Open access • Journal Article • DOI:10.1038/S41576-019-0195-2

Methods for mapping 3D chromosome architecture. — Source link

Rieke Kempfer, Ana Pombo, Ana Pombo

Institutions: Max Delbrück Center for Molecular Medicine, Humboldt University of Berlin

Published on: 01 Apr 2020 - Nature Reviews Genetics (Nature Publishing Group)

Topics: Chromosome conformation capture and Nuclear lamina

Chromatin folding in health and disease: exploring allele-specific topologies and the reorganization due to the 16p11.2 deletion in autism-spectrum disorder.

DISSERTATION

zur Erlangung des akademischen Grades

Doctor of Philosophy
(Ph.D.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

M.Sc Rieke Kempfer, geb. Fischer

Präsidentin
der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen

1. Prof. Ana Pombo
2. Prof. Uwe Ohler
3. Prof. Stefan Mundlos

Tag der mündlichen Prüfung: 07.10.2020

Declaration

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/ Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung gutter wissenschaftlicher Praxis eingehalten wurden.

Abstract

The 3D folding of interphase chromosomes inside the nucleus regulates important nuclear functions, such as transcription and replication, and once disrupted can lead to the manifestation of disease. Different techniques can be used to map 3D genome folding and detect pairwise and multiway interactions of the genome, or map the positions of DNA with respect to subnuclear compartments or the nuclear lamina. Here, I use GAM and Hi-C to explore two aspects of 3D genome topology, the allele specificity of chromatin contacts and long-range contacts between chromosomes, respectively. I detect specific contacts of the parental alleles in mouse embryonic stem cells and interactions between chromosomes in the context of congenital disease and study them with regard to their functionality and importance in mammalian gene regulation.

For detecting chromatin contacts with allele specificity, I produced a GAM dataset containing thousands of nuclear slices, which is part of the research of the 4D nucleome consortium. The collection of this data was accompanied by the development of a high-throughput version of GAM that allows the generation of large datasets. I show that GAM can determine haplotype-specific chromatin contacts with high efficiencies. First explorations of allele-specific chromatin topologies reveal many differences between the parental alleles, including allele-specific compartments A and B, and specific chromatin contacts, for example at the imprinted *H19/Igf2* locus.

For the exploration of inter-chromosomal contacts in disease, I mapped chromatin interactions with Hi-C in the context of a CNV at the human 16p11.2 locus, associated with autism spectrum disorders. Here, I show that the recurrent deletion at the 16p11.2 locus results in the rearrangement and loss of specific inter-chromosomal contacts between the 16p11.2 locus and chromosome 18 and propose a role for these inter-chromosomal contact changes in the upregulation of the nearby *Pcdhb* gene cluster, which comprises protocadherin genes with important functions in neuronal connectivity during development.

Zusammenfassung

Die 3D Struktur von Chromosomen im Zellkern reguliert verschiedene Funktionen in der Zelle, wie Transkription oder DNA Replikation, und Fehler in der 3D Faltung des Genoms können pathogen sein. 3D Genomfaltung kann mit verschiedenen Methoden untersucht werden um paarweiser und komplexerer Chromatinkontakte, sowie die Position von DNA in Relation zu sub-nuklearen Bereichen oder der Kernmembran zu detektieren. Hier verwende ich GAM und Hi-C um zwei Aspekte der 3D Genomtopologie zu untersuchen, die Allelspezifität von Chromatinkontakten und Kontakte zwischen Chromosomen. Ich untersuche allelspezifische Kontakte in murinen embryonalen Stammzellen und Interaktionen zwischen Chromosomen im Zusammenhang mit Autismus Spektrum Störung auf ihre Funktionalität und Relevanz in der Regulation von Genen.

Zur allelspezifischen Detektion von Chromatinkontakten generierte ich einen GAM Datensatz der tausende von nuklearen Cryodünnschnitten enthält. Diese Arbeit gehört zur Forschung des 4D Nucleome Konsortiums. Die Generierung dieser Daten beinhaltete die Entwicklung einer verbesserten Version der GAM Methode zur Produktion von großen Datensätzen in Hochdurchsatz. Hier zeige ich, dass GAM effizient Haplotyp-spezifische Chromatinkontakte bestimmen kann. Erste Untersuchungen von allelspezifischer 3D Genomtopologie zeigten weitreichende Unterschiede zwischen den Allelen, welche „A/B compartments“ und spezifische Chromatinkontakte beinhalten, wie zum Beispiel am Imprinting Locus *H19/Igf2*.

Zur Untersuchung von interchromosomalen Kontakten detektierte ich Chromatinkontakte mit Hi-C im Kontext einer genomischen Deletion am humanen 16p11.2 Locus, assoziiert mit Autismus Spektrum Störung. Ich zeige hier, dass die häufigste Deletion am 16p11.2 Locus zu der Reorganisation und dem Verlust von spezifischen interchromosomalen Kontakten zwischen 16p11.2 und Chromosom 18 führt, und stelle eine Hypothese auf wie diese interchromosomalen Kontakte zur ektopischen Aktivierung von *Pcdh* Genen auf Chromosom 18 führen. Protocadherins haben wichtige Funktionen in neuronaler Konnektivität, ein Prozess dessen Störung zur Manifestierung von Autismus Spektrum Störung beitragen könnte.

Acknowledgements

Several people have supported and helped me to during my PhD with guidance and advice, with contributions to the here presented work, but also with their friendship. First, I would like to thank my supervisor Ana for guiding me though my PhD and helping me manage three interesting but also challenging projects with ups and downs along the way. Thank you for lots of great advice and for helping me stay positive in difficult times.

Many thanks to Uwe and Robert for our annual committee meetings, your time and advice, and my PhD committee for taking the time to review this thesis.

I would like to thank Sasha (Alexander) for working with me on what sometimes seemed to be endless GAM optimisations, and for advice in all possible and impossible challenges that the wet lab provides. Many thanks go to all bioinformaticians that worked with me during my PhD; Ibai, Sasha (again), Ehsan, Christoph, Dominik, Tom, Rob, Mariano, and Markus. All of you are indispensable to making these projects a success, and you did not only do an amazing job but also managed to teach me a basic understanding of data analysis. Many thanks to Rob for all his prior work on GAM and for teaching me GAM when I joined the lab. Thank you, Enric, for joining our lab for the task of setting up an in-house WGA, your initiative and your work made this possible. Special thanks go to Gesa for helping me out when collecting nuclear profiles got too much for just one person, and for being a great friend. Without you the lab would have been only half the fun. I would also like to thank Izabela for her contribution to our final optimisations, which gave me a lot of positive energy when I really needed it, and everyone else in the Pombo lab for their help and/or friendship; Marta, Julietta, Warren, Leo, Doro, Anita, Joao, Elena, Giulia, Tiago, Carmelo, Jenny, Silvia, Luna. You are great colleagues to work with! Thanks to Michaela, Ines, Regina, and the PhD office for having my back with all organisational matters, and to Sasha, Gesa, and Jenny for final proofreading of the thesis. Thanks to great collaborators; Mario, Luca, Francesco, and Antonio for their past and ongoing development of SLICE and for understanding and helping biologists in need of mathematical solutions; Julia for her allele-specific analysis of RNA-seq data; Bing and Miao for providing the F123 mESCs, and for sharing the ChIP-seq data with us before it was published.

Many thanks to my family and friends who supported me, not only during this PhD but all my life, especially Philip who was always there for me, believed in me, and helped me with everything, including this thesis. I'd also like to thank my parents, Maike, Linus, Dani, Sarah, my grandmother, Lisa, and finally Fabio, my great motivator at the end.

Table of Contents

Index of Tables

Index of Figures

Abbreviations

3C	Chromosome conformation capture
4C	Circular chromosome conformation capture
5C	Chromosome conformation capture carbon copy
ASD	Autism spectrum disorder
ATAC	Assay of transposase accessible chromatin
Bp	Base pair
CCR	cluster control region
ChIA-Drop	Chromatin-interaction analysis via droplet-based and barcode-linked sequencing
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP	Chromatin immunoprecipitation
CNV	Copy number variation
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
D’	normalised linkage disequilibrium
DamID	DNA adenine methyltransferase identification
DAPI	4′,6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
EpiSC	Epiblast stem cells
ESC	Embryonic stem cell
EtOH	Ethanol
FA	Formaldehyde
FISH	Fluorescence *in situ* hybridisation
GAM	Genome architecture mapping
H3K27me3	Tri-methylation of the 27th lysine residue of histone 3
H3K4me1	Mono-methylation of the 4th lysine residue of histone 3
H3K4me3	Tri-methylation of the 4th lysine residue of histone 3
Hi-C	High throughput chromosome conformation capture
HPSF	High-purity, salt-free
ICE	Iterative correction and eigenvector decomposition
ICR	Imprinting control regions
iPSC	Induced pluripotent stem cells
LAD	Lamina associated domain
LIF	Leukemia inhibitory factor
LMD	Laser microdissection
MALBAC	Multiple annealing and looping based amplification
Mb	Megabase (pair)
MEF	Murine embryonic fibroblasts
NP	Nuclear profile

NPC	Neuronal precursor cell
NPMI	Normalised point mutual information
o.n.	overnight
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLAC	Proximity ligation-assisted ChIP
PRC	Polycomb repressive complex
PRCa	Active gene with PRC
RNA	Ribonucleic acid
RNApol II	RNA polymerase II
SCC	Spearman correlation coefficient
SLICE	Statistical inference of co-segregation
SNP	Single nucleotide polymorphism
SPRI	Solid phase reversible immobilisation
SPRITE	Split-pool recognition of interactions by tag extension
TAD	Topologically associating domain
TF	Transcription factor
TPM	Transcripts per million
TSA	Tyramide signal amplification
WGA	Whole genome amplification

1. Introduction

The materials in sections 1.1, 1.3 to 1.7, 1.10, and 1.11 of the introduction are from my previously published literature review (Kempfer and Pombo, 2019)

1.1 Summary and aim of the literature review

The nucleus of human cells harbours 46 densely packed chromosomes. Chromosomes are folded into hierarchical domains at different genomic scales, which likely enable efficient packaging and organize the genome into functional compartments. Chromosomes occupy distinct positions within the nucleus, called chromosome territories, which are partitioned into chromosomal compartments, and further into topologically associating domains (TADs) and chromatin loops which are mediated by either CCCTC-binding factor (CTCF) or enhancer-promoter contacts. Chromatin folding is a major feature of gene regulation and it dynamically changes in development and disease and potentially allele-specific. Transcriptional control is mediated through physical contacts between enhancers and target genes, which occurs via loop formation between the respective DNA elements. Functional loops between regulatory regions and genes are thought to occur predominantly within TADs. The expression of genes can also be influenced by their position relative to spatial landmarks inside the nucleus that are enriched for specific biochemical activities, such as the nuclear lamina. The disruption of enhancer-gene contacts and alteration of nuclear sub-compartments play important roles in disease, including congenital disorders and cancer. Importantly, many disease-associated mutations of the linear genomic sequence, particularly in non-coding regions, can only be understood by considering their 3D conformation in nuclear space.

