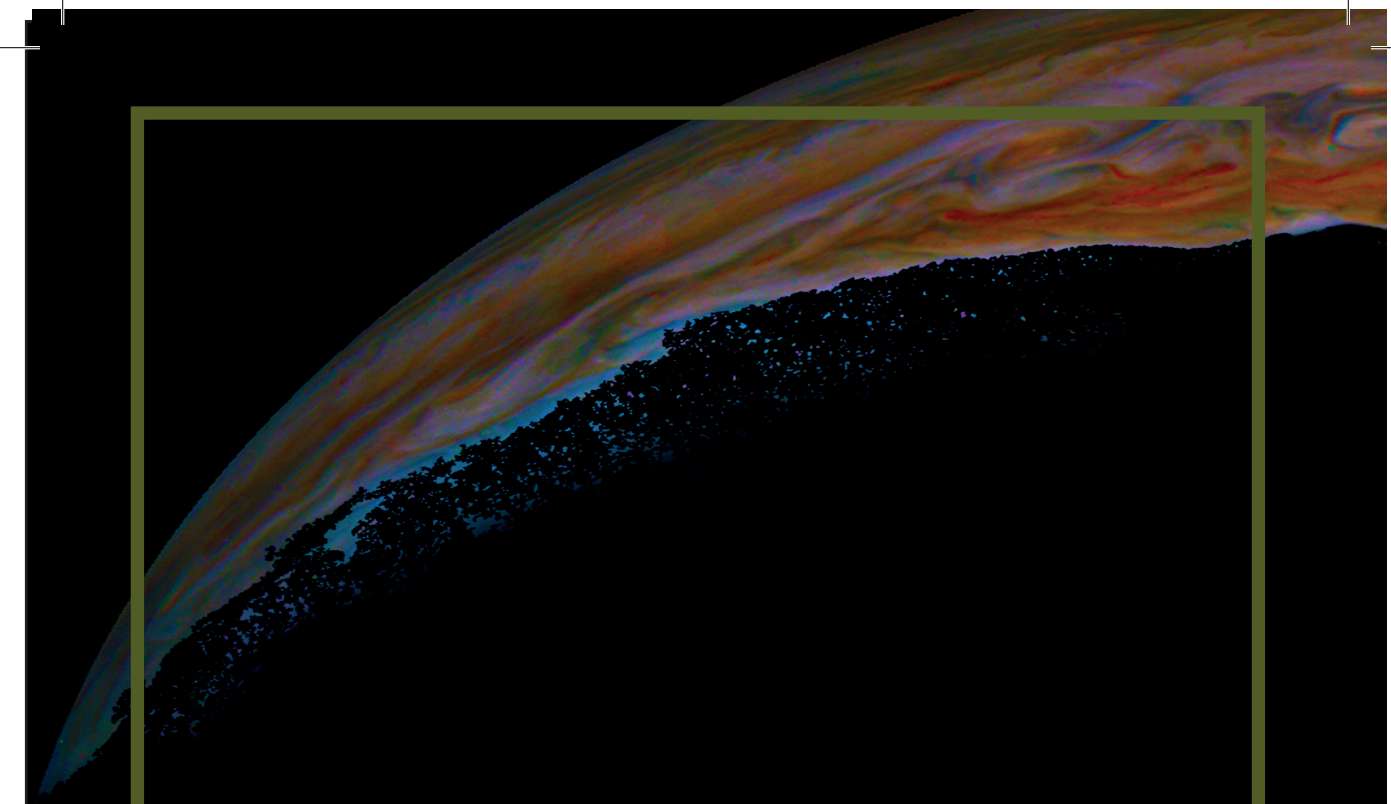
Gate hierarchy	% parent	Cell number
Ungated		108,464
Gate 1	62.4	67,734
Gate 2	95.6	64,727
Gate 3	75.0	48,576
G1	24.8	12,035
G2/M	5.52	2,681

g

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

Supplemental 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). Percentage of events in Gate 1 is indicated. **b.** Dot plot of mESCs passing Gate 1 showing gating strategy to exclude dead cells in Propidium iodide versus FSC. Percentage of events in Gate 2 is indicated. **c.** Dot plot of mESCs passing Gate 2 showing gating strategy to exclude duplet cells in Hoechst versus Hoechst area. 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 the FlowJo software, version 10.1r5.

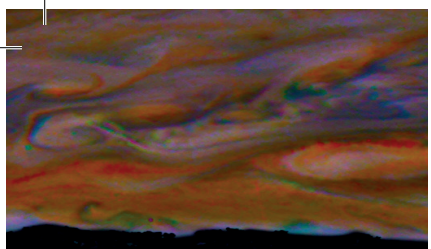


Chapter 4

scDam&T-seq maps lamina associated domains in developing cortex

Corina M. Markodimitraki, Tessy Korthout, Reinier van der Linden, Franka J. Rang, Youri Adolfs, Solee Pop, Samy Kefalopoulou, Onur Basak[#] and Jop Kind[#]

[#]co-corresponding



4

Abstract

The mammalian cerebral cortex is responsible for the processing of sensory information and the control of movement. Its development is governed by a multitude of gene regulatory mechanisms but the role of genome organization at the nuclear periphery in this process has not been addressed. To address this, we utilize in utero electroporation-mediated gene delivery in the developing cortex and apply scDam&T-seq on the targeted cells. We map the genome at the NL in the major cortex cell types and reveal cell-to-cell differences in genome architecture within the cortex stem cell pool.

INTRODUCTION

The mammalian cerebral cortex is a highly structured organ that controls high-order cognitive and sensorimotor information processing. It is composed of a plethora of morphologically and functionally distinct neurons. This diversity arises during development from a limited pool of progenitor cells located in the ventricular zone (VZ).

In mouse, the cerebral cortex is composed of 6 layers which are formed in an inverted fashion during development, with newborn neurons emerging from the VZ before migrating to their destination layer in the cortical plate (CP). This means that deep layer neurons are born first, followed by superficial layer neurons. The apical progenitors (APs) in the VZ pass through different phases of neurogenic competence as development progresses to facilitate this process. Starting from embryonic days (E) 10.5-11.5 they expand their numbers, after which they undergo a neurogenesis phase until E16. During this time APs can generate basal progenitors (BPs) or neurons. BPs have limited differentiation capacity compared to APs and can undergo 1-2 divisions, before generating neurons. Around E16.5 the APs enter the gliogenesis phase, generating astrocytes, oligodendrocytes and ependymal cells. Tight spatiotemporal regulation of corticogenesis is therefore essential.

Gene regulatory mechanisms act in an orchestrated manner to ensure normal cortex development. The Notch signaling pathway for instance, plays an important role in cell fate choice and proliferation of the APs[1-3]. Transcription factors such as Pax6 and Neurogenin1 and 2 contribute to faithful fate acquisition of newborn neurons[3-6]. Cortex development is also regulated by the Polycomb group complex (PcG), for instance by repressing proneuronal genes in APs and promoting the start of the astrogenic stage[7]. DNA modifications are also crucial. Cytosine methylation (5mC) prevents premature differentiation of APs into astrocytes[8-10]. Higher order chromatin organization such as topologically associated domains (TADs) and A/B compartments form an additional layer of gene regulation[11, 12]. In vivo studies show that corticogenesis is characterized by an increase in B-compartment robustness towards a more differentiated cell state, meaning that intra-compartment interactions are favored over inter-compartment interactions[13].

Genomic interaction with the nuclear periphery is a part of what constitutes higher order chromatin organization. The nuclear lamina (NL), a protein meshwork lining the inner side of the nuclear membrane, acts as a repressive environment and can serve as an anchoring dock for the genome. In vitro differentiation studies show that lamina associated domains (LADs) reorganize once mouse embryonic stem cells (mESCs) are pushed towards a neuronal fate[14]. Pluripotency genes move towards the periphery while neuronal genes reposition to the nuclear interior and these rearrangements are often accompanied by a change in transcriptional state. Similar LAD reorganizations have also been observed in other in vitro differentiation studies[15-17]. In Drosophila neuroblasts, the transcription factor *Hunchback* (*hb*) regulates motorneuron competence which is lost once the *hb* locus moves to the NL[18]. Consequentially, the neuroblasts start generating interneurons, a process which cannot be reversed by ectopically expressing *hb*. Another study shows that depletion of the Lamin B1 (LmnB1) protein, a component of the NL, results in aberrant gene expression in adult olfactory neurons and reduces odour detection[19]. These findings indicate that genome-NL interactions are crucial for fate establishment.

In previous work, changes in genome-NL interactions during development were studied in in vitro[20] or in non-mammalian systems[21-23]. Understanding the role of LADs in vivo in a mammalian dif-

ferentiation system is challenging. To detect LADs, DNA adenine methyltransferase identification (DamID) is used[24, 25]. For this, LmnB1 is expressed as a fusion protein with Dam, a bacterial methyltransferase protein that can methylate adenines in a GATC sequence motif. As the Dam-LmnB1 fusion protein is tethered to the NL, only regions that are in proximity to the periphery will get methylated by Dam. The fully methylated GATC sites are then specifically digested with the DpnI restriction enzyme and amplified for next generation sequencing. The lengthy process of establishing a strain of transgenic animals expressing Dam-LmnB1 is one of the reasons why this has not yet been achieved. In addition, expression levels of the fusion protein are crucial to acquire a good resolution, which makes the testing and fine tuning of the system difficult[23] Finally, cell type calling in the brain is key to studying possible differences in their LAD signatures, which adds another layer of difficulty.

Here, we investigate the feasibility of assigning LAD signatures to specific cell types in the developing mouse cortex at a single-cell resolution. We do this by combining two techniques: in utero electroporation (IUE)-mediated gene delivery and scDam&T-seq[26, 27]. IUE allows for spatially and temporally controlled transient gene expression by introducing plasmid DNA in targeted cells and their progeny, thereby bypassing the use of transgenic animals[26]. In addition, scDam&T-seq obtains transcriptome and protein-DNA interaction information from the same single cell by combining the single-cell RNA sequencing method CELseq2 and single-cell DamID (scDamID), respectively[27, 28]. Thus, scDam&T-seq allows for the identification of cell types based on transcriptome information and can reveal their underlying protein-binding signatures. The results presented here provide compelling evidence that we can identify cell types of the developing cortex and reveal their associated LAD signatures. Fine-tuning of experimental conditions will provide further understanding of the role of genome organization in motor cortex development.

RESULTS

To test the efficiency of expression of Dam-LmnB1 constructs by IUE, we designed different vectors expressing either a fluorescent protein (FP), a Dam-LmnB1 fusion protein or a combination thereof (Fig. **1a**). The DamID readout of scDam&T-seq is especially sensitive to the levels of GATC methylation by the Dam fusion protein[25, 29]. Overly high or long expression of the Dam fusion protein can lead to saturated signal, while too low levels or too short time periods of expression can lead to sparse data. We tested different levels of expression by injecting vectors with the Dam-LmnB1 construct under the control of promoters with various strengths. We used the human elongation factor 1 α (EF1α) promoter and the chicken β-actin and CMV early enhancer element (CAG) promoter, both of which are strong, although CAG is reported to behave differently across cell types[30]. We also used the weaker human phosphoglycerate kinase 1 (hPGK) promoter (Fig. **1a-b**). The constructs were either co-injected with a vector expressing GFP (hPGK-HA-Dam-LmnB1 plus GFP, hereafter hPGK-DL+GFP) or co-expressed a mRuby fluorescent protein (EF1α-mRuby-P2A-HA-Dam-LmnB1, hereafter EF1α-DL and CAG-mRuby-P2A-HA-Dam-LmnB1, hereafter CAG-DL), to allow for selection of successfully electroporated cells based on the presence of fluorescent signal using Fluorescence-Activated Cell Sorting (FACS) (Fig. **1c**).

Cells in the developing cortex have different cell cycle durations and this should be kept in mind when expressing the Dam-fusion protein. For our experimental setup we decided on a 48-hour expression window of the constructs (Fig. **1b**). We achieved this by electroporating the motor cortex of embryos on E14 and allowing development for two additional days until collection on E16. At col-

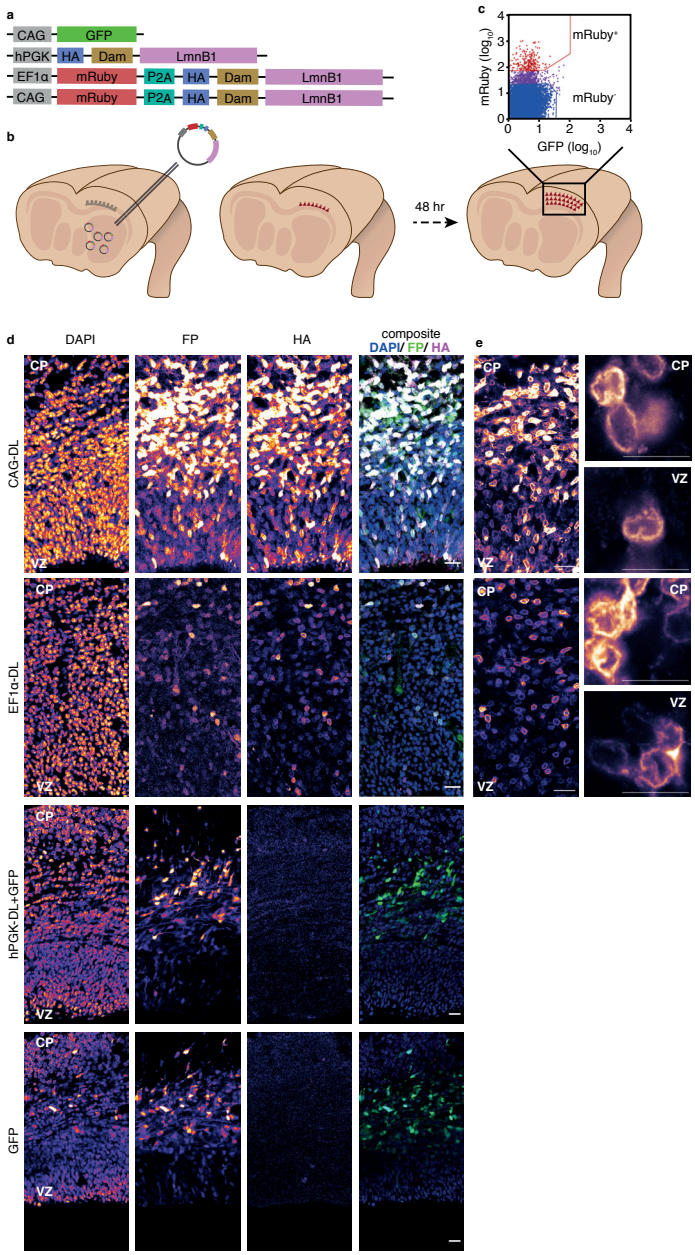


Figure 1 • Experimental setup and Dam-LmnB1 expression in cortex .

a. Constructs used for the IUE. **b.** Schematic of the IUE experimental setup. Constructs were injected into the lumen of the ventricle and guided towards the VZ with an electric voltage on E14. Embryos continued development until brain isolations 48 h after IUE. **c.** Schematic of the FACS experimental setup. GFP- or mRuby-positive and negative cells were sorted in the 384-well plates containing CEL-Seq2 barcoded primers. **d.** DAPI, FP expression and HA staining in coronal cortical sections 48 h after IUE. Scale bar is 20 μm. e. HA staining in coronal cortical sections 48 h after IUE. Scale bars are 20 μm (left) and 10 μm (right). IUE, in utero electroporation; VZ, ventricular zone; CP, cortical plate; FP, fluorescent protein (GFP or mRuby).

lection, we prepared the samples either for immunostaining or for FACS followed by scDam&T-seq.

In order to investigate genome architecture in the motorcortex, enrichment for the different cell types emerging from the VZ between E14 and E16 was crucial. Indeed, confocal microscopy on cortical sections of the electroporated embryos revealed fluorescently labelled cells (GFP or mRuby) in both deep and more superficial layers. (Fig. **1d**). Fluorescent signal was less intense in the VZ and subventricular zone (SVZ) compared to the CP in all conditions, which could be due to the dilution of the plasmid after each cell division of the proliferating APs and BPs. Overall, the fluorescent protein was successfully expressed in all conditions both in deep and more superficial layers of the cortex.