Advances in our understanding of chromosome folding have been limited by a lack of approaches that can map chromatin contacts genome-wide while simultaneously retrieving spatial information, such as molecular distances between different genomic regions or between genomic regions and distinct nuclear compartments. Until recently, studies of 3D genome folding have been limited to two main technologies: imaging, particularly fluorescence *in situ* hybridisation of DNA (DNA-FISH), and approaches based on chromosome conformation capture (3C), namely Hi-C. DNA-FISH was a revolutionary approach which allowed visualisation of the spatial organisation of chromosomes and genes in the nucleus (Gall and Pardue, 1969; Speicher et al., 1996); it provides single cell information, but typically has limited throughput which only allows a small number of genomic loci to be analysed at a time. 3C-based approaches, which depend on proximity ligation of DNA ends involved in a chromatin contact, have helped identify enhancer-promoter contacts; their high-throughput derivatives, such as Hi-C, map chromatin contacts genome-wide at a length scale ranging from hundreds of kilobases to a few megabases. More recently, improvements in imaging techniques have increased the number of loci that can be analysed in parallel (Wang et al., 2016a) and extended the approach to live-cell imaging (Ma et al., 2013; Ma et al., 2016). Orthogonal ligation-free approaches have also emerged, namely genome architecture mapping (GAM; Beagrie et al., 2017), split-pool recognition of interactions by tag extension (SPRITE; Quinodoz et al., 2018), and chromatin-interaction analysis via droplet-based and barcode-linked sequencing (ChIA-Drop; Zheng et al., 2019) have started to reveal novel aspects of chromatin organisation. GAM, SPRITE and ChIA-Drop map chromatin contacts genome-wide and identify topological domains, but also robustly detect a previously unappreciated level of high complexity chromatin contacts that involve three or more DNA fragments and uncover specific contacts that span tens of megabases.

In this literature review I summarize the current knowledge of 3D genome topology at all genomic scales from structural conformation of entire chromosomes to local DNA looping and the impact of these topologies on nuclear functions, such as gene regulation. I describe differences in chromatin folding between the alleles of diploid organisms, and discuss the functionality of chromatin contacts discovered by studying congenital diseases. Further, I describe the main approaches currently used in 3D genome research, highlighting their major advantages and caveats. To recognise the strengths of each technique, it is important to understand the principles and experimental details underlying each method, their intrinsic biases and power to capture specific aspects of 3D genome architecture. I discuss major features of 3D genome organisation which have emerged, at the kilobase scale and above, through the application of these different technologies, while highlighting discrepancies between approaches.

1.2 Epigenetic gene regulation

The precise spatio-temporal regulation of genes is key to every multicellular organism to develop specialised cell types and tissues. Thus, many layers of gene regulation are necessary to orchestrate the complex task of precisely activating and silencing the transcription of genes. For example, the product of transcription, RNA, can be modulated by alternative splicing and post-transcriptional processing to affect its stability, or the efficiency of RNA transport to the cytoplasm, thereby influencing the translation rate of the RNA into protein. Other mechanisms for tuning gene expression occur at the level of transcription itself, via the recruitment of RNA polymerase II to the gene's promoter. This process it enabled and regulated by transcription factors (TFs) which recognise and bind to specific DNA sequences to initiate transcription. The sequences that recruit TFs to the gene are *cis*-regulatory elements, including promoters, enhancers, silencers, and insulators. Amongst these, enhancers play a crucial role in activating transcription, and since their discovery (Banerji et al., 1981) have been studied extensively to explore underlying mechanisms of enhancer-mediated gene activation (Levine, 2010; Ong and Corces, 2011). In most cases, to activate gene expression, the enhancer, which can be located far away from its target gene on the linear DNA sequence, is recognised by the TF and loops out of its linear genomic surrounding to physically contact the target promoter, resulting in the recruitment of RNA polymerase II to the target gene (Schoenfelder et al., 2010a).

1.3 Chromatin contacts between *cis*-regulatory elements

The physical contacts between enhancers and promoters are essential for the transcription of genes (Chen et al., 2018a) and can occur over distances ranging from less than one kilobase up to several megabases (Javierre et al., 2016; Lettice, 2003; Nobrega et al., 2003; Qin et al., 2004; Tolhuis et al., 2002) (Figure 1.1). Genome-wide maps of candidate promoter-enhancer contacts can be created using high-resolution C-methodologies that enrich for contacts mediated by RNA polymerase II or promoter histone marks, or that preferentially capture promoter-based contacts (Fang et al., 2016; Mifsud et al., 2015; Mumbach et al., 2016). Direct pairwise contacts between gene promoters and enhancers have become the most prominent concept of enhancer function, possibly as a result of the increased power of C-technologies to detect local pairwise contacts rather than higher-order conformations. However, other mechanisms of enhancer functions are also emerging, which can involve formation of chromatin hubs, tethering of genes to active chromatin or nuclear environments (Finlan et al., 2008; Kumaran and Spector, 2008; Reddy et al., 2008; Zullo et al., 2012), and phase separation (Nott et al., 2015; Strom et al., 2017). Interestingly, a study in budding yeast suggests homolog pairing as a mechanism for gene activation (Kim et al., 2017). In the diploid yeast genome, upon glucose deprivation of the cell, both copies of the genomic locus containing the gene *TDA1* are relocalized to the nuclear periphery, where the homologues associate with each other and *TDA1* expression is activated. A more classical concept in gene regulation can be observed at developmental loci, where *cis*-regulatory contacts between enhancers and promoters are thought to occur most commonly within TADs (Chetverina et al., 2014; Lupianez et al., 2015; Symmons et al., 2016). Although regulatory landscapes within TADs seem to be a common mechanism, genes themselves also contact each other across TAD boundaries over large genomic distances (Bantignies et al., 2011; Beagrie et al., 2017; Schoenfelder et al., 2015a; Tiwari et al., 2008). Ligation-free methods, such as FISH, GAM and SPRITE, all detect long-range contacts across TAD borders (Beagrie et al., 2017; Fraser et al., 2015; Quinodoz et al., 2018), and detailed analyses of Hi-C ligation frequencies also identify ligation events across TADs, over tens of megabases, that are statistically different from random contacts (Fraser et al., 2015). The functional relevance of these contacts is a compelling question that is beginning to be addressed by developments that allow ectopic chromatin contacts to be engineered in the cell (Deng et al., 2014; Kim et al., 2019). The spatial and functional relationship between gene promoters that contact each other also remains poorly understood. Deletions of several gene promoters in the mouse ESC genome altered the expression of nearby genes (Engreitz et al., 2016). This observation suggests that genes themselves may act as enhancers for other genes, possibly by recruiting *cis*-regulatory signals, and supports the concept that clustering of genes in transcription factories has regulatory functions.

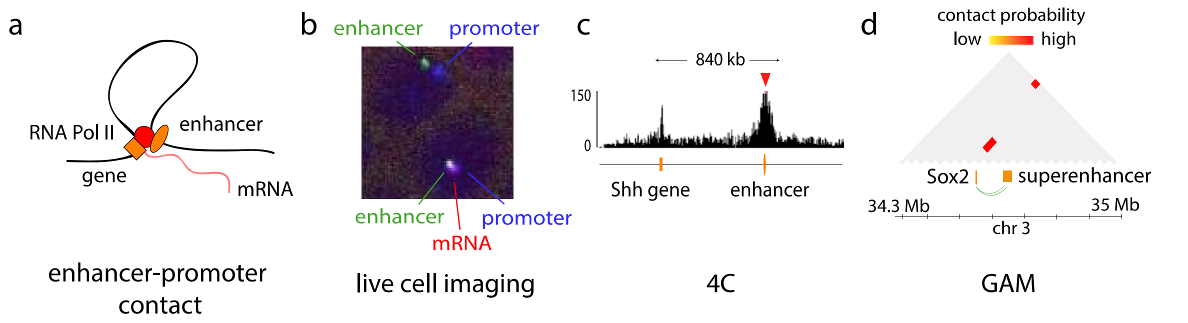


Figure 1.1: Enhancer-promoter contacts and their detection with different methodologies.
(a) Contacts between a gene and its *cis*-regulatory elements occur via loop formation between the enhancer bound by RNA polymerase II and the gene promoter. (b) Snapshot of live cell imaging of contacts between enhancer (green) and promoter (blue) of *eve* with simultaneous imaging of *eve* mRNA expression (red) in the Drosophila embryo (Chen et al., 2018a). (c) The 4C-sequencing track shows the interactions of the ZRS, a limb-specific enhancer of the *Shh* gene, with the *Shh* promoter in the anterior forelimb in mice (Symmons et al., 2016). (d) GAM data can be processed using a mathematic model, statistical inference of co-segregation (SLICE), to extract the most significant enhancer-promoter contacts from the dataset, resulting in a contact matrix with only the high-probability interactions (Beagrie et al., 2017). The most significant interaction at the *Sox2* locus can be found between the *Sox2* gene and one of its well-studied enhancers (Li et al., 2014). Diagram adapted from Kempfer and Pombo (2019).