In addition to the expression and detection of the fluorescent protein, we confirmed both the expression and perinuclear localization of the Dam-LmnB1 fusion protein in all cortical layers. Staining for the HA epitope in the Dam-LmnB1 constructs showed signal in VZ, SVZ and CP for embryos electroporated with the CAG-DL and EF1α-DL vectors (Fig. **1d**). Furthermore, the exogenous Dam-LmnB1 fusion protein was successfully incorporated into the nuclear lamina in these conditions (Fig. **1e**). In contrast, HA-Dam-LmnB1 was not observed in the embryo electroporated with the hPGK-DL+GFP vectors which may have resulted from expression levels below the detection threshold. Overall, IUE of the Dam-LmnB1 construct results in the fluorescent labelling of cells within VZ, SVZ and CP and the correct perinuclear localization of the protein in the CAG-DL and EF1α-DL electroporated embryos. Even though hPGK-DL+GFP was not detectable with the HA staining, we continued with all conditions, because HA staining intensity is not directly translatable to DamID signal.

Following successful expression of Dam-LmnB1 constructs in the developing cortex, we tested whether scDam&T-seq can be employed on these samples. For this, motor cortex samples were dissociated into single-cell solutions from which we sorted 881 mRuby- or GFP-positive cells along with 136 control mRuby-negative cells. We processed the plates following the scDam&T-seq protocol, as described[27, 28]. Filtering based on the number of obtained transcripts (see Materials and Methods) yielded 851 single cells (84% of all sorted cells) with sufficient transcriptomic readout (Supplementary Fig. **1a**). Using Seurat[31, 32], we removed batch effects from the data and performed clustering (Fig. **2a**). This resulted in 5 clusters in which the different experimental conditions were uniformly represented (Supplementary Fig. **1b**). Clustering was not affected by cell-to-cell differences in unique GATC counts or transcript numbers (Supplementary Fig. **1c, d**). However, we did observe an overall lower GATC and transcript count in cluster 4, which was mainly comprised of mRuby-negative control cells (Supplementary Fig. **1e**). On closer inspection, these cells expressed either CAG-DL or EF1α-DL, two conditions with overall low unique gene counts (Supplementary Fig. **1a**).

We assigned cell types to the clusters based on differential gene expression (Fig. **2b-c,** Supplementary Fig. **1f**) and additional differential expression analysis (not shown). Cluster 1 was enriched for cells expressing the AP markers *Vim* and *Sox2*[33] (Fig. **2b-c**). *Pax6*, another progenitor marker was highly expressed in cluster 1, confirming this is mainly comprised of APs[34] (Supplementary Fig. **1f**). Cluster 2 showed predominantly expression of marker genes *Eomes* and *Ezr*, indicating enrichment for BPs[35, 36] (Fig. **2b-c** and Supplementary Fig. **1f**). Cells in cluster 0 were identified as neurons, showing high expression levels for the *Neurod2* gene, which is associated with neuron differentiation and fate establishment[36]. They were also enriched for the neuronal markers *Mapt* and *Tbr1*[36] (Fig.

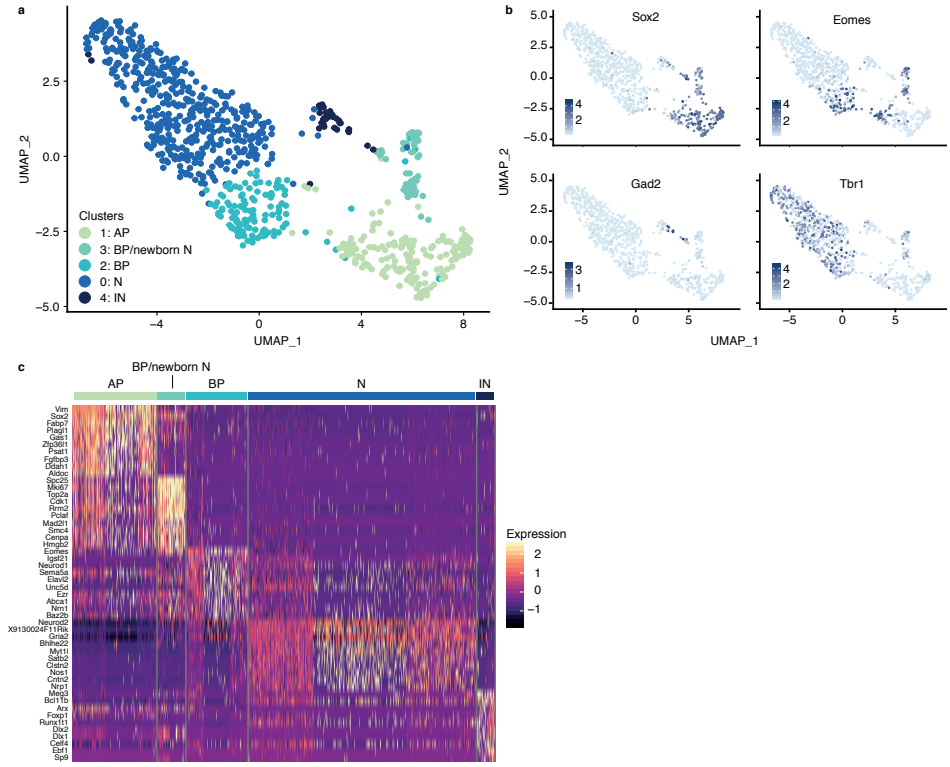


Figure 2 • scDam&T-seq identifies cell types in developing cortex

a. UMAP representation of the transcriptome scDam&T-seq data reveals 5 clusters. **b.** Expression of cell type markers. Colour bars indicate expression level. **c.** Clustered expression heatmap with the top 10 differentially expressed genes per cluster.

2b and Supplementary Fig. **1f**). At this stage of development, we detect neurons as expected according to previous single-cell RNA sequencing experiments[33, 36, 37]. Cluster 4 showed specific enrichment for interneuron markers *Dlx1* and *Dlx2*, suggesting that it contains mainly dorsally migrating neurons originating from the ventral pallium[36] (Fig. **2b-c** and Supplementary Fig. **1f**). Lastly, cluster 3 showed high expression of cell-cycle regulators such as *Cdk1*, indicating proliferation and DNA-damage response genes such as *Pclaf* and *Hmgb2* which have been associated with newborn neurons[33, 36] (Fig. **2c**). Cluster 3 also shows expression of the BP marker *Eomes* and the gene *Sox11*, reported to play a role in early-born neurons[36] (Fig. **2b-c** and Supplementary Fig. **1f**). Thus, cluster 3 possibly represents both proliferating BPs and newborn neurons. Cells in transition might show a large cell-to-cell variety in genome-NL interactions, which could lead to a misrepresentation in the average LAD profile of this cluster. For this reason, we exclude this cluster for further comparative analysis of DamID data between clusters, although this kind of dynamic cluster is interesting for further studies. A differentiation trajectory analysis tool such as Monocle would be suitable for this[38]. In conclusion, scDam&T-seq is able to identify the main cell types based on transcripts of the developing cortex at E16.

We investigated the DamID performance for each of the tested conditions. We applied a minimum

threshold of 25,000 unique GATCs per cell (Fig. **3a**), which resulted in success rates of 78.8, 51.6 and 15.7% for the CAG-DL, EF1α-DL and hPGK-DL+GFP conditions, respectively. As few cells passed the GATC threshold in the latter condition, we rejected it as a suitable IUE condition for DamID. Next, the mean DamID signal was calculated for cells in each condition, which showed similar Dam-LmnB1 enrichment patterns for the three conditions on chromosome 12 (Fig. **3b**). In addition, we were able to visually identify LADs in single cells for all IUE conditions, even observe

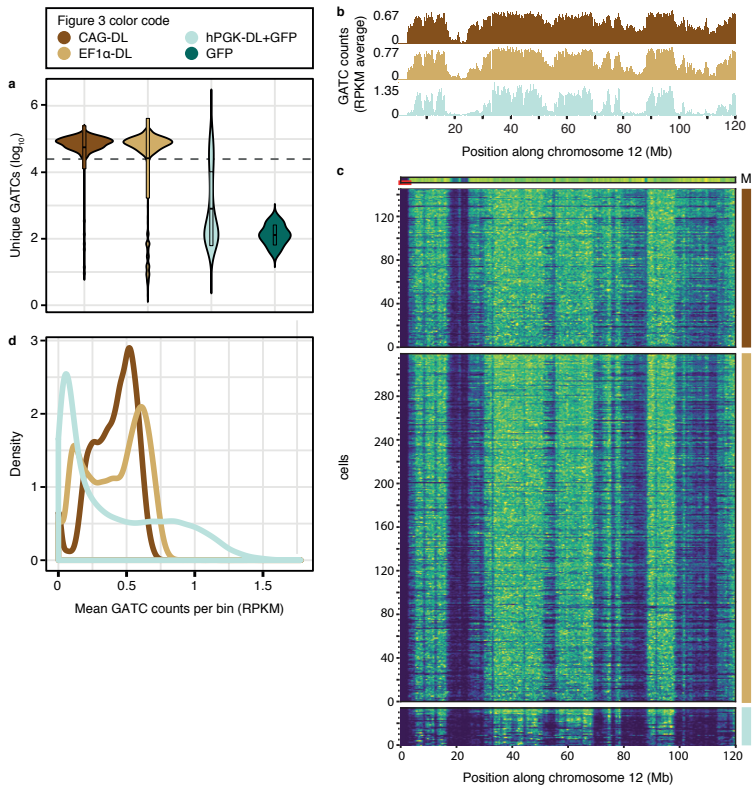


Figure 3 • scDam&T-seq identifies LADs in developing cortex

a. Violin plots showing number of unique GATCs per electroporation condition for GFP- or mRuby-positive cells. mRuby-negative control cells are not shown. Dashed line indicates the DamID threshold of 25,000 unique GATCs per cell. Boxes in the violin plots indicate mean value ± 1 the standard deviation (SD). **b.** RPKM-normalized GATC counts per bin of chromosome 12, averaged over cells of each IUE condition. Averages were calculated from n=145, 319 and 34 single cells for the CAG-DL, EF1α-DL and hPGK-DL+GFP conditions, respectively. **c.** Single-cell DamID heatmaps of chromosome 12 showing the number of observed unique GATC counts per bin for single cells of each IUE condition Signal is scaled to max per cell and color coded with a range; blue signal for bins without GATCs and yellow the bins with the maximum unique GATC number for that cell. Cells are ordered from top to bottom on high to low depth (unique GATC number). Each row represents a single cell, and each column a genomic bin of 100 kilobases (kb). The mappability track indicates the mappable GATC fragments. Red horizontal line on the mappability track represents unmappable genomic regions. IUE conditions are represented by vertical coloured lines on the right side of the tracks. M; mappability. **d.** Density plot of bins genome-wide with RPKM-normalized GATC counts averaged over all cells of each IUE condition. For panels b-d, only cells passing the CEL-Seq2 threshold of >200 and <6,000 unique genes and the DamID threshold of >25,000 unique GATCs were used with a 100 kb bin size.

cell-to-cell variation (Fig. **3c**). This data demonstrates that the combination of IUE and scDam&T-seq allows for the identification of LADs in an in vivo system.

To successfully distinguish LADs from inter-LAD (iLAD) regions which do not contact the NL, a clear difference must be observed between true signal and non-specific background methylation which can be the result of accessibility of the DNA to the Dam fusion protein. Under this condition, one can successfully measure differences in GATC counts between genomic bins contacting the NL and bins that do not. The mean DamID signal for all genomic bins of the CAG-DL condition revealed a narrow signal range and vague distinction between high and low GATC count bins (Fig. **3d**). The EF1α-DL-expressing cells on the other hand, showed a clear bimodal distribution, with low signal bins clearly separating from high signal bins. We therefore used the EF1α-DL dataset for further analysis.

We visualized LAD profiles in APs, BPs and neurons of the EF1α-DL condition based on the clusters that were assigned in the transcriptome analysis. For this analysis we excluded the mixed population of cluster 3 as well as cluster 4 which contained only 3 cells passing the threshold. DamID signal plotted as an average per bin showed similar LAD patterns for the three cell types on chromosome 12, and this was also evident from the signal per single cell (Fig. **4a-b**). This is for most part expected, as LADs can show an overlap of ~ 80% between cell types[14]. The density plot of DamID averages per bin genome-wide, shows a reduction in bins with low mean GATC values in in BPs and neurons (Fig. **4c**). This could either be due to a true gain in genome-NL associations in more differentiated cells, or because AP LADs do not get erased in differentiating non-dividing cells, resulting in the detection of a higher overall LAD number.

We plotted DamID single-cell profiles for chromosome 8 in APs, BPs and neurons (Fig. **4d** and Supplementary Fig. **2a-b**). Besides overall similar LAD patterns between single cells of the AP cluster, we also observed striking differences at megabase-size scale (Fig. **4d**). This variation is likely not due to data quality differences because the depicted cells had similar GATC counts. Similar inspection of the BP and neuron DamID profiles did not show as striking differences, although we did observe some changes in smaller regions (Supplementary Fig. **2b**). Overall, our data suggests that IUE in combination with scDam&T-seq can identify cell types in the developing mouse cortex and reveal underlying differences in LAD signatures, at a single-cell resolution.

DISCUSSION

We demonstrate that the combination of IUE and scDam&T-seq enables the measurement of transcription and LADs simultaneously in single cells of the developing mouse cortex. We identify different cell types based on their transcriptional profiles and observe differences in LADs mainly within the apical progenitor pool.

We tested constructs expressing the Dam-LmnB1 fusion protein under the control of promoters with varying strengths. Our results show that the hPGK-DL+GFP co-electroporated vectors gave the cleanest DamID profiles of all the tested conditions (Fig. **3a-c**). This could be due to the lower amount of injected Dam-LmnB1 vector compared to the CAG- and EF1α-DL vectors (see Methods). Limiting the expression of the Dam fusion protein could thus lead to cleaner DamID profiles. Even though the hPGK-DL+GFP condition showed the least background methylation, only 15.7 % of cells passed the GATC threshold (Fig. **3a**). This could be related to the fact that we selected

cells based on fluorescence, not on presence of Dam-LmnB1 construct per se. This means that a sorted cell would not necessarily express the Dam-LmnB1 protein, as the GFP and the DL were in two different vectors. The bimodal distribution of **Fig. 3a** shows indeed that few cells acquire both vectors and pass the applied GATC threshold. The CAG-DL and EF1α-DL conditions on the other hand, showed much higher single-cell success rates, indicating that electroporation of one

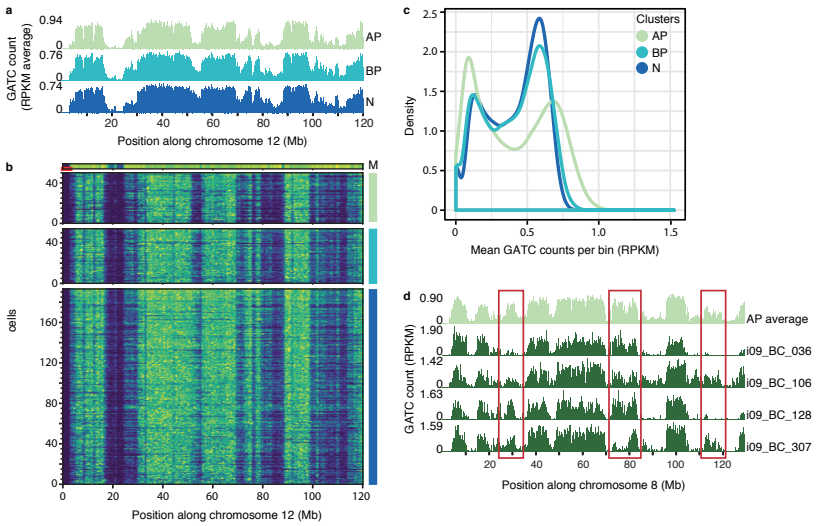


Figure 4 • EF1α-DL-expressing cells reveal heterogeneity in LADs

a. RPKM-normalized GATC counts per bin of chromosome 12, averaged over all cells of the AP, BP and N clusters for cells of the EF1α-DL IUE condition. Averages were calculated from n=50, 54 and 192 single cells for the AP, BP and N clusters, respectively. **b.** Single-cell DamID heatmaps of chromosome 12 showing the number of observed unique GATC counts per bin, as in Fig. 3c, for single cells of the AP, BP and N clusters in the EF1α-DL IUE condition. Clusters are represented by vertical coloured lines on the right side of the tracks. M; mappability. **c.** Density plot of bins genome-wide with RPKM-normalized GATC counts averaged over all cells of each cluster in the EF1α-DL IUE condition. **d.** RPKM-normalized DamID signal per bin of chromosome 8 for four single cells of the AP cluster in the EF1α-DL IUE condition. Top track, RPKM-normalized DamID signal per bin averaged over all cells of the AP cluster for cells of the EF1α-DL IUE condition. Red boxes indicate variable regions between the single cells. Cells shown are within the top 10 cells with the highest counts of unique GATCs per cell in the AP cluster. For panels a-d, only cells passing the CEL-Seq2 threshold of >200 and <6,000 unique genes and the DamID threshold of >25,000 unique GATCs were used with a 100 kb bin size. AP, apical progenitors; BP, basal progenitors; N, neurons; i, Illumina index; BC, unique DamID barcode.threshold of >25,000 unique GATCs were used with a 100 kb bin size.

vector co-expressing the fluorescent protein and Dam-LmnB1 is most optimal. We collected the electroporated area of the cortex 48 hours after IUE which showed Dam-LmnB1 enrichment for all conditions. Embryos electroporated with CAG-DL and EF1α-DL constructs however had an elevated level of background methylation compared to the hPGK-DL+GFP condition (**Fig. 3a-c**). Collecting the tissue earlier could help reduce unspecific Dam methylation.

We identified cell types of the E16 motor cortex and calculated their average Dam-LmnB1 pro-

files. We find that cell types differ in amount of DamID methylation, with APs showing fewer low GATC-count bins compared to the other two cell types (Fig. **4a-c**). The HA-tag staining in both the CAG-DL- and EF1α-DL-electroporated embryos is another indication for this, as there was less HA signal in the VZ where the APs are located compared to the rest of the regions (Fig. **1d**). This could be due to technical or biological reasons, or a combination of both. Technically, it is possible that the constructs were expressed for too long. Fast-cycling cells such as progenitors in the ventricular zone[36] lose GATC methylation after replication and are therefore less affected by long periods of Dam-LmnB1 expression. Neurons, on the other hand, are prone to background methylation if the expression window of the Dam fusion protein is too long, as they are post-mitotic and will thus not deplete GATC methylation during replication. A possible technical alteration would thus be to express Dam-LmnB1 for shorter periods of time, although this could come at the cost of the amount of unique GATCs in APs. Because the cell types differ in cell cycle properties, it is likely that cell-type optimized expression windows are needed. Another option would be to sort cells based on their fluorescence intensity. For this, more extensive analysis is needed to correlate fluorescence intensity with DamID signal. The differences in amount of DamID methylation between the different cell types could also be explained from a biological perspective. Newborn daughter neurons have been found to inherit the transcriptional profile of their mother progenitor cells[36, 39-42]. They acquire their final cell fate only after migration towards their destination layer is complete[36]. It is plausible that besides transcripts, also LADs are transmitted from progenitors to their neuron progeny and are reorganized only once the neuron has stopped migrating. Because neurons scarcely divide once they have started migration[3], LAD changes could result in an accumulation of DamID signal of the AP and neuron transcriptional states.

We were able to obtain clear LAD profiles from single cells for all clusters (Fig. **4d** and Supplementary Fig. **2a-b**). LAD differences within the AP population were clear, possibly indicating distinct neurogenic capacity, or differentiation status. Cell cycle phase differences could also contribute. In contrast, single-cell variation was not as obvious within the BP and neuron clusters, although more extensive analysis is needed to confirm this. However, we might hypothesize that cell-to-cell LAD variation decreases along the differentiation trajectory. This would be in line with research showing that intra-B-compartment interactions increase in neurons of the developing cortex[13]. Chromatin compaction at the NL was also observed in differentiated cells of the hematopoietic system[43]. These findings indicate that heterochromatin condenses as differentiation potential decreases, which could translate to less mobile and variable LADs between neurons. However, lack of single cell heterogeneity in neurons could be due to technical reasons as well, because scDam&T-seq measures accumulated Dam methylation and loci retain the m^6A mark even when they move to the nuclear interior. To uncover if single cell variation is indeed lost in differentiated neurons of the cortex, one could use a snapshot-like technique such as the single-cell chromatin immunocleavage (ChIC)-based method CUT&RUN against the LmnB1 protein[44]. For this, single cells of each cell type could be sorted based on fluorescent markers. Even though single-cell transcriptional resolution would be lost, one would gain insight into the true single-cell LAD signal within each population.

As the Polycomb complex PRC2 has been found to contribute to the faithful acquisition of neurogenic potential of the AP population[36], it would be interesting to investigate the occupancy of the complex or the histone PTM mark it catalyses, H3K27me3, by using our system as well. Our lab has obtained single-cell profiles of H3K27me3 in combination with transcriptome by using a Dam-H3K27me3 nanobody construct (unpublished data). Since LAD borders have been shown to be enriched for the H3K27me3 mark[45], it could be of interest to investigate the potential interplay these

different regulatory mechanisms in the developing cortex.

Altogether, we demonstrate a flexible system in which we utilize IUE and scDam&T-seq to identify cell types in the brain and measure their nuclear organization, with potential future applications for other Dam fusion proteins as well. These results build a roadmap for the further investigation of spatial genome organization during developmental processes.

METHODS

Mice

All experiments were performed in accordance with the institutional guidelines of the University Medical Center Utrecht, approved by Experimental Animal Committee Utrecht (DEC-Utrecht, University Utrecht, Utrecht, The Netherlands) and conducted in agreement with the national and international law (Guideline 86/609/EEC). C57BL/6J mice were obtained from the Jackson laboratories, bred in-house and used for the experiments. All mice were raised with their mothers up to 4 weeks of age, maintained in 12 light-dark cycle at temperature 22±1˚C and fed ad-libitum (211 RMH-TM diet; hope farms). Mice were housed in transparent boxes with wood chips and tissue for nest building. Embryonic day 0.5 was established as the day of the vaginal plug. Both female and male embryos were used throughout the study.

Cloning and vectors used

We used the hPGK-Dam-LmnB1 construct previously described[27]. For the hPGK-DL+GFP and GFP conditions we used the pEGFP-N1 vector (Clontech). We used a custom synthesized pUC57 vector containing the mRuby-P2A-HA-miniAID-Dam-LmnB1 sequence (GENEWIZ). The CAG-DL vector was cloned in the following way: we used the pCAG-tdTomato plasmid (Addgene, #83029) as the target vector and cut out the tdTomato by digestion with KpnI and NotI (NEB). Then, the mRuby-P2A-HA-miniAID-Dam-LmnB1 sequence including the beta-globulin polyadenylation site (BGpA) was amplified by PCR (primers 136/137, see Table 1) and inserted in the cut backbone by Gibson. The EF1α-DL vector was cloned in the following way: pCCL-sin-EF1α-Dam-Lmnb1 was PCR amplified using primers 138/139 to include the EF1α promoter, vector backbone, the WPRE and pA sequences to generate the backbone. mRuby-P2A-HA-miniAID-Dam-LmnB1 sequence was amplified by PCR (primers 140/141, see Table 1) and inserted into the backbone using Gibson. PCR amplification was performed using the KOD DNA polymerase (Merck Millipore) using 2mM MgSO4 at 60˚C, and the final products were sequence verified by restriction digestion and Sanger sequencing (Macrogen).

In utero electroporations

In utero electroporation was performed as previously described[46]. Briefly, E14 pregnant C57BL/6J mice were put under anesthesia using Isoflurane (induction: 3-4%, surgery: 1.5-2%) and injected with 0.05 mg/kg buprenorfin-hydrochloride in saline. Under sterile conditions mice were opened at the abdominal cavity to expose the uterine horns. For the hPGK-Dam-LmnB1+GFP condition a DNA solution was prepared with sterile PBS containing the hPGK-Dam-LmnB1 construct at a 0.6 μg/μl concentration and the GFP construct at a 0.4 μg/μl concentration. For the GFP condition the DNA solution contained the GFP construct at a 0.4 μg/μl concentration. For the CAG-DL and EF1α-DL conditions the DNA solutions contained either construct at a 2.0 μg/μl concentration. For all DNA solutions the constructs were dissolved in MiliQ water and 0.05% Fast Green (Sigma) and were injected at a volume of 1.7 μl unilaterally into the lateral ventricle of embryos using glass micro-pipettes (Harvard Apparatus) and a PLI-100 Picoinjector (Harvard Apparatus). The motor cortex was electroporated with gold-plated tweezer electrodes (Fischer Scientific) and an ECM 830 Electro-Square-Porator (Harvard Apparatus, Holliston, MA) set to 5 unipolar pulses of 50 ms pulse length at 30 V (950 ms interval). The whole procedure from anesthesia to awaking of the mother by release from Isoflurane took approximately 1 hr. The mice were monitored daily until E16.5 when the mothers were sacrificed via cervical dislocation. We used multiple constructs in one litter to reduce the number of mice necessary.

Cryosectioning

Mothers were sacrificed at E16 (48 hr after IUE) and whole brains from the pups where collected for immunostainings in ice-cold L-15 medium (Thermofisher Scientific), then washed in phosphate-buffered saline (PBS, pH 7.4) and fixed by immersion in 4% paraformaldehyde (PFA) in home-made phosphate buffered saline (PBS) at 4˚C overnight. Brains were washed in PBS, cryoprotected in 30% sucrose in PBS at 4˚C and frozen in 2-methylbutane, after which they were embedded in embedding matrix (Thermo Scientific). Brains were sliced into 25 μm coronal sections using a cryostat (Leica, CM1950 Ag Protect) set at -14˚C, mounted on slides (VWR Superfrost plus), air-dried and stored at -80˚C.

Immunohistochemistry

Slides were hydrated in home-made PBS with 0.5% TritonX-100 (Sigma) (PBS-T) for 10 min at room temperature (RT) after which excess PBS-T was removed. Slides were treated with 400 μl blocking buffer (2% donkey serum in PBS-T) at RT for 30 min in a humidifying chamber followed by primary antibody incubation in 250

μl of blocking buffer at RT for 2 hr in a humidifying chamber. Slides were washed three times with PBS before secondary antibody incubation in 300 μl blocking buffer at RT overnight in a humidifying chamber. Sections were washed 3 times with PBS, counterstained with 1 ug/ml DAPI (Sigma) and mounted in 10 ml Prolong Antifade reagent mounting medium (Thermofischer Scientific). Images were acquired using a Leica TCS SPE confocal microscope using the Leica LAS-X software. The following antibodies were used: Rabbit anti-HA epitope C29F4 (1:500, Cell Signaling; 3724), Donkey a-rabbit IgG (H+L) Alexa Fluor 647 (1:1,000, Thermofischer Scientific; A-31573).

Cortex dissociation and EACS

Mothers were sacrificed at E16 (48 hr after IUE) and whole brains from the pups where collected in Leibovitz's L-15 medium (Gibco). Cortices dissected under a stereo microscope (Leica) were dissociated using the Papain Dissociation System (Worthington Biochem) for 10 min at 37˚C, followed by mechanical dissociation, another incubation at 37˚C for 5 min and a final dissociation into single cells. Cells were washed in 10 ml of L-15 medium (Gibco) supplemented with 2% fetal bovine serum (FBS) (company) to stop enzyme digestion. Cells were centrifuged at 200 g for 7 min, resuspended in L-15 medium supplemented with 25 μg/ml DNaseI and filtered through a 40 μm mesh. DAPI was added to a final concentration of 0.5 μg/μl (Sigma) to each sample before sorting. Samples were sorted as single cells in 384-well plates (Biorad) with the BD Influx Cell Sorter. GFP positive cells were sorted for the hPGK-DL+GFP and GFP samples. mRuby positive cells were sorted for the EF1α-DL and CAG-DL samples.

scDam&T-seq

Cells were processed with the scDam&T-seq pipeline described in detail elsewhere[27, 28]. For all dispension steps we used the Nanodrop 2 robot (BioNex). For the primer and adapter dispension we used the Mosquito HTS robot (TTP Labtech). For the hPGK-DL+GFP and the GFP samples we used a set of 96 DamID2 adapters and 384 CEL-Seq2 primers (sequences can be found in [27]) at a 64 nM and 500 nM concentration respectively. For the CAG-DL and the EF1α-DL samples we used a set of 384 DamID2 adapters and 384 CEL-Seq2 primers (sequences can be found in [28]) at a 25 nM and 1500 nM concentration respectively. Sequencing libraries were sequenced 2 x 75 bp paired-end on the Illumina NextSeq.

CEL-Seq2 analysis

We followed the previously described scDam&T-seq data processing pipeline by aligning reads to the mm10 mouse genome[27, 28] and obtained gene count table files for CEL-Seq2. Because the data was sequenced in three different NextSeq runs, we followed the Seurat v3 data integration vignette offered by the Satija lab[31, 32] (https://satijalab.org/seurat/v3.1/immune_alignment.html). In short, we created a Seurat object including all 1,016 sequenced single cells. We then excluded cells that met either of the following conditions: 1) expression of < 200 unique genes, 2) expression of > 6,000 unique genes 3) mitochondrial content >10%. We also excluded genes from the analysis that were expressed in less than 3 cells. Out of the 1,016 cells, 851 cells passed the threshold. We split the object by sequencing run and applied the function NormalizeData(normalization.method = "LogNormalize", scale.factor = 10000) which divides the UMI count of a gene of cell χ, by the total UMI counts of cell χ then scales by a factor 10,000 and the outcome is log-transformed. Next, the top 2,000 variable genes were calculated with the function FindVariableFeatures (selection.method = "vst", nfeatures = 2000). Next, we calculated the integration anchors across the datasets with the function FindIntegrationAnchors(dims = 1:30). Then we integrated the data of the three runs with the function IntegrateData(anchorset = anchors, dims = 1:30) which batch corrects the values. We then scaled the UMI counts with the function ScaleData (default settings). Afterwards we performed dimensionality reduction with the function RunPC and computed 30 principal components. Next, we used the first 30 principal components to perform Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction with the function RunUMAP(reduction = pca, dims = 1:20). Finally we, clustered the cells using a graph-based clustering approach, with the functions FindNeighbours(dims = 1:20) and, based on the Louvain algorithm the function FindClusters(resolution = 0.3).

scDamID analysis

We followed the previously described scDam&T-seq data processing pipeline and obtained 100 kb-binned hdf5 files containing the DamID unique GATC counts[27, 28]. We normalized the GATC counts for each cell by dividing the GATC counts per bin by the GATC sum of all bins. Then we divided this value by a factor 1,000,000 and multiplied by the binsize in kb (100 kb). For figure 3 and 4, only cells passing the CEL-Seq2 threshold (see "CEL-Seq2 analysis") were used. Mappability was calculated as [27].

Table 1: primer sequences

Primer 136	AATTCTGCAGTCGACGGTACGCCACCATGGTGTCTAAGGG
Primer 137	TGCACCTGAGGAGTGCGGCCTTACATAATGGCACAGCTTTTATTG
Primer 138	AAAGCTGTGCCATTATGTAATCTAGACTCGACAATCAACCTCT
Primer 139	CCCTTAGACACCATGGTGGCGGTACCGTCGATCGACTGCA
Primer 140	TGCAGTCGATCGACGGTACCGCCACCATGGTGTCTAAGGG
Primer 141	GGTTGATTGTCGAGTCTAGATTACATAATGGCACAGCTTTTATTG

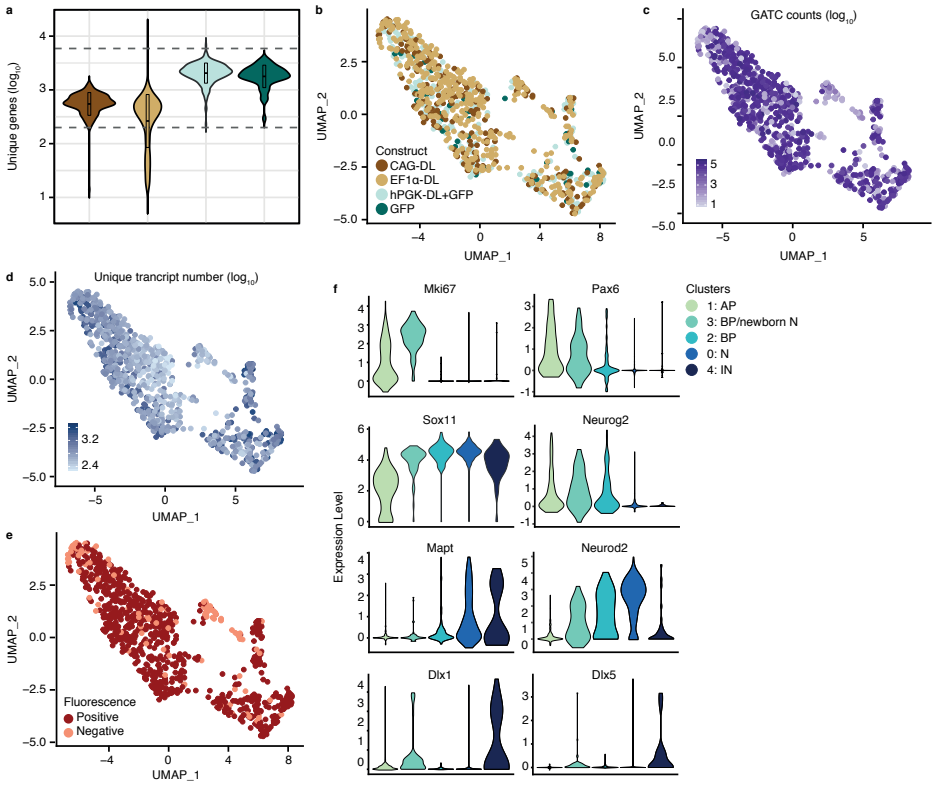
Author contributions

C.M.M. and O.B. conceived the study. C.M.M. performed immunofluorescence, FACS and scDam&T-seq experiments and bioinformatic analysis. O.B. cloned constructs, supervised electroporations, isolated embryos and performed FACS experiments. T.K. and F.J. R. performed bioinformatic anlaysis. R.v.d.L. performed FACS experiments. Y.A. performed electroporations. S.P. assisted with immunostainings and FACS. S.K. designed the constructs. J.K. supervised the study.