1.4 Folding of chromatin into TADs and loop domains

At smaller scales, chromosomes fold into self-associating chromatin domains, termed TADs (Figure 1.2) (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). Chromatin domains had been previously identified by microscopy, but their detailed genomic composition was unclear. Since the discovery of TADs, the segmentation of the genome into megabase-sized domains has been extensively studied in several organisms and with different methodologies, leading to major breakthroughs in the discovery of mechanisms of disease caused by congenital genomic rearrangements (Franke et al., 2016; Hnisz et al., 2016; Lupianez et al., 2015; Spielmann et al., 2018). TADs often enclose clusters of co-regulated enhancers and promoters (Shen et al., 2012; Symmons et al., 2014). Their size has been re-examined with the increasing resolution afforded by improved 3C-based assays, and found to vary from 40 kb to 3 Mb in the human genome (Rao et al., 2014), leading to the proposal of smaller loop domains as a sub-structure of TADs. Loop domains had been detected by microscopy before the emergence of C-technologies as DNA loops between transcriptionally active regions (Jackson et al., 1996). Loop domains derived from 3C-based technologies often coincide with pairs of convergent CTCF binding sites, indicating that CTCF binding can contribute to the partition of specific regions of the genome into self-associating domains (de Wit et al., 2015; Gomez-Marin et al., 2015; Rao et al., 2014; Vietri Rudan et al., 2015). Higher-order contacts between TADs have also been investigated, leading to the identification of metaTADs, which bring together distant TADs in cell-type specific patterns that relate to gene activity (Fraser et al., 2015; Weinreb and Raphael, 2016).

It has been debated whether TADs represent domains that exist predominantly across the cell population, or represent an average of individual preferred contacts. Although interactions observed in single cells by single-cell Hi-C and imaging do not often identify whole TADs, the contacts detected frequently occur within the TAD coordinates defined by population Hi-C (Bintu et al., 2018; Nagano et al., 2013; Stevens et al., 2017). However, this preference might not be as strong as anticipated. Imaging of chromatin contacts in mouse ESCs and oocytes showed that 3D physical distances between regions that flank TAD borders are shorter in 40% of the cases than distances between regions within TADs (Flyamer et al., 2017), leading to highly variable contact clusters in individual cells, that do not coincide with the positions of TADs in the cell population. This observation agrees with the detection of chromatin contacts between regions separated by TAD borders in single cells, which are often found at similar frequencies as contacts within TADs (Finn et al., 2019). However, it is particularly noteworthy that combining the single-cell Hi-C data results in the same TAD coordinates observed in bulk population Hi-C, which supports the idea that TADs represent contact preferences of a cell population, rather than compact domains of chromatin in single cells (Flyamer et al., 2017; Fudenberg et al., 2016).

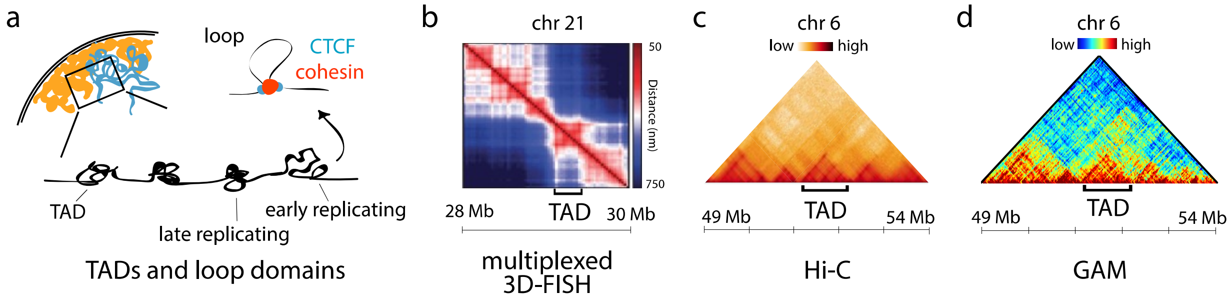


Figure 1.2: TADs and loop domains.
(a) Chromatin folds into topologically associating domains (TADs), which overlap with domains of early and late replication, and DNA loops, that arise from cohesin-mediated interactions between paired CTCF proteins. (b) Multiplex-FISH of consecutive DNA segments in a 2 Mb region in the human genome shows the emergence of TADs in the population-average distance map (Bintu et al., 2018). (c) and (d) In Hi-C and GAM contact maps, TADs are represented by regions of high internal interaction frequencies and demarcated by a drop of local interactions at their boundaries. (c) HiGlass (Kerpedjiev et al., 2018) was used to generate contact maps of previously published Hi-C data from mouse ESCs (Bonev et al., 2017); (d) Heatmaps for GAM were generated from normalised published matrix files for mouse ESC data (Beagrie et al., 2017). Figure adapted from Kempfer and Pombo (2019).

1.5 Organisation of DNA at nuclear bodies

Nuclear bodies are membrane-free organelles enriched for specific nuclear proteins and RNAs, which often have preferred associations with specific genomic regions, thereby influencing the large-scale organisation of chromosomes during interphase (Figure 3). They include the nucleolus, nuclear lamina, splicing speckles, paraspeckles, Cajal bodies, promyelocytic leukemia bodies, Polycomb bodies, replication factories, and transcription factories, all of which have been initially described using microscopy (Dundr and Misteli, 2010; Mao et al., 2011). For example, active ribosomal gene clusters are localized in the nucleolus, where the large ribosomal RNAs are transcribed, processed and assembled into pre-ribosomes (Pederson, 2011). Splicing speckles occupy internal nuclear positions, separate from the nuclear lamina and nucleoli, and bring together gene-dense regions (Brown et al., 2008; Shopland et al., 2003; Spector and Lamond, 2011). Association of specific genes at splicing speckles has not only been shown using imaging techniques, but has been confirmed with SPRITE at the genome-wide level, revealing that regions from different chromosomes come together on the same speckles (Beagrie et al., 2017; Fraser et al., 2015; Quinodoz et al., 2018). Genome-wide mapping of genes associated with speckles has also recently been achieved by TSA-seq (Chen et al., 2018b). Imaging using fluorescence and electron microscopy showed that transcription itself occurs at discrete sites in the nucleus, termed transcription factories, which may organise active transcription units (Iborra et al., 1996; Pombo et al., 1999; Xie et al., 2006), with only a small proportion of transcriptional activity (~5-10%) being found immediately adjacent to the most prominent splicing speckles (Xie et al., 2006). Interestingly, TSA-seq allows the fraction of the genome that associates closely with slicing speckles to be defined, and it revealed genomic regions that contain highly transcribed genes and super-enhancers (Chen et al., 2018b), in agreement with previous imaging data (Shopland et al., 2003). Co-expressed genes can share the same transcription factory, which may coordinate with mechanisms of coordinated gene regulation via chromatin folding (Ferrai et al., 2010; Osborne et al., 2004; Osborne and Eskiw, 2008; Schoenfelder et al., 2010b), but it remains unclear whether transcription factories are strictly specialised. Recent findings show that several factors involved in the transcription process, such as RNA polymerase II (Boehning et al., 2018), or transcriptional co-activators BRD4 and MED1 (Sabari et al., 2018), can form condensates by liquid-liquid phase separation, a process that may allow for concentration of TFs and generate transcription factories. Moreover, the formation of nuclear condensates has been suggested as a general principle of nuclear body formation (Banani et al., 2016). Clustering of distant genomic regions is not only mediated by transcription, but also occurs in the context of gene repression. Chromatin contacts at Polycomb bodies, which are repressive nuclear compartments, are a prominent example of gene clustering. In Drosophila, Polycomb-repressed *Hox* genes come together over a genomic distance of 10 Mb when they interact with a Polycomb body (Bantignies et al., 2011). Other studies have reported long-range intra- and inter-chromosomal contacts between Polycomb-bound genes in human teratocarcinoma cells (Tiwari et al., 2008) and in mouse ESCs (Mifsud et al., 2015).

Understanding how the preferential associations of genomic regions with specific nuclear domains relate with 3C-derived chromatin contacts remains a major challenge. Comparisons of genome-wide maps of lamina-associated domains (Guelen et al., 2008) and Hi-C contacts show a strong coincidence between the transcriptionally inactive B compartment and the nuclear lamina (Dixon et al., 2012; Fraser et al., 2015; Rao et al., 2014) or late replication domains (Pope et al., 2014). Repressive histone marks that define the heterochromatic B compartment are also strongly enriched at genomic regions that associate with the nucleolus

(Nemeth et al., 2010), suggesting that the compacted B-compartment is situated both at the nuclear periphery and clustered around the more central nucleoli, separated by the active, open A-compartment. However, the bimodal separation of A and B compartments derived from C-technologies should not be naively inferred as strictly active or silent chromatin. Genes can be activated in all areas of the nucleus, including at the nuclear lamina (Kumaran and Spector, 2008) or at centromeric regions (Sabbattini et al., 1999), and gene positioning at the periphery does not always result in gene inactivation (Finlan et al., 2008; Wang et al., 2018). Heterochromatin domains also contain active sites of transcription (Grewal and Elgin, 2007). Consequently, a strict separation of compartments A and B, as defined by C-approaches, seems unlikely, especially considering that contacts between compartments can be found in Hi-C maps (Dixon et al., 2012), and that these contacts are even more robustly identified with use of orthogonal methods such as GAM, that do not rely on weak fixations (Beagrie et al., 2017). These observations suggest that long-range gene regulation mechanisms are complex, and not only depend on pairwise contacts between genomic regions, but may also be influenced by the local nuclear environment where each region is located (Pombo and Branco, 2007). The ongoing challenge of disentangling the direct functional relationships between the positioning of genomic regions in the nucleus, their local and long-range contacts, and the state of gene activity is being addressed by analysing chromatin contacts at the single-cell level and with allele specificity, for example using DNA-FISH (Finn et al., 2019) or single-cell Hi-C (Flyamer et al., 2017; Nagano et al., 2017).