REFERENCES

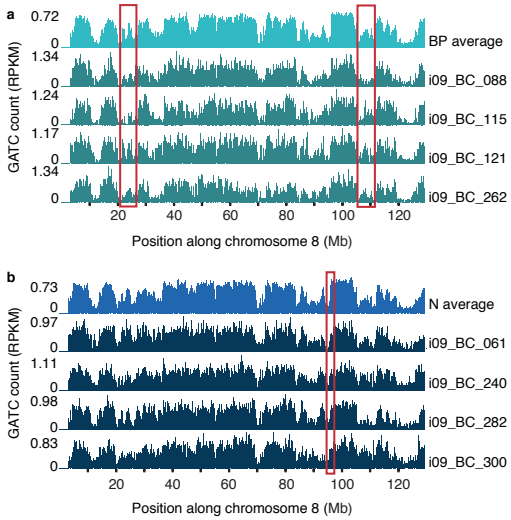
1. Gaiano, N., Nye, J.S. & Fishell, G. Radial glial identity is promoted by Notch1 signaling in the murine forebrain. Neuron 26, 395-404 (2000).
2. Lutolf, S., Radtke, F., Aguet, M., Suter, U. & Taylor, V. Notch1 is required for neuronal and glial differentiation in the cerebellum. Development 129, 373-385 (2002).
3. Mukhtar, T. & Taylor, V. Untangling Cortical Complexity During Development. J Exp Neurosci 12, 1179069518759332 (2018).
4. Muzio, L. et al. Emx2 and Pax6 control regionalization of the pre-neuronogenic cortical primordium. Cereb Cortex 12, 129-139 (2002).
5. Schuurmans, C. et al. Sequential phases of cortical specification involve Neurogenin-dependent and -independent pathways. EMBO J 23, 2892-2902 (2004).
6. Fode, C. et al. A role for neural determination genes in specifying the dorsoventral identity of telencephalic neurons. Genes Dev 14, 67-80 (2000).
7. Hirabayashi, Y. et al. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. Neuron 63, 600-613 (2009).
8. Takizawa, T. et al. DNA methylation is a critical cell-intrinsic determinant of astrocyte differentiation in the fetal brain. Dev Cell 1, 749-758 (2001).
9. Fan, G. et al. DNA methylation controls the timing of astrogliogenesis through regulation of JAK-STAT signaling. Development 132, 3345-3356 (2005).
10. He, F. et al. A positive autoregulatory loop of Jak-STAT signaling controls the onset of astrogliogenesis. Nat Neurosci 8, 616-625 (2005).
11. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289-293 (2009).
12. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38, 1348-1354 (2006).
13. Bonev, B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557-572 e524 (2017).
14. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell 38, 603-613 (2010).
15. Robson, M.I. et al. Tissue-Specific Gene Repositioning by Muscle Nuclear Membrane Proteins Enhances Repression of Critical Developmental Genes during Myogenesis. Mol Cell 62, 834-847 (2016).
16. Poleshko, A. et al. Genome-Nuclear Lamina Interactions Regulate Cardiac Stem Cell Lineage Restriction. Cell 171, 573-587 e514 (2017).
17. See, K. et al. Lineage-specific reorganization of nuclear peripheral heterochromatin and H3K9me2 domains. Development 146 (2019).
18. Kohwi, M., Lupton, J.R., Lai, S.L., Miller, M.R. & Doe, C.Q. Developmentally regulated subnuclear genome reorganization restricts neural progenitor competence in Drosophila. Cell 152, 97-108 (2013).
19. Gigante, C.M. et al. Lamin B1 is required for mature neuron-specific gene expression during olfactory sensory neuron differentiation. Nat Commun 8, 15098 (2017).
20. Borsos, M. et al. Genome-lamina interactions are established de novo in the early mouse embryo. Nature 569, 729-733 (2019).
21. Towbin, B.D. et al. Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. Cell 150, 934-947 (2012).
22. Gonzalez-Sandoval, A. et al. Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. Cell 163, 1333-1347 (2015).
23. Meister, P., Towbin, B.D., Pike, B.L., Ponti, A. & Gasser, S.M. The spatial dynamics of tissue-specific promoters during C. elegans development. Genes Dev 24, 766-782 (2010).
24. Pickersgill, H. et al. Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat Genet 38, 1005-1014 (2006).
25. 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).
26. Saito, T. & Nakatsuji, N. Efficient gene transfer into the embryonic mouse brain using in vivo electroporation. Dev Biol 240, 237-246 (2001).
27. Rooijers, K. et al. Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. Nat Biotechnol 37, 766-772 (2019).
28. Markodimitraki, C.M. et al. Simultaneous quantification of protein-DNA interactions and transcriptomes in single cells with scDam&T-seq. Nat Protoc (2020).
29. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134-147 (2015).
30. Qin, J.Y. et al. Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. PLoS One 5, e10611 (2010).
31. 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).

32. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902 e1821 (2019).
33. Telley, L. et al. Sequential transcriptional waves direct the differentiation of newborn neurons in the mouse neocortex. Science 351, 1443-1446 (2016).
34. Gotz, M., Stoykova, A. & Gruss, P. Pax6 controls radial glia differentiation in the cerebral cortex. Neuron 21, 1031-1044 (1998).
35. Arnold, S.J. et al. The T-box transcription factor Eomes/Tbr2 regulates neurogenesis in the cortical subventricular zone. Genes Dev 22, 2479-2484 (2008).
36. Telley, L. et al. Temporal patterning of apical progenitors and their daughter neurons in the developing neocortex. Science 364 (2019).
37. Loo, L. et al. Single-cell transcriptomic analysis of mouse neocortical development. Nat Commun 10, 134 (2019).
38. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32, 381-386 (2014).
39. Zahr, S.K. et al. A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification. Neuron 97, 520-537 e526 (2018).
40. Yoon, K.J. et al. Temporal Control of Mammalian Cortical Neurogenesis by m(6)A Methylation. Cell 171, 877-889 e817 (2017).
41. Yoon, K.J., Vissers, C., Ming, G.L. & Song, H. Epigenetics and epitranscriptomics in temporal patterning of cortical neural progenitor competence. J Cell Biol 217, 1901-1914 (2018).
42. Ozair, M.Z. et al. hPSC Modeling Reveals that Fate Selection of Cortical Deep Projection Neurons Occurs in the Subplate. Cell Stem Cell 23, 60-73 e66 (2018).
43. Ugarte, F. et al. Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells. Stem Cell Reports 5, 728-740 (2015).
44. Ku, W.L. et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. Nat Methods 16, 323-325 (2019).
45. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948-951 (2008).
46. Cunha-Ferreira, I. et al. The HAUS Complex Is a Key Regulator of Non-centrosomal Microtubule Organization during Neuronal Development. Cell Rep 24, 791-800 (2018).



Supplementary figure 1 • Quantitative measurements of scDam&T-seq across conditions

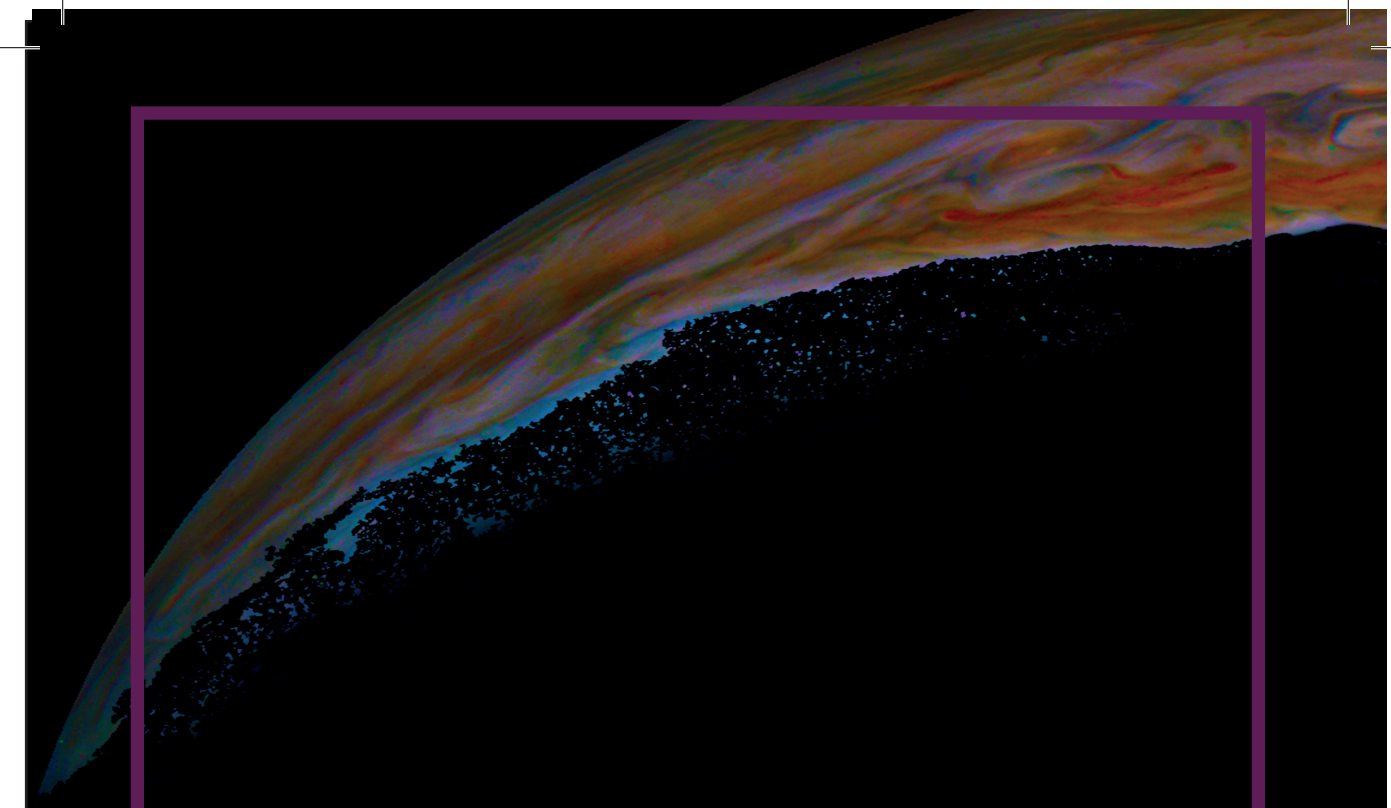
a. Violin plots showing number of unique genes per electroporation condition for GFP- or mRuby-positive cells. mRuby negative control cells are not shown. Dashed lines indicate the maximum and minimum threshold of 6,000 and 200 unique genes, respectively. Boxes in the violin plots indicate mean value ± 1 the SD. **b.** IUE conditions projected on UMAP shown in Fig. 2a. **c.** Unique GATCs (log_{10}) projected on UMAP shown in Fig. 2a. **d.** Unique transcripts (log_{10}) projected on UMAP shown in Fig. 2a. **e.** Fluorescent positive or negative cells projected on UMAP shown in Fig. 2a. **f.** Violin plots showing enrichment of known marker genes per cell cluster. Only cells passing the CEL-Seq2 threshold of >200 and <6,000 unique genes are used.



Supplementary figure 2 • LADs in EF1α-DL-expressing BPs and Ns

a. RPKM-normalized DamID signal per bin of chromosome 8 for four single cells of the BP cluster in the EF1α-DL IUE condition. Top track, RPKM-normalized DamID signal per bin averaged over all cells of the BP cluster for cells of the EF1α-DL IUE condition. Red boxes indicate variable regions between the single cells. **b.** RPKM-normalized DamID signal per bin of chromosome 8 for four single cells of the N cluster in the EF1α-DL IUE condition. Top track, RPKM-normalized DamID signal per bin averaged over all cells of the N cluster for cells of the EF1α-DL IUE condition. Red box indicates a variable region between the single cells. For panels a-b, cells shown are within the top 10 cells with the highest counts of unique GATCs per cell of the respective cluster. Only cells passing the CEL-Seq2 threshold of >200 and <6,000 unique genes and the DamID threshold of >25,000 unique GATCs are shown. Bin size is 100 kb. i, Illumina index; BC, unique DamID barcode.

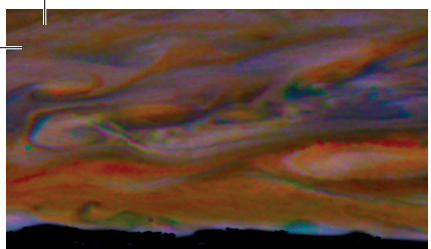
4



Chapter 5

Discussion

Corina M. Markodimitraki



Abstract

This thesis addresses the need for technical advancements in the field of single-cell genomics and genome organization. We do this by introducing an integrated single-cell method for protein-DNA and transcriptional measurements named scDam&T-seq in Chapter 2 as a proof-of-concept. The scDam&T-seq protocol is presented in detail in Chapter 3 along with modified versions based on the original scDamID method. In Chapter 4 we apply scDam&T-seq to the developing mouse cortex, thereby introducing a strategy that allows the study of genome organization in vivo.

5

scDam&T-seq

Method recap and novelty

In **Chapter 2** we introduce scDam&T-seq, a single-cell method to simultaneously measure protein-DNA contacts and transcriptomes without separating the nucleic acids. This method offers the possibility to understand how dynamic interactions between proteins and the DNA affect the transcriptome within the same cell. We used mouse embryonic stem cells (mESCs) expressing the fusion construct Dam-LmnB1, which results in the tethering of the Dam protein to the nuclear periphery. We identified regions in the genome that contact the nuclear lamina (NL) named lamina associated domains (LADs), at a single-cell resolution. We next studied the transcriptomes of the cells in relation to the positioning of their genome in the nucleus.

Main findings

To interpret results of our study it is important to clearly define the term contact frequency (CF). A genomic location *A* can localize at the NL or at the nuclear interior in any of the cells within a pool. If locus *A* contacts the NL in most of the cells, we can say that it frequently associates with the NL, categorizing it as a high CF locus. If, on the other hand, locus *A* is located in the nuclear interior more often and associates with the NL in just a handful of cells, then we would call this a low CF locus. We showed that high CF genomic regions displayed overall lower transcription levels than low CF regions, adhering to results of a previous study[1]. However, until now it was impossible to directly attribute transcriptional activity to nuclear positioning in the same cell. We addressed this question by applying scDam&T-seq on mESCs and obtained simultaneous LAD and transcriptome measurements. The results revealed an overall genome-wide trend of lower transcriptional activity for genomic regions in contact with the NL versus regions not contacting the NL in single cells. Regions most affected by nuclear positioning were regions of low CF. These regions showed a greater fold change in expression levels in relation to differential nuclear positioning (this is discussed further in the section "Biological outlook" of this discussion). Thus, the results of our study bring forward new insights into the relationship between NL association and transcription.

Technical outlook: Improving scDam&T-seq

Tight control of Dam fusion protein expression

One of the main characteristics of the DamID technology is that it allows the detection of accumulative signal of DNA-protein contacts. This can be considered an advantage in the case of transient protein-DNA contacts which are difficult to detect by antibody-based methods. The flipside is that approaches aimed at directly linking cause and effect can be complicated to interpret. For example, the effect we measured on transcription due to genome-NL contacts would probably be more pronounced had the induction time for the Dam-LmnB1 protein been shorter than the 12 hours used in our experiments. One way to tackle this would be to couple Dam to the endogenous LmnB1, by CRISPR-Cas9-mediated knock-in. This way, any competition between exogenous Dam-LmnB1 and the endogenous LmnB1 locus would be avoided, possibly contributing to a faster deposit of the m6A mark. Such an endogenous knock-in approach could be adopted for any other protein. The additional benefit of coupling Dam to an endogenous protein is the continued expression during differentiation. Experiments in our laboratory have shown that mESC lines with a virally integrated Dam-fusion protein become silenced, possibly through epigenetic mechanisms. In other words, expressing a Dam-fusion protein endogenously could be advantageous on many levels.

Dam mutants for reduced background methylation

Unlike Dam-LmnB1, using DamID for untethered proteins remains a challenge. Contrary to Dam-

LmnB1 which is tethered to the nuclear periphery, freely diffusing Dam-fusion protein can methylate the genome non-specifically, often in accessible regions, which leads to high background signal. To correct for this, an unfused Dam protein control is typically used. However, this cannot be expressed in the same cell as the Dam fusion protein, providing no internal control. The generation of a separate cell line expressing the unfused control Dam protein can be painstaking as it brings along time-consuming optimization experiments such as induction time and expression level titrations. An alternative strategy is offered by the targeted mutagenesis of the Dam enzyme[2]. Dam mutants such as R116A or N126A coupled to the transcription factor Tcf7l2 retained their capacity to methylate GATCs in mESCs but produced considerably less background signal than the wild-type Dam in bulk experiments[2]. Such an approach would streamline the use of scDam&T-seq for proteins that freely diffuse in the nucleus and euchromatin-interacting proteins.

Nascent RNA detection in differentiation

Application of scDam&T-seq in its current form might not be best suited for a fast-evolving system such as a differentiation trajectory, because some measured transcripts might no longer represent the transcriptional state of the cell at the moment of collection. Measurement of nascent RNA synthesis could elevate the current temporal resolution of mRNA detection. Nucleotide analogs such as the uridine analog 5-ethynyluridine (EU) are used to label nascent RNA, which are then pulled down with biotin beads after a copper-mediated click reaction[3]. Recent work in single cells shows that enrichment for both labelled (nascent) and unlabeled (mature) transcripts is possible[4, 5]. This approach would circumvent the capture of mRNA molecules with a long half-life that do not represent the state of the cell at the moment of collection. It would also allow to draw more concrete conclusions about cause and effect. Alternatively, the ratio of measurements of unspliced (nascent) to spliced (mature) transcripts could unveil both present and future transcriptional states[6]. This would be especially interesting to apply to in vivo development, as was shown previously on the developing mouse hippocampus[6].

Technical outlook: Expanding the repertoire of scDam&T-seq

Investigating LADs and lineage acquisition with scDam&T-seq and L122A

Exciting biological questions could be answered with scDam&T-seq or adaptations thereof. A possible expansion of the protocol could come from the use of the Dam mutant L122A, which recognizes hemi-methylated GATCs and deposits the m6A mark on the unmethylated strand[7, 8]. Thus, Dam L122A can re-establish fully methylated GATC sites after replication and maintain them in daughter cells. One could use this property of the L122A enzyme to explore the contribution of a protein of interest to lineage acquisition during differentiation. For example, to study LADs, one could explore how pluripotent cells contribute to cell clonalities further in development. Do certain LAD signatures accompany the development of specific lineages? To explore this, Dam-LmnB1 expression could be induced for a certain period of time during development and shut down again, with the help of an inducible system for fast protein degradation[9, 10]. This way, only during this time period would Dam-LmnB1 mark genomic regions close to the NL. Next, one could induce expression of the L122A Dam which would maintain the m6A methylation patterns established in the founder cells throughout their progeny. Cells could be collected after a period of time and processed with scDam&T-seq. Transcriptional information could be used to assign cell types while the DamID data could reveal if the different cell types and lineages reveal LAD signature-specific clonalities. Such an approach could elucidate the potential role of genome organization in cell fate choice. LADs have recently been studied in the mouse pre-implantation embryo[11]. By mRNA injection of an auxin-degradable Dam-LmnB1 construct and short-term culture under auxin conditions,

m6A methylation of LADs was tightly controlled. A co-injection of the L122A Dam construct with the piggyBac transposase would ensure a stable genomic integration of the mutant and its expression even after multiple cell divisions[12]. This system would be suitable for collection of stages viable in vitro. Alternatively, one could use mESCs expressing Dam-LmnB1 and L122A Dam to in vitro differentiate them into synthetic embryos termed gastruloids. These cellular aggregates represent the gastrulation phase of embryogenesis[13] and they circumvent the need for a transgenic animal.

-"What do we want?" -"More measurements!"

The advance of omics sequencing technologies demonstrates increasing interest in the integration of multiple measurements. An example is scNMT-seq which measures chromatin accessibility, transcription and 5mC in one cell[14]. Similarly, scDam&T-seq can be integrated to additionally measure modifications such as 5-cytosine methylation (5mC) or 5-hydroxy-cytosine methylation (5hmC) with the use of restriction enzymes sensitive to DNA modifications. One example is the 5hmC-sensitive endonuclease AbaSI, which can be used for lineage reconstruction by inferring DNA strand bias[15]. Another option is the 5mC-sensitive enzyme MspJI. Assays based on such enzymes could be integrated into the scDam&T-seq pipeline by ligation of universal T7 adapters and in silico separation of reads based on the restriction site sequence. Such a multiomics technique could be used in combination with in utero electroporation (IUE) as described in **Chapter 4**. During corticogenesis, progenitors undergoing mitosis on the apical size of the ventricular zone (VZ) (apical progenitors; APs) give rise to the neuronal layers that make up the adult cortex, which consists of neurons, astrocytes and oligodendrocytes. Successive waves of 5mC genome demethylation in these progenitors coincide with stages of neurogenesis, astrogenesis and oligodendrogenesis[16]. Measurements of LADs, 5mC and transcription might shed light into the role of these regulatory mechanisms and perhaps their interplay. Conditional depletion of enzymes such as the 5mC maintenance DNA methyltransferase 1 (DNMT1), could offer additional insights. Such an experiment would require co-electroporations of a plasmid encoding for Dam-LmnB1 and a plasmid encoding for guide RNAs that target DNMT1, in a mouse ubiquitously expressing the Cas9 protein[17]

Measuring past and present states

It would be interesting to measure the past and present state of protein-DNA contacts in a cell, for example to study LADs which have been shown to interact with the NL in a dynamic way[1, 18]. As DamID relies on the accumulation of signal it could serve to detect past states, while antibody-based methods which offer a more "snapshot" view could uncover present states. This could be achieved if DamID were to be combined with another method for detecting protein-DNA contacts such as the Chromatin Immunocleavage Sequencing (ChIC-seq) technique which is applicable to small populations and single cells[19, 20]. This method maps protein-DNA contacts with antibodies which are recognized and bound by a proteinA-MNase fusion protein. MNase then digests the sequences around the antibody binding site. ChIC could be merged with DamID, as both rely on the amplification of free genomic ends. By ligating universal adapters containing T7 promoters to both the DamID and proteinA-MNase-cut free ends, one could theoretically detect both DamID and ChIC signal. Such an approach would not only be interesting to unveil past and present states of protein-DNA contacts such as NL interactions, but also to measure two separate proteins simultaneously. For example, one could look into the interplay between LADs and Polycomb-group regulation in this way.

Biological outlook

The power of scDam&T-seq

We demonstrated that scDam&T-seq uncovers chromatin states such as LAD architecture, chro-

matin accessibility and protein-DNA binding and directly links these to the transcriptional profile of the cell through an in silico sorting strategy. This aspect of the method allows the study of protein-DNA interactions in a complex tissue by cell type identification based on the transcriptome. Indeed, in **Chapter 4** we called cell types and identified progenitor cells and neurons in the developing murine cortex. Another system that could be studied in a similar manner is the pre-implantation embryo. Protein-DNA interactions could be measured by mRNA injection of a Dam-POI construct in a zygote or 2-cell stage embryo, as was done previously for LADs and chromatin accessibility[11]. Other interesting applications could include healthy or diseased human cells which could be reprogrammed into induced pluripotent stem cells (iPSCs) to study human stem cell regulatory networks. Finally, applying scDam&T-seq on gastruloids could be interesting[13] as this system could be used for compound or knock-out screenings. Overall, scDam&T-seq could be applied to a range of systems to relate protein-DNA contacts with transcription.

Contact frequencies and pluripotency

Single-cell LAD profiles have thus far been obtained from mESCs or human cancer cells, showing similar trends in CFs[1, 21]. To some extent, cancer cells and stem cells are alike in that both can display genomic plasticity and this ensures the differentiation potential of the stem cells[22, 23]. Upon commitment of an embryonic stem cell to a cell fate however, the genome undergoes dramatic changes. Research addressing the reorganization of LADs during differentiation has been limited to the population level, providing no insight into single-cell NL associations[24]. This aspect of LAD biology is interesting because multiple studies show that the genome undergoes changes in compaction upon differentiation of pluripotent cells. Differentiation of mESCs into neural precursor cells is accompanied by chromatin condensation at the NL[25]. Additionally, the constitutive heterochromatin mark H3K9me3 concentrates at the periphery of the nucleus and intra-topological associated domain (TAD) interactions and interactions between B-compartments grow stronger[26-28] . We could hypothesize that LADs, in a similar fashion, become more defined and robust, displaying different CFs than in mESCs and possibly less cell-to-cell variation. In support of this hypothesis, results from DamID experiments on an in vitro differentiation model showed that the dynamic range with which the genome interacts with the NL in mESCs is lower than in neural progenitors and astrocytes[24]. This hints towards more single cell variation in NL association in pluripotent states than in differentiated cells, although technical explanations in this study were not ruled out. To investigate CFs in different states of pluripotency, one could push mESCs into a differentiated state and collect cells at different time points of the trajectory. By using scDam&T-seq transcriptome data, the differentiation stage could be assigned, and one could look at LADs in different points of fate acquisition. Such experiments would reveal how the organizational changes at the NL shown in previous studies, translates to NL associations in single cells.

Low CF regions and alternative silencing mechanisms

We showed that NL association results in gene repression, especially for regions often residing in the nucleoplasm. The effect size we observed however, was small. It could be explained by a combination of technical shortcomings (discussed above) as well as biological processes. Biologically, why would some loci, but not others, be more prone to transcriptional shutdown upon contact with the NL? It is possible that certain characteristics of these regions, such as their epigenetic landscape, deem them sensitive to NL contact. In particular, histone marks have been linked to LADs and their anchoring at the NL, as discussed in the introduction of this thesis, and are therefore interesting for follow-up studies. Intriguingly, our results showed a decreased occupancy of the constitutive heterochromatin mark H3K9me3 for regions localizing more often at the nucleoplasm. In contrast, they harbored the H3K27me3 modification, which marks facultative heterochromatin and which is

regulated by the Polycomb group complex. These results are in line with previous research on human cancer cells [1]. It is tempting to hypothesize an interplay for transcriptional repression between Polycomb regulation and NL association. Could less frequent localization at the nuclear periphery, and thus lower transcriptional repression, be compensated by an alternative silencing mechanism such as the Polycomb complex?

Polycomb regulation and LADs

Whether NL tethering and the Polycomb complex act in a complimentary way to silence genes could be further explored. We could use our dataset to look into the regions sensitive to NL association. Are they enriched for lineage-specific genes for example, known to be enriched for Polycomb proteins in mESCs[29, 30]? Further experiments could provide more mechanistic insight. For example, one could perturb Polycomb regulation in mESCs by depleting writers of the H3K27me3 mark such as the EZH2 protein, a subunit of the Polycomb group complex. If the two mechanisms act in a complimentary way, one might expect regions now depleted of Polycomb regulation, moving to the NL to be silenced. Or, reversely, depletion of NL components such as the Lamin B receptor (LBR) might result in a more active Polycomb regulation. How do these perturbations correspond to NL contacts? Questions like these could also be examined by studying LADs with scDam&T-seq in a differentiation system. Alternatively, DamID for LAD detection could be combined with scChIC (described in the section "Technical outlook" of this discussion) for Polycomb proteins or H3K27me3 to reveal any interplay between the two within the same cell.

LADs & NADs

Another direction that could be explored is the role of the nucleolus in genome anchoring. Similar to the NL, the genome also interacts with the nucleolus, a nuclear compartment where ribosome biogenesis takes place[31, 32]. Interestingly, LADs and nucleolus-associated domains (NADs) seem to overlap, and after mitosis, LADs are reshuffled and can return to the nuclear lamina or move to the nuclear interior and in some cases contact the nucleolus[31-34]. Perhaps the two nuclear compartments complement each other in order to keep the DNA away from active RNA polymerase II transcription hubs, possibly due to limited storage space within each individual compartment. Such a hypothesis could be tested by measuring both NADs and LADs in the same cell with a combination of scDamID for LmnB1 contacts with scChIC for a nucleolar protein such as nucleophosmin (NPM) to map nucleolus-associated regions. Such a combinatorial technique could also be implemented in perturbation experiments like depletion of laminar proteins. Would LADs reshuffle to the nucleolus, showing a "rescue" phenotype? An exciting system to study is the pre-implantation embryo as the nucleolus has a distinctive organization compared to other cell types (reviewed in [35]).

Contact frequencies & the nature of the interaction with the NL meshwork

One could compare the structure of the protein meshwork lining the inner nuclear membrane in mammals with that of a bird's nest. The laminar proteins Lamin A/C and B1, B2 are interwoven to create the filamentous structure[36]. To further investigate NL association and transcriptional repression, one could attempt to link them to the manner in which the genome associates with the NL. Data produced with super-resolution microscopy indicate that LADs are embedded in the lamin meshwork in HT1080 human fibrosarcoma cells[1]. Additionally, histone marks have been associated to NL association. In C. elegans for example, H3K9me3 mediates interaction of the genome with the NL through the CEC-4 protein[37, 38]. A similar mechanism in mammals is not unlikely. It is intriguing to hypothesize that low CF regions, devoid of H3K9me3 -and possible mediator proteins- could interact with the NL meshwork in a different manner than high CF regions enriched

for H3K9me3. Insight into these kinds of questions could be gained by CRISPR-Cas9-mediated knock-outs of laminar components such as LBR, LmnA or LmnB1/B2 and screening of the cells for altered contact frequencies and disrupted transcriptional repression. Thus, scDam&T-seq could be employed to understand the nature of the interaction between the genome and the NL.

MAPPING LADs IN DEVELOPING BRAIN

Method recap and novelty

In **Chapter 4** we applied scDam&T-seq to identify LADs in mouse cortical development. We in utero electroporated the motor cortex using vectors with promoters of different strengths to titrate the expression levels of Dam-LmnB1 and applied scDam&T-seq on the isolated tissue. The most suitable condition was chosen based on the ability to detect LADs, background m^6A methylation and single-cell success rate. We detected LADs in all conditions and called cell types based on the scDam&T-seq transcriptome data. To our knowledge, this is the first report linking LADs to specific cell types in single cells in vivo.

Main findings

Cell type-specific genome architecture and protein-DNA interactions in development have been studied in bulk or single cells in vivo. One study investigated A/B compartments and TADs by isolating cortex cells expressing fluorescently tagged markers[28]. Bulk DamID experiments in developing Drosophila brain identified binding sites of the Polycomb and Trithorax group complexes as well as HP1 by using cell type-specific promoters expressing the Dam fusion proteins[39]. Seminal work on mouse-preimplantation development mapped LADs in single cells of 2-cell and 8-cell stage embryos[11]. These approaches however, have several drawbacks such as not including transcriptional information, which is vital to identify intermediate states within a rapidly developing system. Additionally, in two of the three studies, measurements were derived from population averages, disregarding variability between single cells. We successfully targeted the motor cortex in developing embryos with Dam-LmnB1-coding vectors and performed scDam&T-seq. We identified APs, basal progenitors (BPs) and neurons and revealed genome-wide LAD profiles for each cell type. The single-cell resolution allowed us to observe heterogeneity in LAD signatures particularly within the AP cluster. This is the first report of a genome organization study in single cells in a living organism, paving the way for the study of other protein-DNA interactions as well.