1.6 Chromatin folds into hubs and compartments

The organisation of chromosomes into sub-chromosomal domains has been extensively studied. For example, in mammalian cells, chromatin domains were observed in relation to replication origins, which contain many replicons and maintain their domain co-association across subsequent cell cycles (Jackson and Pombo, 1998). The compartmentalization of chromosomes into early and late replicating domains (Ferreira et al., 1997; Visser et al., 1998; Zink et al., 1999) was also shown to be linked to transcriptional activity, with sites of active transcription occurring predominantly in early replicating domains (Sadoni et al., 1999). More recently, these observations have been largely confirmed by genome-wide assays to map replication and transcription, in which transcriptionally active and early replicating chromatin domains organise into separate sub-compartments, distinct from late replicating domains (Hiratani et al., 2008; Pope et al., 2014; Schwaiger et al., 2009). Early analyses of nuclear organisation by electron and confocal microscopy had shown that chromatin occurs in highly condensed (heterochromatic) and less condensed (euchromatic) states (Monneron and Bernhard, 1969), and revealed that transcription occurs in euchromatic areas of the nucleus (Verschure et al., 1999). With the emergence of whole genome C-methodologies, such as Hi-C, the mapping of active and repressed chromatin states has become possible at the genome-wide scale, providing powerful insights into how gene expression relates to chromatin compaction. Application of principal component analysis to Hi-C data revealed a strong segregation of ligation events into two distinct compartments (A and B compartments) according to the activity state of the genomic regions (Lieberman-Aiden et al., 2009). These compartments can also be seen in contact maps generated by ligation-free approaches (Beagrie et al., 2017; Quinodoz et al., 2018) (Figure 1.3). Comparisons with linear maps of protein occupancy on chromatin helped reveal a strong relationship between compartment A and transcriptionally active, open chromatin, as defined by DNase hypersensitivity, and compartment B with closed chromatin, defined by repressive epigenetic marks on heterochromatin (Lieberman-Aiden et al., 2009). Increased depth of Hi-C datasets has since allowed smaller sub-compartments to be detected, which capture finer differences in replication timing as well as preferred associations with the nucleolus or the nuclear lamina (Rao et al., 2014).

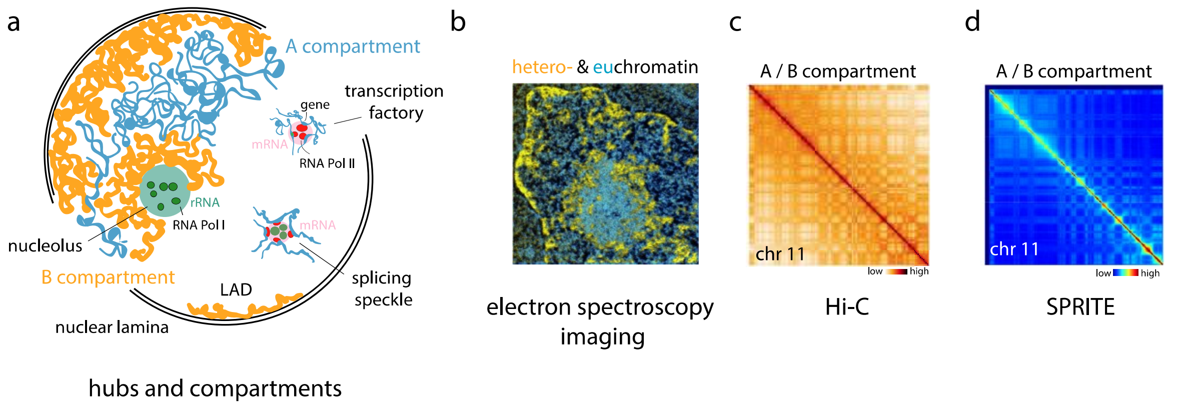


Figure 1.3: Nuclear bodies and compartments.
(a) DNA inside the nucleus separates into hubs of active (compartment A) and inactive (compartment B) chromatin, clustering around the nucleolus, splicing speckles, transcription factories, and other nuclear bodies not represented here. (b) Electron spectroscopy imaging of the mouse epiblast shows distribution of heterochromatin (yellow) around the nucleolus (light blue) and at the nuclear periphery. Decondensed euchromatin (dark blue) is positioned more centrally in the nucleus. Nucleic acid-based structures are stained yellow, protein-based structures blue (Ahmed et al., 2010). (c) and (d) Hi-C and SPRITE contact maps of mouse chromosome 11 show the separation of chromatin into discrete contact hubs (A and B compartments), which are visible as checkerboard-like contact patterns. (c) HiGlass (Kerpedjiev et al., 2018) was used to generate contact map for previously published Hi-C data from mouse ESCs (Bonev et al., 2017); (d) SPRITE contact map was generated from normalised published matrix files from mouse ESCs (Quinodoz et al., 2018). Figure adapted from Kempfer and Pombo (2019).

1.7 Chromosome territories and inter-chromosomal contacts.

FISH imaging shows that specific chromosomes occupy discrete non-random nuclear spaces during interphase, termed chromosome territories (Cremer and Cremer, 2010) (Figure 1.4). Chromosome territories show cell-type dependent preferences in terms of both their radial position within the nucleus and their position relative to other chromosomes (Branco et al., 2008; Branco and Pombo, 2006; Parada and Misteli, 2002). Specific contacts can be detected at the interface between chromosome territories (Hacisuleyman et al., 2014; Maass et al., 2018a; Maass et al., 2018b); overall, an estimated 20 % of the volume of chromosome territories intermingles with other chromosomes, often at their peripheries, both in human primary lymphocytes and in *Drosophila melanogaster* cells (Branco and Pombo, 2006; Rosin et al., 2018). The extent of intermingling between chromosome territories directly correlates with translocation probabilities upon ionizing radiation damage, highlighting that the physical proximity between chromosomes affects their stability in response to DNA damage (Branco and Pombo, 2006; Maharana et al., 2016; Zhang et al., 2012). The organisation of chromosomes into discrete territories is also inferred from 3C-based and ligation-free approaches, as higher interaction frequencies are detected within chromosomes than between them (Figure 1.4). 3C-based technologies have also detected contacts between chromosomes, and these have been successfully validated by imaging (Cairns et al., 2016; Fanucchi et al., 2013; Hacisuleyman et al., 2014; Lomvardas et al., 2006; Mifsud et al., 2015; Nagano et al., 2017; Schoenfelder et al., 2010b; Spilianakis et al., 2005).

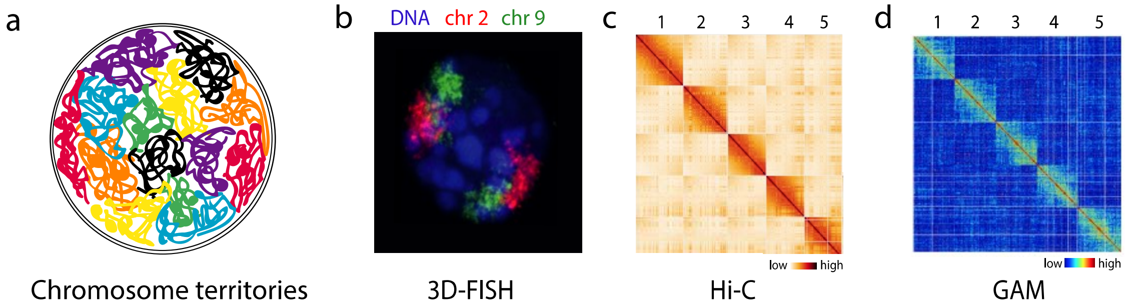


Figure 1.4: Chromosomes occupy discrete territories in the nucleus, which were first detected using imaging techniques.
(a) Illustration of chromosome territories. (b) The 3D-fluorescence *in situ* hybridisation (3D-FISH) image shows the positions of the chromosome territories of chromosome 2 (red) and chromosome 9 (green) within DAPI-stained nuclei (blue) from mouse embryonic stem cells (ESCs) (Mayer et al., 2005). (c) Chromosome territories are also detected as regions of high frequency intrachromosomal interactions on contact maps generated by 3C-based methods (such as Hi-C) and (d) ligation-free approaches (such as genome architecture mapping (GAM)). (c) HiGlass (Kerpedjiev et al., 2018) was used to generate contact maps for previously published Hi-C data from mouse ESCs (Bonev et al., 2017); (d) Heatmaps for GAM were generated from normalised published matrix files for mouse ESC data (Beagrie et al., 2017). Figure adapted from Kempfer and Pombo (2019).

1.8 Allelic differences in 3D chromatin topology

Diploid organisms carry two copies of each chromosome (with some exceptions, such as the sex chromosomes of male humans), one allele inherited from each parent. Genes can be expressed from both alleles (biallelic), or only from one of the two copies (monoallelic). When genes are expressed constitutively monoallelic, in all somatic cell types and tissues of the organism, they are called imprinted genes (Reik and Walter, 2001). While these are only 100 to 150 genes in human and mice (Kelsey and Bartolomei, 2012), the majority of monoallelic gene expression is cell-type specific, and depending on the cell type monoallelic expression affects up to 10 % of all genes (Reinius and Sandberg, 2015). Thus, allele-specific gene regulation must take place to activate only one of the two copies of a gene. At imprinted genes, allelic regulation is thought to involve long non-coding RNAs, as well as allele-specific DNA methylation at so called imprinting control regions (ICRs) (Bartolomei, 2009). Another proposed mechanism for allelic gene expression is differential chromatin folding between the maternal and the paternal allele, a phenomenon that could be observed for example at the immunoglobulin heavy chain locus, where the allelic recombination of V_H, D_H and J_H gene segments controls the expression of specific lymphocyte receptors (Holwerda et al., 2013).