Technical outlook

What about other lineages?

We mapped LADs by expressing the Dam-LmnB1 constructs for 48 hours starting on developmental day 14 (E14). We thereby identify neurons and BPs born during that time period, but are limited to this specific phase of development because vectors remain episomal and are eventually lost upon cell division. To investigate later developmental phases such as gliogenesis, during which astrocytes, ependymal cells and oligodendrocytes are born, we would need to electroporate at a later time point. An alternative strategy would be to make use of the *piggyBac* plasmid transposon system which offers genomic integration of a transgene of interest by using the *piggyBac* transposase (PBase)[40]. This would allow the labelling of multiple neural progenitor lineages and an extension of the collection time point.

Biological Outlook

LADs and lineage acquisition

Neural progenitor identity progresses during development, passing through successive phases of self-renewal, neurogenesis and gliogenesis[41]. Could genomic nuclear positioning be indispensable for the fate changes in neural stem cells? Studies in the fruit fly indicate that genomic nuclear positioning is key for faithful lineage acquisition. During Drosophila embryogenesis, neural stem cells termed neuroblasts produce neurons by transitioning through consecutive phases of neurogenic potential. They initially generate motorneurons after which they switch to producing interneurons and this change is facilitated by the repression and NL repositioning of the gene coding for the transcription factor (TF) *Hunchback* (*hb*)[42]. Depletion of Lamin results in insufficient silencing of the *hb* gene, thereby extending the motorneuron competence window[42]. The mammalian homolog of *hb*, the protein Ikaros (Ikzf1), is similarly expressed in APs during early corticogenesis and contributes to the generation of deep layer neurons which are born between embryonic day (E) 11.5 and 13.5. Studies have shown that nuclear localization of Ikzf1, regarded also a key B cell developmental gene, is distinct between cell types. Ikzf1 was found to reside in the nuclear interior in primary pro-B-cells while in mouse embryonic fibroblasts it associated with the NL, suggesting a cell-type specific localization of this locus[43]. It would therefore be interesting to study LADs at an earlier developmental time point either by in utero electroporating a Dam-LmnB1 vector or by constructing a transgenic mouse expressing Dam-LmnB1.

scDam&T-seq for study of epigenetic landscapes in brain development

Epigenetic regulation is pivotal for normal cortical development as writer, eraser and reader proteins regulate key developmental genes by catalysing the addition or removal of histone marks and recognising them, thereby initiating downstream regulatory mechanisms. The Polycomb repressive complex 2 (PRC2) for instance, represses the promoter of proneuronal gene Neurogenin 1 (Neurog1), thereby shifting the AP competence towards gliogenesis[44]. Upon deletion of the PRC2 subunit Enhancer of Zeste homolog 2 (EZH2), the methyltransferase catalysing the H3K27me3 mark, deep layer neuron production is increased, extending neurogenesis and delaying astrogenesis[44]. Conditional knockout of EED, another PRC2 component, causes insufficient deep layer neuronal production and premature production of superficial layer neurons, resulting in a thinner ventricular zone (VZ)[45]. As LAD borders are enriched for H3K27me3 in certain cell types[43, 46] it would be interesting to study the interplay between Polycomb regulation and genome organization at the NL in a normal or perturbed developmental trajectory. This could be achieved by IUE of vectors coding for gRNAs against PRC2 subunits as well as Dam-LmnB1 in a Cas9-expressing mouse[17]. Applying scDam&T-seq afterwards could reveal if LAD organization is affected by loss of PRC2 regulation. The heterochromatic marks H3K9me2 and H3K9me3 are enriched in LADs and the latter becomes progressively enriched at the nuclear periphery in more differentiated cell types [27, 46]. In a similar manner as suggested for the PRC2 complex subunits above, it would be interesting to investigate how the H3K9me2/3-occupied heterochromatin and LADs relate during cortex development. Conditional knockouts of the G9a or SUV39H1 proteins catalyzing the H3K9me2 and H3K9me3 mark respectively, and application of scDam&T-seq could provide more insight into this.

LADs, LINEs and somatic mosaicism in brain

Brain neurons in several mammals including primates and mouse display genomic diversity and a source for this somatic mosaicism is the elevated activity of retrotransposable mobile genomic elements in neural precursor cells. Long interspersed nuclear elements (LINEs) can retrotranspose to new genomic locations and in neural progenitor cells they often do so into neuronal genes, as they

are active and in an open chromatin conformation[47]. Retrotransposition can impact local gene activity, RNA splicing and premature polyadenylation of the RNA[47]. In neurons, this could for instance influence synaptic activity or stimulus. Constitutive LADs (cLADs) overlap between cell types and are considered to make up a structural backbone, maintaining the three-dimensional organization of the interphase genome[24, 48]. cLADs display high NL contact frequencies and are characteristically richer in LINEs than facultative LADs (fLADs) which are more cell-type specific[1, 48]. Do cLADs in neuronal cells show similar characteristics? Or do they lack LINEs? If LADs do contain LINEs in brain neurons, it is tempting to investigate the nature of the genome-NL interactions.

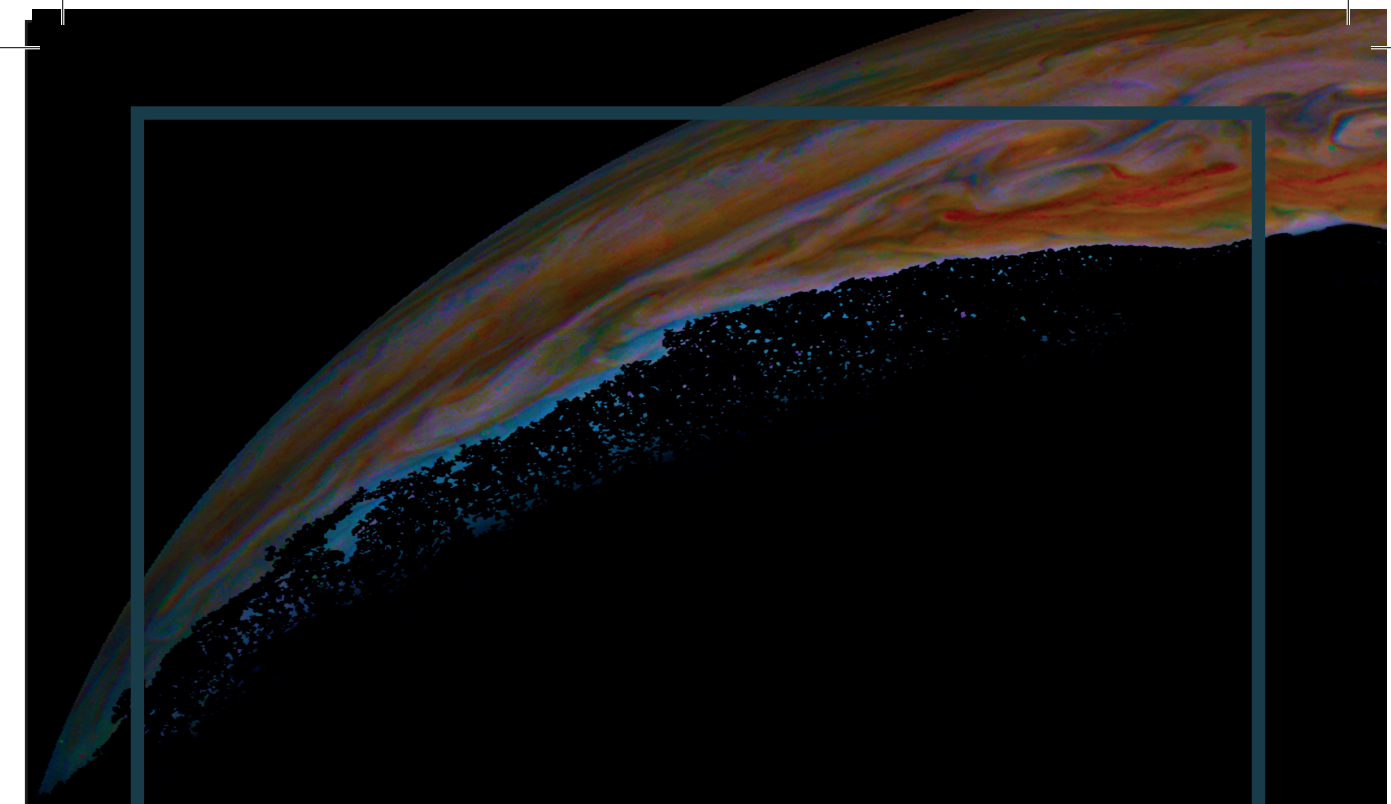
SUMMARY & CLOSING REMARKS

The work presented in this thesis is part of a broader scientific trend towards integration of multiple measurements in single cells. We introduce the single-cell molecular tool scDam&T-seq which quantifies protein-DNA interactions and directly links them to transcriptional outcome. Attempting to address a longstanding question in the field of genome organization about the effect that perinuclear localization of the genome has on transcription, we measured LADs and mRNAs in mESCs, revealing an overall trend of lower transcriptional output upon NL association, an effect whose size however, differed amongst genomic regions. We further discovered that scDam&T-seq could be employed to uncover patterns of chromatin accessibility and to map binding domains of the PRC1 complex in differentiation. We complemented this work by applying scDam&T-seq to the developing mouse cortex, and linked LAD signatures to specific cortical cell types, thus demonstrating that our method can be used in vivo. I hope that the tools presented in this thesis can help others further along the search for answers and be a source of knowledge for the development of novel technologies.

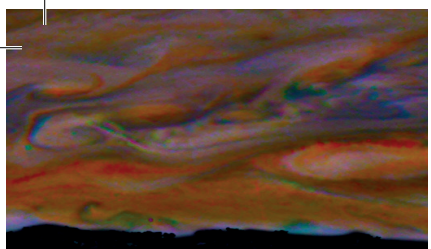
REFERENCES

1. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134-147 (2015).
2. 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).
3. Jao, C.Y. & Salic, A. Exploring RNA transcription and turnover in vivo by using click chemistry. Proc Natl Acad Sci U S A 105, 15779-15784 (2008).
4. Battich, N. et al. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. Science 367, 1151-1156 (2020).
5. Hendriks, G.J. et al. NASC-seq monitors RNA synthesis in single cells. Nat Commun 10, 3138 (2019).
6. La Manno, G. et al. RNA velocity of single cells. Nature 560, 494-498 (2018).
7. Elsawy, H. & Chahar, S. Increasing DNA substrate specificity of the EcoDam DNA-(adenine N(6))-methyltransferase by site-directed mutagenesis. Biochemistry (Mosc) 79, 1262-1266 (2014).
8. Horton, J.R., Liebert, K., Bekes, M., Jeltsch, A. & Cheng, X. Structure and substrate recognition of the Escherichia coli DNA adenine methyltransferase. J Mol Biol 358, 559-570 (2006).
9. 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).
10. Nabet, B. et al. The dTAG system for immediate and target-specific protein degradation. Nat Chem Biol 14, 431-441 (2018).
11. Borsos, M. et al. Genome-lamina interactions are established de novo in the early mouse embryo. Nature 569, 729-733 (2019).
12. Suzuki, S., Tsukiyama, T., Kaneko, T., Imai, H. & Minami, N. A hyperactive piggyBac transposon system is an easy-to-implement method for introducing foreign genes into mouse preimplantation embryos. J Reprod Dev 61, 241-244 (2015).
13. van den Brink, S.C. et al. Symmetry breaking, germ layer specification and axial organisation in aggregates of mouse embryonic stem cells. Development 141, 4231-4242 (2014).
14. 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).
14. 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).
15. Sanosaka, T. et al. DNA Methylome Analysis Identifies Transcription Factor-Based Epigenomic Signatures of Multilineage Competence in Neural Stem/Progenitor Cells. Cell Rep 20, 2992-3003 (2017).
16. Platt, R.J. et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell 159, 440-455 (2014).
17. van Schaik, T., Vos, M., Peric-Hupkes, D. & van Steensel, B. Cell cycle dynamics of lamina associated DNA. bioRxiv, 2019.2012.2019.881979 (2019).
18. Ku, W.L. et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. Nat Methods 16, 323-325 (2019).
19. Skene, P.J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife 6 (2017).
20. Rooijers, K. et al. Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. Nat Biotechnol 37, 766-772 (2019).
21. Meacham, C.E. & Morrison, S.J. Tumour heterogeneity and cancer cell plasticity. Nature 501, 328-337 (2013).
22. Sanchez Alvarado, A. & Yamanaka, S. Rethinking differentiation: stem cells, regeneration, and plasticity. Cell 157, 110-119 (2014).
23. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell 38, 603-613 (2010).
24. Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res 20, 155-169 (2010).
25. Ahmed, K. et al. Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. PLoS One 5, e10531 (2010).
26. Ugarte, F. et al. Progressive Chromatin Condensation and H3K9 Methylation Regulate the Differentiation of Embryonic and Hematopoietic Stem Cells. Stem Cell Reports 5, 728-740 (2015).
27. Bonev, B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557-572 e524 (2017).
28. Schuettengruber, B., Bourbon, H.M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell 171, 34-57 (2017).
29. Kloet, S.L. et al. The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. Nat Struct Mol Biol 23, 682-690 (2016).
30. Nemeth, A. et al. Initial genomics of the human nucleolus. PLoS Genet 6, e1000889 (2010).
31. van Koningsbruggen, S. et al. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol Biol Cell 21, 3735-3748 (2010).
32. Kind, J. et al. Single-cell dynamics of genome-nuclear lamina interactions. Cell 153, 178-192 (2013).
33. Vertii, A. et al. Two contrasting classes of nucleolus-associated domains in mouse fibroblast heterochromatin. Genome Res 29, 1235-1249 (2019).
34. 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).
35. Gesson, K., Vidak, S. & Foisner, R. Lamina-associated polypeptide (LAP)2alpha and nucleoplasmic lamins in adult stem

cell regulation and disease. Semin Cell Dev Biol 29, 116-124 (2014).
36. Towbin, B.D. et al. Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. Cell 150, 934-947 (2012).
37. Gonzalez-Sandoval, A. et al. Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. Cell 163, 1333-1347 (2015).
38. Marshall, O.J. & Brand, A.H. Chromatin state changes during neural development revealed by in vivo cell-type specific profiling. Nat Commun 8, 2271 (2017).
39. Chen, F., Maher, B.J. & LoTurco, J.J. piggyBac transposon-mediated cellular transgenesis in mammalian forebrain by in utero electroporation. Cold Spring Harb Protoc 2014, 741-749 (2014).
40. Mukhtar, T. & Taylor, V. Untangling Cortical Complexity During Development. J Exp Neurosci 12, 1179069518759332 (2018).
41. Kohwi, M., Lupton, J.R., Lai, S.L., Miller, M.R. & Doe, C.Q. Developmentally regulated subnuclear genome reorganization restricts neural progenitor competence in Drosophila. Cell 152, 97-108 (2013).
42. 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).
43. Hirabayashi, Y. et al. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. Neuron 63, 600-613 (2009).
44. Telley, L. et al. Temporal patterning of apical progenitors and their daughter neurons in the developing neocortex. Science 364 (2019).
45. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948-951 (2008).
46. Erwin, J.A., Marchetto, M.C. & Gage, F.H. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci 15, 497-506 (2014).
47. 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).



Addendum &



Samenvatting

Ons lichaam bestaat uit biljoenen cellen, die in alle soorten en maten komen. Het zijn de bouwstenen die met elkaar samenwerken om een volledig functionerend organisme te vormen. Een neuron heeft een karakteristiek boomachtig uiterlijk en is gespecialiseerd in het doorgeven van elektrische pulsen om bijvoorbeeld pijn te signaleren of beweging te sturen. Vetcellen (of adipocyten) zijn daarentegen rond en hun primaire functie is vetopslag. Deze twee celtypes, evenals de rest van de 200 verschillende celtypes in het menselijk lichaam, bevatten dezelfde genetische informatie, maar zijn functioneel en morfologisch verschillend. Hoe leidt dezelfde DNA-code tot verschillende uitkomsten? Het draait allemaal om blootstelling.

Niet alle DNA-code is te allen tijde beschikbaar zodat de cel deze kan lezen en erop kan reageren. Sommige delen van het DNA kunnen dichtgepakt (opgevouwen) zijn, terwijl andere meer loszitten. Die delen zijn gemakkelijker te lezen door de cel. Wanneer de cel de genetische code "leest", vermenigvuldigt die deze regio, waardoor zogenaamde messenger-RNA (of mRNA) transcripten ontstaan. De cel gebruikt de mRNA transcripten als blauwdrukken om eiwitten te produceren. Een neuron heeft andere eiwitten nodig dan een adipocyt, wat betekent dat de toegankelijke DNA-delen en de resulterende mRNA-transcripten verschillen tussen de twee celtypen. Door de mRNA-transcripten van een cel te meten, kan een promovendus dus begrijpen naar welk celtype ze kijkt.

Het DNA wordt opgeslagen in een klein bolvormig compartiment van de cel, de cel kern. Nabij het midden van de kern is het DNA flexibeler en opener, terwijl het dichter is gepakt aan de rand van de kern. Onderzoek toont aan dat de manier waarop het DNA in de kern wordt gevouwen (dichtgepakt of open) belangrijk is voor de celfunctie en morfologie. Ten eerste omdat de regio's die zich aan de periferie van de kern bevinden en de regio's die zich in het midden van de kern bevinden, kunnen verschillen tussen celtypen. Ten tweede omdat de DNA-code die de cel nodig heeft, zich vaak in het midden van de kern bevindt waar deze wordt "gelezen". Daarentegen wordt code die niet nuttig is voor de celtaken, "opgeslagen" aan de periferie van de kern en blijft deze meestal ongebruikt.

In mijn proefschrift heb ik mij gericht op het onderzoeken van de relatie tussen DNA vouwing en mRNA productie. Over het algemeen blijkt uit onderzoek dat DNA aan de periferie van de kern niet door de cel wordt gelezen en bijna geen mRNA produceert. Tot nu toe was dit echter niet bevestigd. Om de 3D DNA vorm te koppelen aan de mRNA productie, moet men beide dingen in dezelfde cel meten. Tot nu toe ontbrak zo'n techniek. In ons lab hebben we de methode scDam&T ontwikkeld die zowel DNA vouwing als mRNA in dezelfde cel meet. Onze metingen bevestigden dat de meeste DNA regio's niet door de cel worden gelezen als ze zich aan de rand van de kern bevinden. We ontdekten ook dat sommige regio's meer geneigd zijn om "uitgezet" te worden zodra ze zich aan de nucleaire periferie bevonden dan andere regio's. De reden hiervoor blijft open voor onderzoek. We hebben scDam&T-seq vervolgens toegepast op de hersenen van ontwikkelende muizen om te begrijpen hoe cellen van dit complexe orgaan hun DNA vouwen. We ontdekten dat de cellen die de neuronen in de hersenen produceren verschillen in DNA vouwing vertonen, ook al lijkt hun mRNA vergelijkbaar te zijn. Nader onderzoek zou kunnen uitwijzen waarom dit het geval is. In conclusie, dit proefschrift presenteert voornamelijk technische oplossingen op het gebied van de biologie wat DNA architectuur bestudeerd. Nu is het mogelijk om de relatie tussen DNA-vouwing en mRNA-productie te bestuderen en om deze techniek toe te passen op een levend dier.

Summary

Our bodies are composed of trillions of cells, coming in all forms and shapes. They are the building blocks that cooperate with each other to form a fully functioning organism. A neuron has a characteristic tree-like appearance and is specialized in transmitting electrical pulses to signal pain or to command movement for example. Fat cells (or adipocytes) on the other hand, are round and their primary function is fat storage. These two cell types, as well as the rest of the 200 different cell types in the human body, contain the same genetic information, yet are functionally and morphologically different. How does the same DNA code result in different outcomes? It is all about exposure.

Not all DNA code is available for the cell to read and act upon at all times in all cell types. Some parts of the DNA can be tightly packed (folded), while others are more lose and those parts are easier for the cell to read. When the cell "reads" the genetic code, it multiplies this region, creating so-called messenger RNA (or mRNA) transcripts. The cell uses the mRNA transcripts as blueprints to produce proteins. A neuron needs different proteins than an adipocyte, which means that the accessible DNA parts and the resulting mRNA transcripts are different between the two cell types. Therefore, by measuring the mRNA transcripts of a cell, a PhD student can understand what cell type she is looking at.

The DNA is stored in a spherical compartment in the cell called the nucleus. Near the center of the nucleus, the DNA is more flexible and open, while at its periphery, the DNA is more densely packed. Studies show that the manner in which the DNA is folded (open or packed) in the nucleus is important for cell function and morphology. Firstly, because the regions that locate at the periphery and the ones residing at the nuclear interior can differ between cell types. Secondly, because the DNA code that is needed by the cell is often located in the center of the nucleus where it is "read". In contrast, code that is not useful for the cell's tasks is "stored" at the periphery of the nucleus, and mostly goes unused.

In my thesis I have focused on exploring the relationship between DNA folding and mRNA production. In general, evidence shows that DNA at the periphery of the nucleus is not read by the cell and produces almost no mRNA. However, until now, this was not confirmed because of the lack of molecular tools to do so. To link the DNA folding with the mRNA production, one has to measure both things in the same cell. In our lab we developed the method scDam&T-seq that measures both DNA folding and mRNA in the same cell. This allowed us to confirm that most DNA regions are not read by the cell when they are at the periphery of the nucleus. We also found that some regions are more prone to be "shut off" at the nuclear periphery than other regions. The reason behind this remains open for investigation. We also applied scDam&T-seq to the brain of developing mice to understand how cells of this complex organ fold their DNA. We found that the cells which produce the neurons in the brain show differences in DNA folding even though their mRNA seems similar. Further research could reveal why this is the case. In conclusion, this thesis presents mainly technical solutions to the field of DNA architecture. Now one can study the relationship between DNA folding and mRNA production and apply this technique to a living animal.

Περίληψη

Το σώμα μας αποτελείται από τρισεκατομμύρια κύτταρα, σε όλες τις μορφές και σχήματα. Είναι τα δομικά στοιχεία που συνεργάζονται μεταξύ τους για να σχηματίσουν έναν πλήρως λειτουργικό οργανισμό. Νευρικά κύτταρα για παράδειγμα, έχουν μια χαρακτηριστική εμφάνιση που μοιάζει με δέντρο και είναι ειδικευμένα στη μετάδοση ηλεκτρικών σημάτων που σηματοδοτούν πόνο ή που ελέγουν την κίνηση. Τα λιπώδη κύτταρα (ή τα λιποκύτταρα) από την άλλη πλευρά, είναι στρογγυλά και η κύρια λειτουργία τους είναι η αποθήκευση λίπους. Αυτοί οι δύο τύποι κυττάρων, καθώς και οι υπόλοιποι από τους 200 διαφορετικούς τύπους κυττάρων στο ανθρώπινο σώμα, περιέχουν την ίδια γενετική πληροφορία, αλλά είναι λειτουργικά και μορφολογικά διαφορετικοί. Πώς καταλήγει ο ίδιος κωδικός του DNA σε διαφορετικά αποτελέσματα; Όλα σχετίζονται με την έκθεση του DNA.

Ο γενετικός κώδικας του DNA δεν είναι διαθέσιμος στο κύτταρο ανά πάσα στιγμή ώστε να «διαβαστεί». Ορισμένες περιοχές του DNA είναι σφιχτά οργανωμένες, ενώ άλλες είναι πιο «χαλαρές» και αυτές οι περιοχές διαβάζονται ευκολότερα από το κύττάρο. Όταν το κύτταρο «διαβάζει» τον γενετικό κώδικα, πολλαπλασιάζει αυτήν την περιοχή, δημιουργώντας το λεγόμενο αγγελιοφόρο RNA (ή mRNA). Το κύτταρο χρησιμοποιεί τα μεταγραφή mRNA ως σχεδιαγράμμα για την παραγωγή πρωτεϊνών. Ένα νευρικό κύτταρο χρειάζεται διαφορετικές πρωτεΐνες από ένα λιποκύτταρο, πράγμα που σημαίνει ότι τα προσβάσιμα μέρη του DNA και τα προκύπτοντα αγγελιοφόρα mRNA είναι διαφορετικά μεταξύ των δύο τύπων κυττάρων. Επομένως, μια διδακτορική φοιτήτρια μπορεί να αναγνωρίσει τον τύπο του κυττάρου μετρώντας το mRNA του.

Το DNA αποθηκεύεται σε ένα σφαιρικό οργανίδιο μέσα στο κύτταρο που ονομάζεται πυρήνας. Κοντά στο κέντρο του πυρήνα, το DNA είναι πιο εύκαμπτο και ανοιχτό, ενώ στην περιφέρειά του, το DNA είναι πιο πυκνό. Μελέτες δείχνουν ότι ο τρόπος με τον οποίο το DNA διπλώνεται σε στον πυρήνα είναι σημαντικός για τη λειτουργία των κυττάρων και τη μορφολογία τους. Πρώτον, επειδή οι περιοχές που βρίσκονται στην περιφέρεια και εκείνες που κατοικούν στο εσωτερικό του πυρήνα μπορούν να διαφέρουν μεταξύ των διαφορετικών ειδών κυττάρων. Δεύτερον, επειδή μέρη του DNA που χρησιμοποιούνται από το κύτταρο βρίσκονται συχνά στο κέντρο του πυρήνα όπου «διαβάζονται». Αντίθετα, μέρη του DNA που δεν είναι χρήσιμα για τις εργασίες του κυττάρου αποθηκεύονται στην περιφέρεια του πυρήνα και συνήθως δεν χρησιμοποιούνται.

Στη διατριβή μου έχω επικεντρωθεί στη διερεύνηση της σχέσης μεταξύ της αναδίπλωσης του DNA και της παραγωγής του mRNA. Έως τώρα μελέτες έδειχναν ότι σε γενικές γραμμές, το DNA στην περιφέρεια του πυρήνα δεν διαβάζεται από το κύτταρο και δεν παράγει σχεδόν καθόλου mRNA. Ωστόσο, μέχρι τώρα, αυτό δεν είχε επιβεβαιωθεί. Και αυτό γιατί το να σχετίσει κανείς την αναδίπλωση του DNA με την παραγωγή mRNA, πρέπει να μετρήσει και τα δύο πράγματα στο ίδιο κύυταρο. Έως σήμερα, ωστόσο, δεν υπήρχε κάποια τεχνική που να επιτρέπει κάτι τέτοιο. Στο εργαστήριό μας αναπτύξαμε τη μέθοδο scDam&T-seq που μετρά τόσο την μορφολογία του DNA όσο και το mRNA στο ίδιο κύτταρο. Οι μετρήσεις μας επιβεβαίωσαν ότι οι περισσότερες περιοχές που βρίσκονται στην περιφέρεια του πυρήνα του DNA δεν διαβάζονται από το κύτταρο. Στη συνέχεια εφαρμόσαμε αυτήν την τεχνική στον εγκέφαλο αναπτυσσόμενων ποντικών για να κατανοήσουμε πώς τα κύτταρα του πολύπλοκου αυτού οργάνου διπλώνουν το DNA τους. Διαπιστώσαμε ότι τα κύτταρα που παράγουν τους νευρώνες στον εγκέφαλο αναδιπλώνουν το DNA τους διαφορετικά, παρόλο που το mRNA τους φαίνεται παρόμοιο. Εν κατακλείδι, η παρούσα διατριβή παρουσιάζει κυρίως τεχνικές λύσεις στον τομέα βιολογίας που ασχολείται με την αναδίπλωση του DNA. Τώρα μπορεί κανείς να μελετήσει τη σχέση μεταξύ αναδίπλωσης DNA και παραγωγής mRNA και να εφαρμόσει αυτήν την τεχνική σε ένα ζωντανό ων.

Acknowledgements

There are many people that shaped the period of my PhD, and I would like to thank them in this chapter. I hope my words can do my feelings justice! Here it goes.

Jop. After a few email exchanges you invited me for an interview in the fall of 2015. Two full days of talking and ideas, led to an acceptance email around Christmas. I was over the moon. I was so excited when I started in February 2016. Excitement was our overlap, excitement kept pushing me forward at the bench, at cell culture. Thanks for always making time when I needed advice. Thanks for putting together this group of incredible people that created a community, without which I would be lost. I learned a lot on this journey.

My promotor, **Alexander van Oudenaarden**. Thank you for input, especially for the last chapter of this thesis. I also appreciate the opportunity you gave me to do my internship in your lab during my masters program. It was detrimental to making it through the scDam&T project. My PhD committee, **Wouter de Laat** and **Frank Holstege**. Thank you for your support and your advice throughout the years. Your judgements helped shape this thesis. Your support helped me through difficult times. Thanks to my reading committee: **Wouter de Laat**, **Michiel Vermeulen**, **Rik Korswagen**, **Jeroen Bakkers** and **Francesca Mattiroli**. Thank you for critically going through this thesis. Michiel, special thanks for taking me in your lab during my first year of masters studies. It ignited my interest in epigenetics and gene regulation, which shaped the rest of my academic track.

Lekker met de meiden. I cannot believe how lucky I have been to be surrounded every day by all of these incredible people. You were my family for the past 4.5 years. **Kim**. Saving my a** since 2016. Thank you for listening to me, your advice and the times you comforted me. I will forever admire your incredible strength, perseverance, endless wisdom and sudden and unexpected bursts of happy dancing. Hope the gecko brings you good things. You more than deserve it! **Sara**. The queen of a ~~rudeness~~ directness, more Dutch than the Dutch. It took me a while to get used to it ("-How much did you pay for your haircut? -12 euro -I can see that"), but your heart-stopping "Marko's!" out of nowhere, your excellent cooking skills and dorky jokes made up for all that. **Bram**. Your jokes and broodje kroket were a comforting constant that first year. I'm happy you have found your place! **Isabel**. The big mama. One of the most modest and hard-working people out there. Your down-to-earth-ness, "nuchterheid" and hands-on attitude really helped me and kept me grounded. Thank you for your willingness to help out at any time, our unending conversations, the morning coffee ritual, your R scripts, your jokes! And thank you for your guest list at your wedding. I owe you a big one! **Sandra**. Cooler than cool. A multi-tasking, 900-mePCR-reactions-at-the-same-time, force of nature. I really enjoy all our conversations regarding movies, music, books, museums and events. Thank you for many things. For all your support and understanding during the difficult times, for your calmness, your listening, for laughing at my bad jokes and for making revisions a fun activity! Thanks for setting a standard for a way of living life, you're iconic. And thanks for being my paranymph! **Samy**. Τα σιγανά τα ποταμάκια να φοβάσαι. Indeed. Your incredible strength, diligence and self-reflection capabilities are admirable. Thank you for the spanakopites, the fun Greek(lish) chatting every day, the loud disagreements and for setting a shining example of what it means to set boundaries. Thank you for listening to me and for swearing (lab) and sweating (yoga and dancing) alongside with me. Thanks for the long conversations about the human psyche & the bit less sophisticated conversations during cell culture duty. Thanks for being my paranymph!

Silke (Efficient) Lochs. Queen of schedules & doodles & colourful socks. You're such an energetic and hands-on person, watching you run around the lab doing experiments was tiring and motivating at the same time. Thank you for co-creating our beautiful cell line that turned out to be a flop in the end, but you know-we had fun (pipetting competition excluded). Thank you for showing your vulnerable side, for your cute "Corinaki's", the bad parties and worse hangovers. I know you will get off the PhD boat safely! **Franka**. Lovely nerd, badass, the groups buffer. Thank you for always giving a helping hand, be it for revisions or for opening the door with me to two more scientific chapters. Thank you for your talent to make a protocols paper interesting, it was a real pleasure to work with you! Thank you for your honesty, openness and support, for always finding the silver lining and for showing no restraints in being your authentic self, it's truly something to look up to! **Leila**. Thanks for your kindness, your openness and for always being up for a conversation. **Tess**. It was hard to imagine that I would come across a person that talked with more enthusiasm and hand gestures than I did, but then you came along. Thanks for your insightful conversations about career paths and PhD tracks, for analysing all the different data that came your way, for the fun car rides when I was immobile with my broken leg! I will always think of you while downing a Mythos! **Koos**. Although we do not surf on the same wavelength, I want to thank you for the effort you put into our project, the tutorials and for the raw tofu "meatball" which I ate thinking it was a cookie! **Ramada**. Thanks for your happy presence around the lab, your readiness to help me with technical issues (am I that old?) and your unending knowledge on tv-series! You bring out this shoulder bouncing move I did not know I had in me! **Ellen**. Thanks for your no bulls**t attitude, your willingness to help and your surprising stories about your scooter adventures and chickens! **Chris**. Thanks for all the fun conversations and your kindness. I wish you, Marisol and Elias a great new start in Mexico! **Pim**. Although we had a tiny overlap, I know you are on your way to do great in your PhD. You are already on top of the knowledge, good luck with your project! The students: **Ioanna**. The quiet force. I had forgotten about the Greek scientific terms before you came along and started talking about πέψη and αλληλούχιση! I really enjoyed guiding you, watching you grow so fast as a scientist, that I almost could not keep up. Thank you for your patience during my dramatic moments, your calmness helped counterbalance them! **Roberto**, it was fun to have another person calling me Marko in the lab. Thanks for being a bright presence and for all the jokes! **Robin**, too bad we didn't overlap in the end. Good luck with the postdoc! Ork ork!