Several genome-wide studies of allelic chromatin folding have mapped chromosome structures of haplotypes. In Hi-C studies of human cells, only minor differences between the chromatin folding of alleles could be observed, however, these studies report examples of differential chromatin folding at imprinted loci (Dixon et al., 2015; Rao et al., 2014). The imprinted genes *H19* and *IGF2*, transcribed maternally and paternally, respectively, display differential formation of CTCF-mediated loops at their genomic locus, an observation that was made in bulk Hi-C data (Rao et al., 2014), and was confirmed later in single cells (Tan et al., 2018a). Studies of CTCF-mediated chromatin contacts using ChIA-PET (Tang et al., 2015) or contacts between transcriptionally active genomic regions (using Hi-ChIP with H3K27ac) (Mumbach et al., 2017) report larger differences in chromatin folding between alleles. In this case, bias is introduced when differences between alleles are directly connected to differential occupancy of CTCF or histone marks. This may not necessarily lead to changes in chromatin contact frequencies, but instead, only reflects changes in which genomic regions are being captured in that particular experiment. More unbiased studies have explored allelic chromatin folding, by using high-throughput imaging with Oligopaints (Nir et al., 2018). In this study, consecutive 8 Mb long stretches of chromatin are imaged at low resolution, reporting differences in the shape of the visualised genomic regions, reflected in differential ellipticity between the alleles. This implies that allelic differences in chromatin topology go beyond a few imprinted loci, as previous studies suggested. A study in mouse ESCs mapped chromatin contacts and found genome-wide changes in Hi-C compartments between each allele. The observed changes did not connect to imprinting or monoallelic gene expression, but rather to asynchronous replication timing between alleles (Rivera-Mulia et al., 2018). Summarising, most studies that explored allelic chromatin contacts have failed to find a connection between monoallelic expression and regulation via genome folding outside of imprinted loci. However, careful explorations of Hi-C data in human induced pluripotent stem cells (iPSCs) and iPSC-derived cardiomyocytes identified genome-wide changes in DNA loop formation between the alleles (Greenwald et al., 2019). Although these very subtle changes in interaction frequencies are magnitudes lower than the differential loops between cell types, they correlate with monoallelic gene expression, thus providing the first evidence for a global regulation of monoallelic gene regulation via chromatin loop formation.

1.9 3D genome folding in disease

The functionality of different chromatin contacts is a key question when studying the 3D nucleus. Although several studies have suggested the importance of changes in chromatin structures throughout development and between different cell types (Bonev et al., 2017; Dixon et al., 2015; Fraser et al., 2015), direct and inevitable proof of the importance of chromosome folding for the development of an organism was found by studying topologies in the context of congenital disease. Structural rearrangements at the human *EPHA*4 locus that are associated with various limb malformations have been shown to cause alterations in chromatin folding at the genomic locus of *Epha4*, when remodelling the disease in mice (Lupianez et al., 2015). Here, disruption of TAD structures at the locus leads to ectopic interactions of *Epha4* enhancers with genes located in the adjacent TADs, resulting in misregulation of developmental genes and consequently in limb malformations. A number of disease-associated studies of 3D chromatin folding have shed light into the importance of enhancer-promoter contacts by revealing mechanisms like enhancer adoption, meaning the exposure of a gene to the signal of an ectopic enhancer due to TAD disruptions, to be the cause of rare developmental disorders, such as adult-onset demyelinating leukodystrophy (Giorgio et al., 2015) or female to male sex reversal (Franke et al., 2016). TAD disruptions also play important roles in cancer, such as T cell acute lymphoblastic leukemia, where microdeletions at TAD boundaries lead to activation of proto-oncogenes (Hnisz et al., 2016). Removal of these TAD boundaries in human embryonic kidney cells was sufficient to activate the proto-oncogenes in healthy cells. The same mechanism could be observed when studying cancer-associated mutations at CTCF sites. Deletions of CTCF sites in human embryonic stem cells at the boundaries of TADs lead to ectopic interactions of key developmental enhancers and their target genes, causing misexpression of the genes inside these TAD (Ji et al., 2016). These findings indicate once more the importance of TAD structures for gene regulation and consequently the functionality of the cell. At smaller scales, the disruption of enhancer-promoter contacts has been shown for various disease studies, where translocations or copy number variations (CNVs) disrupt chromatin looping in the proximity to the mutation, thus resulting in aberrant gene expression by interrupting or altering promoter-enhancer contacts (Hyon et al., 2015; Lettice et al., 2011; Schmitz et al., 2014; Watson et al., 2016).

However, above the level of TADs chromosomes fold into larger scale chromosomal compartments and establish contacts between genomic regions bridging tens of megabases

(Beagrie et al., 2017; Fraser et al., 2015; Quinodoz et al., 2018), or even crossing chromosome territories (Cairns et al., 2016; Fanucchi et al., 2013; Hacisuleyman et al., 2014; Lomvardas et al., 2006; Mifsud et al., 2015; Nagano et al., 2017; Schoenfelder et al., 2010b; Spilianakis et al., 2005). While the reports about such contacts are increasing, the question about the functionality and importance of long-range chromatin contacts above the TAD level is still debated. While there are some examples about the role of inter-chromosomal contacts in gene expression (Apostolou and Thanos, 2008; Horta et al., 2018; Lomvardas et al., 2006; Spilianakis et al., 2005), so far there is no causal relationship known between alterations in inter-chromosomal contacts and the misregulation of genes. Studying long-range and inter-chromosomal contact changes in disease is the first step towards a better understanding of these contacts. So far disease-related changes in contacts between chromosomes have been discovered in epithelial and breast cancer cells (Barutcu et al., 2015) and in human cells harbouring large CNVs, such as the 2q37 deletion (Maass et al., 2018b), the genetic cause of brachydactyly mental retardation syndrome , 22q11.2 and 1q21.1 CNVs, associated with neurological disorders (Zhang et al., 2018a), and 16p11.2 CNVs, associated with autism spectrum disorder (ASD) (Loviglio et al., 2016). While all those studies report changes in inter-chromosomal contacts as well as genome-wide changes in gene expression, those two events could not be directly connected. The question therefore is if these contacts are regulating genes and if they, when disturbed, are causal for the manifestation of the disease.

1.10 Techniques to study 3D genome folding

Here, I discuss methods to detect chromatin contacts in fixed and live cells, bulk experiments and single cell, using genome-wide sequencing approaches or single-locus imaging techniques. Further, I explain experimental differences and limitations of the currently used techniques to map chromosome topology. An overview of all described methods can be found at the end of this chapter (Table 1.1).

1.10.1 Imaging-based detection of contacts

The visualisation of nuclear structures and specific genomic sequences is key to understanding how chromatin is organised in the nucleus. A variety of light and electron microscopy techniques can be used to identify nuclear compartments or image the physical positions of specific genomic loci in the nucleus of fixed or live cells. The most commonly used techniques for detecting chromatin contacts are DNA-FISH in fixed cells, Lac and Tet operator systems or, more recently, CRISPR-based imaging in live cells (Figure 1.5).

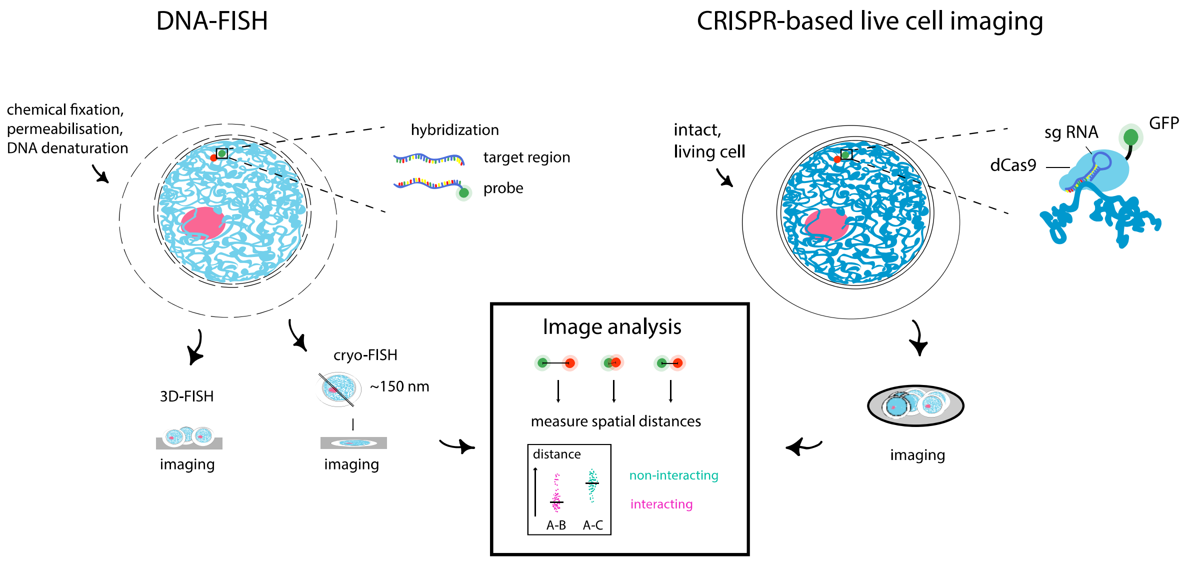


Figure 1.5: Imaging-based approaches to visualise chromatin contacts.
(a) DNA- fluorescence *in situ* hybridisation (DNA-FISH) uses fluorescently labelled probes that hybridise to specific genomic loci in the nucleus. Typically, cells are fixed and permeabilised, and upon denaturing of the DNA, FISH probes hybridise to their complementary target region. The FISH procedure can be performed in whole cells, embryos, thick tissue slices (3D-FISH) or in thin cryosections of cells (cryo-FISH). (b) CRISPR-based live cell imaging can be performed in the intact, living cell. Typically, a dead Cas9 (dCas9) is fused to green fluorescent protein (GFP) and the fusion protein is recruited to the target region by small guide (sg) RNAs, which are complementary to the region of interest. (c) For all the techniques, chromatin contacts are assessed by the spatial distances between the fluorophores targeting the regions of interest. To determine the specificity of a contact, spatial distances between interacting loci should be compared to distances between non-interacting loci in a number of cells. The distribution of distances, the mean distance and median distance can all inform about the quality and abundance of the contact in a cell population. Figure adapted from Kempfer and Pombo (2019).

1.10.1.1 Measuring contacts with DNA-FISH

FISH uses fluorescently tagged DNA sequences (such as oligonucleotides) as probes to hybridise to complementary target regions of interest in the genome (Figure 1.5). For hybridisation to occur, a single-stranded probe needs to be able to enter the nucleus, which is usually achieved by permeabilising the cell with a detergent or organic solvent such as methanol. To ensure the probe can bind to its target, the DNA is most often denatured by heat and formamide treatment. The genomic regions highlighted by the hybridised fluorescent probes are then visualised under a microscope.

DNA-FISH is typically used to measure the physical distances between two or a few differentially labelled genomic regions of interest. A chromatin contact is often defined by a distance threshold, which is usually set arbitrarily according to the scale of genomic distances between the regions of interest and the resolution of the microscope. Thus, chromatin contacts have been inferred when fluorescent signals co-localise within a spatial distance of 50 nm to 1µm (Barbieri et al., 2017; Barutcu et al., 2018; Finn et al., 2019; Maass et al., 2018a), although it is unclear whether the larger separations, close to the diameter of a whole chromosome, represent true interactions or non-random positioning. DNA-FISH can also be used to visualise chromatin compaction (Boettiger et al., 2016) or positioning of genomic regions with respect to nuclear structures, such as the nuclear lamina (Luperchio et al., 2017). The overall distributions of spatial distances between loci or relative to the nuclear periphery found across the cell population are usually summarised by the frequency of co-localisation (that is, the frequency with which chromatin contacts are detected across the cell population), but other metrics, such as mean or median distances, are also used. The data is compared to the physical distances between other (control) genomic regions (which are often separated by similar genomic distances) or, in some cases, to nuclear diameter or volume. These metrics can help distinguish specific chromosomal conformations but can also be ambiguous, depending on the choice of control probes or if allelic differences or other forms of heterogeneity are present within the cell population.

The accuracy and power to detect different nuclear structures or contacts also depend on how well the organisation of the target DNA and nuclear compartments are preserved during the FISH procedure, on the resolution of the microscope and on the size of the target genomic sequence. FISH experiments use probes made of a collection of small DNA fragments that are either synthesized (oligos) or produced from larger DNA molecules (plasmids, fosmids, bacterial artificial chromosomes or whole mammalian chromosomes) by nick-translation, resulting in overlapping fragments of 100-500 bp. The probes often cover genomic sequences ranging in length from 30 kb up to entire chromosomes. Targeting larger genomic regions has the advantage of giving higher signal-to-noise ratios in locus detection, due to increased local fluorescence and higher target specificity. This allows accurate detection of contacts between larger genomic regions, such as TADs (Beagrie et al., 2017; Nora et al., 2012) and chromosomes (Branco and Pombo, 2006). However, with standard FISH techniques short-range interactions between chromosomal regions that are less than 100 kb apart are hard to detect, making it difficult to quantify the finer-scale chromatin folding below the TAD level, such as enhancer-promoter interactions.

Higher resolution imaging of chromatin contacts can be achieved using cryoFISH, in which standard FISH probes are hybridised to thin (~100-200 nm) cryosections from cells fixed using conditions optimised to preserve nuclear ultrastructure; the signal is then visualised using fluorescence or electron microscopy (Barbieri et al., 2017; Beagrie et al., 2017; Branco and Pombo, 2006; Ferrai et al., 2010; Simonis et al., 2006). More recently, the short length and specificity of fluorophore-tagged oligonucleotides known as Oligopaints (Beliveau et al., 2012) have made it possible to target 15 kb loci using conventional microscopy (Boyle et al., 2011), or 5 kb regions using super-resolution microscopy (when combined with a second labelling step to enhance the fluorescence signal) (Beliveau et al., 2015). Oligopaints are not derived from cloned genomic regions, but are instead generated from synthetic libraries of short (~60-100 bp) oligonucleotides, which are produced by massively parallel synthesis (Gnirke et al., 2009). Once generated, the library pool can be amplified in a flexible manner, using different primer pairs to give rise to different sets of FISH probes. The ease of design of the Oligopaint probes has opened new possibilities to study chromatin folding, such as being able to visualise chromatin in different epigenetic states in the resolution of tens of nanometers (Boettiger et al., 2016). Oligopaint-based FISH has also been used in combination with high-throughput imaging to generate low-resolution contact maps (e.g. at the TAD level) of whole chromosomes (Wang et al., 2016a) and high resolution (30 kb) contact maps for stretches of DNA 1.2–2.5 Mb in length (Bintu et al., 2018). In addition, molecular beacon FISH probes have emerged as way to target genomic regions as short as 2.5 kb (Ni et al., 2017). In an unbound state, these probes form a hairpin loop that minimizes off-target fluorescent signal by bringing together the fluorescent label and a quencher. Thus, the technique improves the visualisation of small genomic regions by reducing the background signal of the unbound probe.

1.10.1.2 Live-cell imaging of nuclear structures

Chromosome folding is a highly dynamic process that varies greatly throughout the cell cycle (Gibcus et al., 2018; Stevens et al., 2017). Our ability to study these chromatin dynamics have been revolutionised by technologies based on genome editing that allow specific genomic loci to be targeted in live cells. Early iterations of this approach were rather laborious; cell lines needed to be created in which the target locus was tagged with DNA binding site arrays that recruit a fluorescently-tagged cognate DNA-binding protein (such as the Lac operator-repressor (Belmont and Straight, 1998; Robinett et al., 1996), Tet operator-repressor (Lucas et al., 2014) and ANCHOR (Germier et al., 2018) systems). Now, loci can be targeted in live cells with a version of the CRISPR (clustered regularly interspaced short palindromic repeats) system that uses an endonuclease-deficient form of Cas9 (dead-Cas9 (dCas9)) fused with a fluorescent protein (Chen et al., 2013). The tagged dCas9 is recruited to the genomic locus of interest via its interactions with sequence-specific small guide RNAs (Figure 1.5). For simultaneous labelling of two genomic regions, small guide RNAs can be differentially modified to act as scaffolds that bring fluorescent proteins to the target loci. For example, fusion proteins that comprise a fluorescent protein and either tandem dimer MS2 coat-binding protein (tdMCP) or tandem dimer PP7 coat-binding protein (tdPCP) can be directed to target loci by guide RNAs containing MS2 or PP7 aptamers, respectively. Since both proteins have a comparably high exchange rate, this approach is also well suited to long term live cell imaging, as it compensates for photo bleaching (Fu et al., 2016; Shao et al., 2016; Wang et al., 2016b). However, most CRISPR-based methods are currently limited to the detection of repetitive sequences because they rely on a single species of guide RNA, which hybridises to identical genomics sequences, to direct simultaneous binding of dozens of copies of the fluorescent protein to achieve a strong fluorescent signal. A notable exception is chimeric array of gRNA oligonucleotides (CARGO); by delivering 12 different guide RNAs into a single cell, this technique was able to efficiently label a non-repetitive 2-kb genomic region (Gu et al., 2018).

1.10.2 Ligation-based detection of contacts

3C-based methods extract chromatin interaction frequencies between genomic loci via chromatin crosslinking and proximity ligation (Figure 1.6). Following formaldehyde fixation to capture protein- and RNA-mediated contacts, chromatin is fragmented using a restriction enzyme, and the crosslinked restriction fragments are ligated (Dekker et al., 2002). The purified ligation fragments are called a 3C library. The ligation frequency between two loci of interest can be quantified by PCR using appropriate primer pairs. Thus, 3C focuses on interactions between two loci (‘one versus one’) and requires prior knowledge of the targets of interest. However, the 3C library contains all ligation products for the genome investigated and the 3C workflow can therefore be adapted to enable genome-wide analysis of chromatin contacts. Chromosome conformation capture-on-chip (Simonis et al., 2006) or circular chromosome conformation capture (Zhao et al., 2006), both called 4C, enrich for interactions of one region with the remaining genome (‘one versus all’). Chromosome conformation capture carbon copy (5C) (Dostie et al., 2006) captures contacts of a larger genomic stretch at high resolution (‘many versus many’). Finally, Hi-C (Lieberman-Aiden et al., 2009) captures all ligation events across the entire genome (‘all versus all’). In the following, I focus on the most commonly used versions (Figure 1.6).