My **foster lab, the AvOs**, old generation and new. Thanks for accepting my happy skipping around your lab and chatting you up. Thanks for all the fun times (office ball throwing, corridor chair rolling, barbeques and many other things) and all the advice regarding science, fun and food. The veterans: **Lennart** (goeiemorgen zonnestraaltje! Thanks for all the advice regarding labwork and for the jokes), **Kay** (slap!), **Dylan** (Verplettersdrang en bostrollen! Thanks for the advice & grumpiness), **Chloe** (micro-seal), **JC** (hoezo?!), **Adi** (thanks for the karaoke moments), **CatalAnna** (your energy is out of this world, thanks for all the fun times during borrels!). The new wave: **Helena**, fierce kween, I love that we have mutual interests on all things feminism andother things. I love our conversations, you are awesome! I know you will do great in the PhD :D. **Freddy** γεια σου ρε! Thanks for your kindness, craziness and your hilariously inappropriate jokes. **Jake**. Thanks for sharing your wisdom regarding instant noodles and the musical zoom borrels. **Mike**. "-Good morning! -There's nothing good about it." 'Nough said. You look good hugging giant plants. **Joe** thanks for your Englishness! **Anna v.O.**, thanks for the hilarious stories! **Susanne**, thanks for the advice during the "laatste loodjes" and for your uninhibited expression of self. **Maria**, thanks letting me pull jokes on you, for the hang outs, the cuteness and the fun! **Buys**, thanks for the conversations about so many interesting subjects, I learn a lot from you! **Peter**, the not-so-German German, thanks for all your

advice and for the many conversations. **Vincent** for your calm presence in the lab. **Marloes**, thanks for our conversations and good luck with the stupid cells! You can do it gurl. **Francis**, thanks for the encouragement and for the hangouts. Your "let's do this!" attitude was really what helped me at the last part of the PhD. **Nune**, thanks for co-organising the Dino borrel! It was a lot of fun. **Jeroen**, thanks for the 100% sureness of your statements :P **Nico**, thanks for the bouldering sessions ("hight is not an excuse!") and your annoying jokes! **Marijn**, thanks for the courgettes! **Vivek** I will always laugh at you trying to expand your pc screen empire. **Judith**, although technically not part of the AvOs anymore but still you roamed around, pipetting wildly, in this lab in the first few years. Thanks for your nerdiness, the conversations, your helpfulness, and the sing-alongs! It is always fun to be around you. Special thanks to the one and only Zen Master, **Reinier**! Thank you for the many unending hours at the FACS together, for your patience and our conversations. Thank you for your calmness, it always brought me back to basics.

Team "mission Wiener schnitzel". **Abel** thanks for your truly amazing jokes, all the dinners and amazing parties (Hungarian cake dance), the shishas (conspicuous little suitcase), the hangouts, the running, the friendship! **Mauro**, thanks for the barbeques, all the "bdm-ts!", your down-to-earthness, your encouragement, for finding solutions where I only saw dead ends, the dinners, the laughter and the late-night sandwiches with cheese and bacon. I hope we can all repeat a "Vienna-like" excursion soon, it's always so much fun with you guys!

The Hubrecht community, making the atmosphere lighter (or heavier, depending on the PhD phase), but always a supportive one. **Sanne**, thank you for your friendship. I really enjoyed our gardening sessions where we could unleash our frustrations during weeding and we could talk to each other without inhibitions. I know you have the strength in you for the next step. Go for it! **Bas**, without you the garden sessions would not exist. Thanks to you I was able to escape the lab only a few meters further, swimming in a sea of courgettes! I realise now this sentence can be interpreted in many ways. Oh well. Your calm attitude and jokes always make it a pleasure to hang out together. Too bad our plans for a smashing PhD defence party drowned in the corona-wave but I am still glad we were co-pilots in these last months. **Spiro**, thank you for your excellent sense of humor, for inviting me to the original fightclub (I heard the word is out now, no rule breaking here) and for your openness in our conversations. For surprise-hosting me after the Christmas dinner and the fun suprise flights. As soon as we are allowed to, I shall join you for some crazy party moves, preferably next to Amsterdam canals! **Marta**, your cute cuddles, sweetness and openness are refreshing. I am so happy we can hang out in Amsterdam together now! Let's keep the fun going (twerk twerk twerk). **Geert** thanks for the many yoga sessions and for the nice conversations we always have, good luck in Rotterdam! **Anna**, thanks for the spontaneous fun camping trips and hangouts!

The rest of the Hubrecht TGIF/VMB borrel/picnics/barbeques/bouldering people that were always around and made the time at the Hubrecht a fun rollercoaster ride: **Tim** (it's always fun to have honest and open conversations with you!), **Lotte** (A for effort for your Halloween costumes and dance moves), **Gaby** (thumbs up! heh heh), **Stijn** (love our conversations about far-away lands and travel), **Deepak** (awesome dance moves), **Bram** (WAT ZEG JE?? Ik versta je niet!!), **Iris**, Max, **Annabel** (respect to your presentation-accompanying C. elegans moves and your never-ending socialising! It was awesome to have you as a tent buddy at the best kept secret festival), **Sven** (best hair of the Hubrecht), **Dennis**, **Jens** (salamander high five!), **Lolo** and **Cli** (yes because you are a pair. Thanks for bringing me back to the south when I had too much of Holland and for making the atmosphere relaxed!), **Ajit** (thanks for your sweetness and energy, man you are incredible!) **Ari-**

anna, **Maartje** and **PJ**, **Javi**, **AJ** (aka the beer commander), **Juri**, **Gita**, **Cayetano**, **Niels, Sjoerd, Banafsheh, Timo, Anna, Carlos, Wouter, Alice, Kadi, Erik, Christa, Wim, Saman, Maya, Charlotte, Iliana, Lotte, Arwa, Hesther, Bas M, Eirinn, Erica** (awesome karaoke)**, Carien, Bas C, Wim** (technically PMC but you know, part of the borrels) **and Caro.**

Also thanks to the support staff: **Anko** thanks for your help at the microscopes, **Peter-Erik** thanks for always responding to my emails in a flash and all your help and **Thea**, thanks for the happy good mornings! **Melanie**, thanks for all your tips on the thesis layout, for your enthusiasm on organising the Week van de Wetenschap, and for all the awesome podcast tips! **Annemiek**, thank you for your enthusiasm with the organisation of the AvO-Kind retreat and for your super clear instructions on wrapping up this thesis, so that I had everything under control. I really appreciate your support!

The UMC folks: **Onur**, so many yummy dinners, drinks, parties, boulder sessions, running, coffees. Thanks for your friendship, all the figs and delicious other things you brought back from Turkey, the long talks and the support, for the amazing idea we came up with and which we turned into a reality! Our project motivated me once more to do science, it was a ride! Thanks for your efforts on this and your never-ending enthusiasm. **Youri**, thanks for your hard work on the project. **Solee** thanks for your efforts with our experiments! Special thanks for the **Coffer lab** for letting me run around your lab and use your western blot expertise for a whole summer. Thanks **Simona** for your kindness, for being a calm presence cracking up hilarious jokes, we had some good laughs! **Irem**, I had the best time during my masters internship under your supervision. Thank you for guiding me with kindness and patience. Your way of supervision helped me supervise my own student in my PhD. Thanks for igniting my interest in psychology, I learned a lot the last years!

And now the mic goes to my life outside the lab. Meester **Julie**, wat een geluk dat we elkaar gevonden hebben tijdens de UIT! Dankjewel voor je oneindige steun, je nuchterheid, je bereidheid om weer eens een heel ingewikkeld mensenrechten/juridisch verhaal aan mij uit te leggen, voor je enthousiasme voor all things cultuur, het epische Tanzania avontuur en alle gezelligheid. Do the bird! **Γιάννα**, σε ευχαριστώ πολύ για τη στήριξή σου, που ποτέ δεν αμφέβαλες ότι θα τα καταφέρω, για τις αγκαλιές, τις περιπέτειες και τα κρασάκια! Θα είσαι πάντα η πάπια μου. Γλυκιά μου **Δήμητρα**, μον αμούρ και glamourness my God, ευτυχώς μπορούσα να μοιραστώ μαζί σου τις ιστορίες από το εργαστήριο και όλα τα συναισθήματά μου. Σε ευχαριστώ για την φιλία σου και όλες τις βόλτες! **Αγγελική**, πάντα μπορούσαμε να μοιραστούμε τα εσώψυχά μας η μία με την άλλη. Σε ευχαριστώ για τη στήριξη και τις όμορφες συζητήσεις, πάντα μαθαίνω κάτι καινούριο από σένα! **Κασσάνδρα**, ακόμα και αν δε βρισκόμαστε πια συχνά, σε ευχαριστώ για την ανεμελιά που τόσο εύκολα μου μεταδίδεις, τα φοβερά αστεία σου και για τις ανοιχτές και ειλικρινείς συζητήσεις! **Θανάση** σε ευχαριστώ για την στήριξή σου, και που ποτέ δεν αμφέβαλλες για μένα, την αλήθεια μου και τις ικανότητές μου. Ευχαριστώ εσένα και την **Κορνέι** για τις ευχάριστες στιγμές με την οικογένειά σας! **Eefke** en **Renee** bedankt voor het prettige samenwonen, alle gezelligheid en open gesprekken!

I don't like to quote ancient Greeks but they did have this wise saying: "A healthy mind in a healthy body": thanks to **Nicoline, Ruth, Henk** and especially **Evi** for all the listening, support and for shining a light down the path of life. Without you I would be lost. I hope the stigma of mental health will slowly disappear (from the academy) and that the mental health of all that are in it will become a serious priority in the future with some real policy changes being implemented from top down.

I also want to thank the following people and entities who helped me get through the tough times of the PhD and made me a stronger person: Deborah Frances-White and the Guilty Feminist Podcast, Brene Brown, Esther Perel, Beyonce, Damn Honey de podcast, Irvin Yalom, London Grammar, Eefje de Visser, Merol, Netflix, de Albert Heijn stoommaaltijden, Akis Petretzikis, the Olympos yoga sessions with Jane, boulderhal Sterk, my running shoes, Parnassos cultural center, the Spaghetteria, meneer Smakers, Camping Ganspoort, Tivoli, Paradiso, de Melkweg, Spotify, Ali Farka Toure, Lizzo, Alikiba, Diamond, Ludovico Einaudi, Ben Howard, Headspace, de Intratuin en de volkstuin on the Uithof, youtube and many others.

Θεία Άννα και **θείε Σταύρο**, **γιαγιά Μαρία** και **παππού Δημήτρη**, ευχαριστώ για τη στήριξή σας και για την αγάπη σας! **Κατερίνα**, **Guy**, **Μανούσο**, **Naomi**, thanks for you love, support, you always up for hosting me and for the fun summers! Little **Anna**, I am sure you will just slay at whatever you do. **Νονά** Βαγγελιώ και νονέ Γιάννη, σας ευχαριστώ πολύ για τη στήριξη και την αγάπη σας όλα αυτά τα χρόνια! Lieve **oma** en **opa,** dank u wel voor uw betrokkenheid bij mijn vooruitgang in mijn PhD, uw steun en liefde. **Dorine** en **Jaap** dank voor jullie openheid en warmte! Ik hoop nog vele gezellige momenten met jullie door te brengen. **Joris**, je bent de definitie van nuchterheid en kalmte, dankjewel voor al je grappen over mensenlengte, je liefde en steun! **Liesbeth**, **Frits**, **Nienke** en **Wessel**, dank jullie wel voor jullie openheid en betrokkenheid!

Lieve **mama**, dankjewel voor je al je steun, al je aanmoediging en liefde! **Μπαμπά** σ' ευχαριστώ για τη στήριξή σου, την αγάπη σου και τα ωραία αστεία σου! **Δημήτρη**, σ' ευχαριστώ για την αγάπη σου, την αιώνια υπομονή σου (πραγματικά, πώς το κάνεις?!), τη παντοτινή στήριξή σου, όλα τα αστεία και τις όμορφες στιγμές. Χωρίς το γέλιο και την όρεξή σου, η ζωή θα ήταν πολύ διαφορετική. Ξέρω ότι θα τα καταφέρεις στο επόμενο σου βήμα. Είμαι δίπλα σου!

Lieve **Ruben**, een dank je wel is een te klein woord. Je hebt voor mij gekookt, mij in Amsterdam rondgefietst (woops), mij getroost, mij aangemoedigd als geen ander, jezelf door mij laten meesleuren naar concerten, lezingen, rare films en heftige documentaires, de Amstel in, en mij altijd geprobeerd te begrijpen en te helpen. Ik waardeer en kijk op naar je eindeloze geduld en bereidheid om de positieve kant van alles te zien, je speelsheid, je advies op al mijn science-gerelateerde issues, je hands-on manier van dingen doen en je growth mindset in het leven. Ik ben enorm gegroeid samen met je, naast je, op zoveel verschillende vlakken en wil dat blijven doen. Blijf je vrolijke, zorgeloze zelf, ik heb zoveel zin in het leven met jou (en al onze planten)!

Λάουρα, δε μπορώ να φανταστώ πώς θα είχε πάει όλο το διδακτορικό (και γενικότερα η ζωή) χωρίς εσένα δίπλα μου. Θαυμάζω την επιμονή σου, τη θέλησή σου, όλη την πρόοδο που έχεις κάνει ανά τα χρόνια. Δε μπορώ να βρω τα σωστά λόγια να εκφραστώ, αλλά ευτυχώς δε χρειάζεται να πω τίποτα γιατί ξέρεις τι θέλω να πω (μου δίνεις μια στιγμή το εεεεεε ξέρεις τώρα). Όσο για το διδακτορικό σου.. δώσ'του να κατάλάβει, το 'χεις! Σ' ευχαριστώ για όλα.

That's it folks, it's been a ride!
Corina out.
[mic drop]

Curriculum Vitae

Corina Maria Markodimitraki was born on April 4th 1990 in Heraklion of Crete, Greece. She attended and graduated from the 8th General High school of Heraklion before enrolling at the University of Crete, at the faculty of Biology in Heraklion in 2008. She obtained her Bachelors degree in Biology with a specialization in Biomolecular Sciences and Biotechnology in 2012. One month later she moved to Utrecht, The Netherlands to start the masters program Molecular and Cellular Life Sciences at the University of Utrecht. During this time she joined the lab of Michiel Vermeulen for her major internship and worked in the lab under the supervision of Irem Baymaz. Later she joined the lab of Alexander van Oudenaarden for her minor internship and performed experiments under the supervision of Siddharth Dey. In 2014 she graduated from the masters program and travelled to Indonesia, Fiji and New Zealand where she worked as part of the Willing Workers On Organic Farms program and at the Ministry of Education in Wellington. She returned to Utrecht in the fall of 2015 and in February 2016 she joined the lab of Jop Kind in the Hubrecht Institute. The results obtained during her time as a PhD student are presented in this thesis.

Publication list

Rooijers, K.*, Markodimitraki, C.M.*, Rang, F.J., de Vries S.S., Chialastri, A., de Luca, K.L., Mooijman, D., Dey., S.S.# & Kind, J.# "Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells."
Nature Biotechnology 37, 766-772 (2019)
(*equal contribution, #co-corresponding)

Markodimitraki, C.M.*, Rang, F.J.*, Rooijers, K., de Vries S.S., Chialastri, A., de Luca, K.L., Lochs, S.J.A., Mooijman, D., Dey., S.S.# & Kind, J.# "Simultaneous quantification of protein-DNA interactions and transcriptomes in single cells with scDam&T-seq."
Nature Protocols 15, 1922-1953 (2020)
(*equal contribution, #co-corresponding)