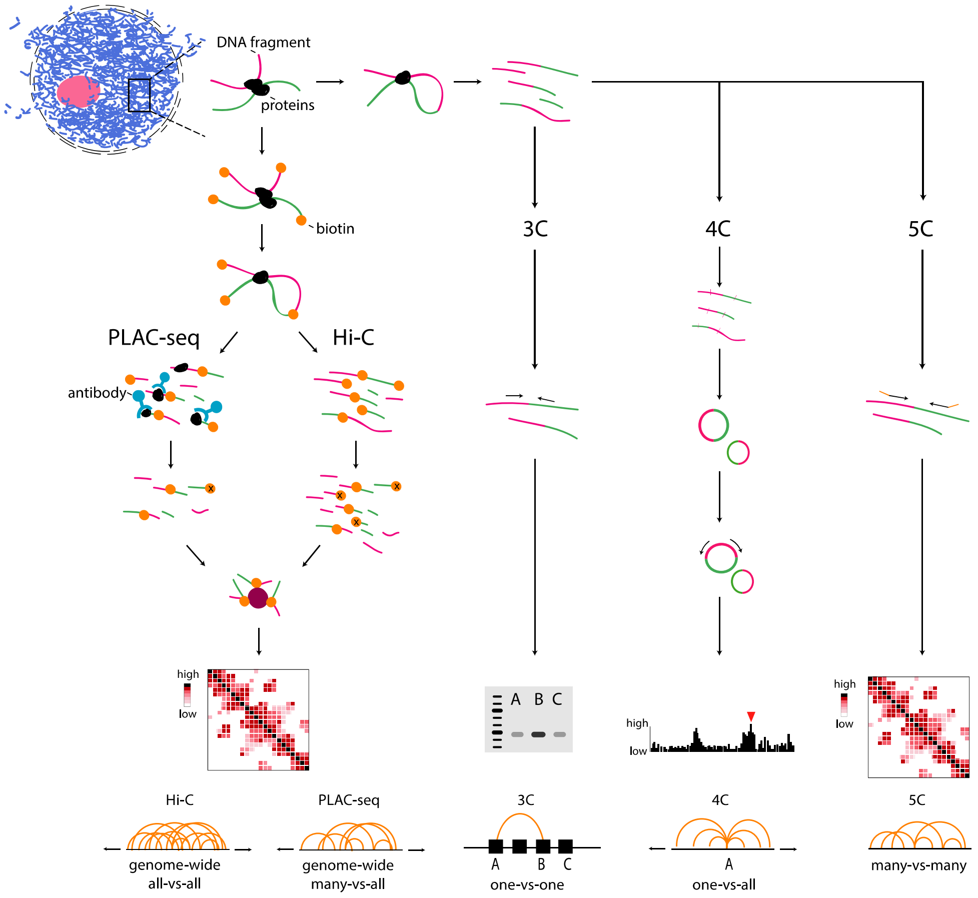


Figure 1.6: Chromosome conformation capture and its derivatives.
Chromosome conformation capture (3C)-based assays measure contact frequencies of pairs of DNA loci by proximity ligation of crosslinked and fragmented chromatin. All 3C-based assays involve fixation of the chromatin, isolation of nuclei, and DNA fragmentation (e.g. with a restriction enzyme). The obtained crosslinked chromatin fragments are then processed for 3C, 4C or 5C, which map chromatin contacts for preselected regions, or for genome-wide assays, such as Hi-C and PLAC-seq. In 3C, 4C, and 5C, the crosslinked chromatin fragments are ligated and the DNA is purified. In 3C, the interactions between two chosen genomic regions are detected by amplification with primers for the two regions of interest. PCR products are analysed semi-quantitatively on an agarose gel, or by real-time quantitative PCR. Interactions are defined by higher ligation frequencies compared to control regions of similar genomic distance. In 4C, interactions of one viewpoint with the whole genome are measured. The ligated and purified DNA is fractionated with a secondary restriction digest, and the digested, smaller DNA fragments are circularised, and amplified with primers facing outwards from the viewpoint (the restriction fragment containing the region of interest). The PCR products are sequenced by paired end sequencing, providing the sequence information and frequency of every chromatin contact of the viewpoint. In 5C, the ligated and purified DNA is directly amplified using primers for all restriction fragments within a consecutive genomic region, usually hundreds of kilobases, up to several megabases. The PCR products are sequenced and provide information about the ligation frequencies of all fragments within the region of interest. In Hi-C and PLAC-seq, prior to ligation digested DNA fragments are labelled with biotin. Then DNA fragments are ligated, and fragmented further by sonication. In PLAC-seq, DNA fragments bound to a protein of interest are pulled-down by immunoprecipitation. Then, in PLAC-seq and Hi-C the DNA is purified, biotinylated nucleotides are removed from unligated fragment ends, and all ligated DNA fragments are pulled down with streptavidin beads. After pull-down, DNA fragments are sequenced and provide information about the interaction frequencies of all pairs of loci in the genome (Hi-C), or specifically the interactions mediated by a protein of interest (PLAC-seq). Figure adapted from Kempfer and Pombo (2019).

1.10.2.1 Mapping all contacts at a single locus with 4C

A straightforward and cost-effective method to obtain additional information from a 3C library is 4C. Here, primers for a region of interest (such as a promoter) are used to amplify all ligation partners of the locus under investigation (called the 'viewpoint') (Figure 1.6). The amplified ligation products are sequenced (to a depth of 1 to 5 million reads per library (van de Werken et al., 2012b)) and used to analyse genome-wide interaction partners of the region of interest at the resolution of a few kilobases. 4C has been widely used to investigate *cis*-regulatory landscapes of genes, especially in development and disease (Franke et al., 2016). It is well suited for detecting short range regulatory interactions (Symmons et al., 2016), but has also been applied to detect contacts spanning long genomic distances, including whole chromosomes (Loviglio et al., 2016; Simonis et al., 2006).

1.10.2.2 Mapping all contacts occurring within a large genomic region with 5C

In 5C, large genomic regions spanning up to several megabases are amplified from the 3C library by using a complex mix of forward and reverse primers (Figure 1.6). For example, 5C analysis of a 4.5 Mb chromosomal region around the *Xist* gene revealed the presence of TADs (Nora et al., 2012). 5C has the advantage of producing high resolution data at an affordable sequencing depth (~60 million reads per library to obtain 15 – 20 kb resolution for a 1 Mb region (Kundu et al., 2017)). However, the resolution of 5C is dependent on the ability to design forward and reverse primers for all possible restriction fragments across a given locus; in the absence of appropriate primers some mappable fragments will be excluded from the contact map. 5C is often used for the analysis of large genomic regions of several megabases (Kundu et al., 2017; Nora et al., 2012).

1.10.2.3 Mapping all contacts at one or more loci with capture-based methods

Alternatively, a 3C library can be enriched for one or more genomic targets using capture-based methods, such as Capture-C (Hughes et al., 2014), Capture Hi-C (Mifsud et al., 2015) and CAPTURE (Liu et al., 2017). In these approaches, biotinylated oligonucleotides complementary to a genomic region of interest are used to pull-down specific ligation products from the library, which are then amplified and sequenced. Capture-based methods allow the enrichment of interactions for one or more loci of interest from the same 3C library. As a result, these approaches can be used to detect interactions of one viewpoint, but also of entire genomic regions (Franke et al., 2016), or groups of targets (Andrey et al., 2017; Schoenfelder et al., 2018).

1.10.2.4 Mapping all genome-wide contacts with Hi-C and its derivatives

Hi-C is the most commonly used genome-wide approach to map chromatin contacts from a 3C chromatin preparation (Lieberman-Aiden et al., 2009). In this approach, the ends of cross-linked DNA restriction fragments are labelled with biotin and then ligated. After ligation, the exonuclease activity of T4 DNA polymerase is used to remove the biotin label from the ends of unligated fragments. Ligated fragments, which retain the biotin label, are enriched using streptavidin beads, to minimize the number of unligated DNA molecules in the sequencing library (Figure 1.6). Depending on the enrichment efficiency, about 50 to 70% of sequencing reads map to pairs of ligated restriction fragments in Hi-C libraries (Belton et al., 2012). In tethered chromosome capture (TCC (Kalhor et al., 2011)), an early modification of Hi-C, the detection of unspecific ligation events between non-crosslinked material is minimized by tethering the crosslinked and biotinylated chromatin to streptavidin beads before ligation. This approach detects more long-range intra-chromosomal contacts and contacts between chromosomes than standard C-technologies (Kalhor et al., 2011). By contrast, genome conformation capture (GCC (Rodley et al., 2009)), an approach developed at the same time as Hi-C, sequences all DNA present in the 3C library, without preselection of ligated fragments. Although currently much more expensive, especially for large genomes, GCC has the advantage of allowing direct normalisation of DNA abundance, thereby controlling for biases in sequencing and for the presence of genomic alterations, such as CNVs. Methods for detection and normalisation of CNVs have also recently been developed for Hi-C (Dixon et al., 2018; Servant et al., 2018; Vidal et al., 2018).

Many other variants of genome-wide C-methods have been reported, ranging from technical optimizations of the original Hi-C protocol (such as DNase Hi-C (Ma et al., 2015, 2018a) and *in situ* Hi-C (Rao et al., 2014)) and advances to improve resolution (such as Micro-C (Hsieh et al., 2019; Hsieh et al., 2015; Hsieh et al., 2016)), to protocols based on the enrichment of contacts mediated by specific proteins or open chromatin regions (open chromatin enrichment and network Hi-C (OCEAN-C (Li et al., 2018))). Currently, the most commonly used version is *in situ* Hi-C. In the original Hi-C protocol, SDS is used to disrupt the nuclear membrane and ligation of crosslinked DNA therefore occurs partially in solution. *In situ* Hi-C omits this SDS step, allowing ligation of chromatin fragments within the presumably more native environment of the intact nucleus. As a result, the number of random ligation events is reduced and signal-to-noise ratios are improved, thereby reducing the required sequencing depth for higher-resolution contact maps. However, detailed analyses of the nuclear fragments that contribute to contacts in the original version of Hi-C showed that large portions of the chromatin were thought to remain inside the partially digested nucleus during ligation (Gavrilov et al., 2013). Nonetheless, the *in situ* Hi-C protocol is faster and easier than the original version (Rao et al., 2014), mainly because it does not require extensive dilution of the crosslinked chromatin prior to DNA ligation. Consequently, all subsequent steps can be conducted in smaller volumes, allowing more efficient ligation and DNA extraction. Easy Hi-C is another recent approach to simplify Hi-C (Lu et al., 2018). It avoids biotin enrichment and can be used with lower cell numbers than standard Hi-C.

1.10.2.5 Mapping genome-wide contacts in single cells with single-cell Hi-C

Standard Hi-C generates average contact maps from millions of cells, without any possibility to understand heterogeneity of the cell population. Single-cell Hi-C overcomes this limitation by allowing Hi-C contact maps to be produced from individual cells isolated during the process of generating Hi-C libraries (Nagano et al., 2013; Nagano et al., 2015). This approach allows rare cell types to be studied (Flyamer et al., 2017) and helps chromosome structures to be determined at specific stages of the cell cycle (Nagano et al., 2017). The single-cell Hi-C protocol involves *in-situ* proximity ligation of crosslinked and digested chromatin, followed by isolation of single nuclei from the cell suspension and generation of sequencing libraries from each nucleus (Nagano et al., 2017; Nagano et al., 2015). Single-cell combinatorial indexed Hi-C (sciHi-C) adopts a different approach; instead of isolating single cells, DNA within each nucleus is tagged with a unique combination of barcodes (Ramani et al., 2017). First, cells are fixed, lysed and digested with a restriction enzyme. Then the cell suspension of digested, but intact, nuclei is split into 96-well plates, indexed with individual barcodes, pooled, and split again. After several rounds of indexing, *in situ* proximity ligation and library preparation are performed on pooled nuclei, allowing high-throughput generation of single-cell Hi-C libraries.

One of the major challenges in single-cell Hi-C is the efficient recovery of contacts; inefficient digestion and ligation and incomplete recovery of input material result in contact maps that represent only a proportion of the contacts that may exist in a single cell. Modifications of the original protocol increased the average number of contacts detected in one cell from ten thousand up to a range of hundreds of thousands (Nagano et al., 2017; Stevens et al., 2017), but this remained a fraction of the possible contacts in the genome (2-5% of possible ligation products). Recently, the development of Dip-C has increased the number of detectable contacts to an average of one million per cell by omitting biotin incorporation and including a whole genome amplification step in the protocol (Tan et al., 2018a).

1.10.2.6 Combining C-based approaches with chromatin immunoprecipitation

C-methods can be used to study chromatin contacts mediated by specific proteins, such as chromatin modifiers, architectural proteins, members of the transcription machinery, or cell type specific transcription factors. To explore contacts that coincide with chromatin occupancy of specific proteins, Hi-C libraries can be enriched by chromatin immunoprecipitation (ChIP) prior to ligation (Figure 1.6). Early methods, such as ChIP-loop (Horike et al., 2005) and enhanced 4C-ChIP (e4C) (Schoenfelder et al., 2010b), required that chromatin is solubilised to enable specific immunoprecipitation prior to ligation. However, standard 3C conditions often do not fully solubilise chromatin, as nuclei stay mostly intact after SDS treatment (Gavrilov et al., 2013), resulting in low signal-to-noise ratios. Other approaches, such as chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), included sonication of the nuclei, as is more typically used for ChIP (Fullwood et al., 2009). Although sonication allows efficient precipitation of chromatin, its influence on the outcome of the subsequent proximity ligation remains unclear. Challenges in implementing ChIA-PET have led to other strategies for combining ChIP with Hi-C, namely Hi-ChIP (Mumbach et al., 2016) and proximity ligation-assisted ChIP-seq (PLAC-seq) (Fang et al., 2016). Instead of performing protein pulldown followed by ligation of DNA fragments, Hi-ChIP and PLAC-seq start with *in situ* Hi-C and proximity ligation before sonication and immunoprecipitation. In this order, the ligation occurs in intact nuclei under optimal conditions, before chromatin contacts specific to the protein of interest are enriched. Regardless of these increased efficiencies, the results from immunoprecipitated C-libraries should be interpreted carefully, due to the bias towards the genomic regions that are bound by the protein for which the library is enriched (Davies et al., 2017).

1.10.2.7 Genomic resolution of genome-wide C-methods

A major consideration for any genome-wide technique is the genomic resolution. Hi-C data represents interaction frequencies between genomic regions in a contact matrix, consisting of equally sized genomic bins. The bin size (resolution) depends almost entirely on the sequencing depth. Resolutions of 30 kb or lower are often preferred to study chromatin domains and compartments, but also long-range contacts between large genomic regions (e.g.

TADs); using standard Hi-C, this requires sequencing depths of approximately 200 - 400 million reads in mammalian genomes. However, billions of reads become necessary for high (1kb)-resolution datasets of the human genome that can provide detailed insights into 3D genome topology (Rao et al., 2014). Recently, a computational approach, HiCPlus, has applied deep learning to infer high-resolution contact matrices from low-resolution Hi-C data, reducing the sequencing depth to obtain a certain resolution by 1/16 (Zhang et al., 2018b).

1.10.3 Ligation-free detection of contacts

The reliance of 3C-based approaches on the ligation of the ends of DNA fragments found in a cluster of contacts favours the detection of ‘simple’ chromatin contacts which involve two or a small number of genomic regions. In each cluster of contacting DNA fragments isolated during a 3C-based experiment, each DNA fragment can only capture two other contacting fragments (O'Sullivan et al., 2013). In complex DNA clusters, each fragment ligates with one or two other fragments out of many simultaneously contacting fragments. Therefore, the interactome of a DNA fragment is diluted across the cell population by the choice of only one or two other fragments during ligation. Recently, three ligation-free approaches have been developed for genome-wide mapping of chromatin contacts, GAM (Beagrie et al., 2017), SPRITE (Quinodoz et al., 2018), and ChIA-Drop (Zheng et al., 2019). These methods are orthogonal to ligation-based approaches and are starting to provide new insights into 3D genome topology. Other ligation-free approaches – tyramide signal amplification (Chen et al., 2018a) (TSA-seq) and DNA adenine methyltransferase identification (Marshall et al., 2016; van Steensel and Henikoff, 2000; Vogel et al., 2007) (DamID) – map chromatin with respect to nuclear landmarks (such as the nuclear lamina or various nuclear bodies), thereby helping to define chromatin positions in 3D space.

1.10.3.1 Mapping contacts with nuclear structures with DamID and TSA-Seq

DamID is an in vivo genome-wide method for detecting interaction sites between a protein of interest and DNA. The DNA binding domain of the protein of interest is fused to the DNA adenine methyltransferase (Dam) protein from *E. coli* (Marshall et al., 2016; van Steensel and Henikoff, 2000; Vogel et al., 2007), which specifically methylates adenines in the sequence GATC. When the fusion protein is expressed at low levels in cells, GATC sequences within or close to DNA binding sites of the protein of interest are marked by methylation. After DNA extraction, the methylated GATC sites are cut with a methylation-sensitive restriction enzyme and adapters are added to the restriction fragments to ensure only methylated binding sites are amplified and sequenced. In an interesting adaptation called Targeted DamID (TaDa) (Marshall et al., 2016), expression of the Dam fusion protein is restricted to a specific cell type of interest, using targeted expression systems (e.g. Gal4), which allows detection of DNA-protein interactions, e.g with RNA polymerase II, in a cell-type specific manner without prior isolation or sorting of cells. DamID has been successfully used to study DNA interactions with proteins, such as Lamin B1, which resulted in the genome-wide mapping of Lamina-associated domains (LADs), and provided spatial information about chromatin with regard to the nuclear periphery (Guelen et al., 2008; Peric-Hupkes et al., 2010). However, interactions between chromatin and other nuclear compartments, such as splicing speckles, are not readily detected with DamID because most of the DNA surrounding the compartments is not directly binding to the tagged proteins (Spector and Lamond, 2011).

TSA-Seq addresses this problem by using tyramide signal amplification (TSA) to measure the distances between chromatin and nuclear compartments (Chen et al., 2018b). In this approach, horseradish peroxidase (HRP) is conjugated to an antibody that binds to a protein specific to the nuclear compartment of interest, where it catalyses the production of biotin-conjugated tyramide free radicals, which diffuse and bind to nearby macromolecules – including DNA. Biotin-labelled DNA can be subsequently selected by biotin pull-down and sequenced to identify all genomic regions that were close enough to the protein of interest to be labelled. With TSA-seq spatial distances of chromatin with splicing speckles were determined genome-wide, mapping the distances of all genes to their nearest speckle (Chen et al., 2018b).

Another recent adaptation of DamID, called DamC, detects 4C-like contacts between a target region and the surrounding DNA regions, up to a few hundred kilobases (Redolfi et al., 2019). In DamC, Dam is fused with the reverse tetracycline receptor (rTetR), which binds to Tet operator sites inserted at the genomic region of interest. The Dam fusion protein methylates the target and its interaction partners *in vivo*. When combined with high-throughput sequencing, DamC reveals chromatin contacts independently of crosslinking or ligation, but requires engineering of the cells of interest unlike the other C-methods and ligation free approaches. The comparison with 4C and Hi-C showed high similarities at the level of TADs and CTCF loops at many genomic sites, however, some alterations of loops and sub-TAD structures could also be observed (Redolfi et al., 2019).

1.10.3.2 Mapping all genome-wide contacts with GAM

In GAM, nuclei are sectioned in random orientations from a population of fixed and sucrose-embedded cells using ultra-thin cryosectioning (220-230 nm thickness). Single nuclear slices are then isolated directly from the cryosection by laser microdissection. GAM thus avoids cell extraction or sorting, both of which can disrupt cellular and nuclear structures, which can be especially important when analysing complex tissues. The DNA from every slice is extracted, whole genome amplification is performed and indexed sequencing adapters are added before the DNA from all slices is pooled for sequencing (Figure 1.7). From the sequencing data for several hundred nuclear sections, each from a single cell, chromatin contacts between pairs of DNA loci can be inferred by counting their co-segregation frequency (that is, how often are the two loci contained in the same nuclear sections). Genomic regions that are closer in 3D space are more frequently found in the same nuclear slices. To detect statistically significant interactions, GAM was combined with a mathematical model, statistical inference of co-segregation (SLICE) (Beagrie et al., 2017). The most specific chromatin contacts detected with SLICE were found to contain active genomic regions, such as active enhancers and actively transcribed genes, with these contacts extending over megabases up to entire chromosomes (Beagrie et al., 2017). SLICE separately models the random interactions that depend on genomic distance and the specific interactions that occur at a given physical distance (for example, below 100 nm; ref. (Beagrie et al., 2017)); it interrogates which pairs of loci co-segregate more often in the collection of slices than expected from random contacts, and quantifies the frequency of specific interactions in the cell population. GAM also allows genome-wide interactions between three or more DNA loci to be detected simultaneously, and found long-range contacts between TADs containing super-enhancers and highly-transcribed TADs (Beagrie et al., 2017). The resolution of GAM datasets depends on the number of nuclear slices collected. With 400 nuclear slices, sequenced with ~1 million reads per slice, it was possible to achieve a resolution of 30 kb for pairwise chromatin contacts (Beagrie et al., 2017), comparable to a Hi-C library with similar sequencing depth (Dixon et al., 2012). Larger GAM datasets comprising a few thousand nuclear slices will help define the maximal resolution that can be practically afforded by GAM.

1.10.3.3 Mapping all genome-wide contacts with SPRITE and ChIA-Drop

SPRITE (Quinodoz et al., 2018) and ChIA-Drop (Zheng et al., 2019) detect chromatin interactions by tagging crosslinked chromatin complexes. Similar to 3C-based approaches, these methods rely on mild fixation and fragmentation of chromatin inside the nucleus – but unlike 3C-based approaches, they do not use proximity ligation. Instead, in SPRITE, the crosslinked chromatin fragments are split across a 96-well plate, where each well contains a unique barcode (Figure 1.7). The indexed chromatin complexes are re-pooled, followed by sequential rounds of splitting, barcoding and pooling. The DNA (and RNA) molecules within an individual chromatin complex are identified after sequencing by their unique combination of barcodes added using this split-pool strategy; only DNA fragments that were crosslinked with each other will display the same combinations of barcodes. In ChIA-Drop (Figure 1.7), crosslinked and fragmented chromatin is separated into single chromatin complexes by droplet formation using a microfluidics device. Each droplet contains reagents for barcoding and amplification, and barcoded complexes are pooled and sequenced, as in SPRITE. SPRITE detects TADs and loop domains, both of which are features of Hi-C contact maps. However, SPRITE also detects additional genome-wide features of nuclear architecture, such as the association of specific genomic regions with nucleoli and splicing speckles. The predominant chromatin hubs around these nuclear bodies contain genomic regions from different chromosomes, an observation that coincides with single cell imaging (Pombo and Branco, 2007), but that had not been made using 3C-based assays. SPRITE also detects long-range contacts between regions containing active genes and super-enhancer regions that were first recognised as being multiway-specific interactions in studies using GAM (Beagrie et al., 2017).

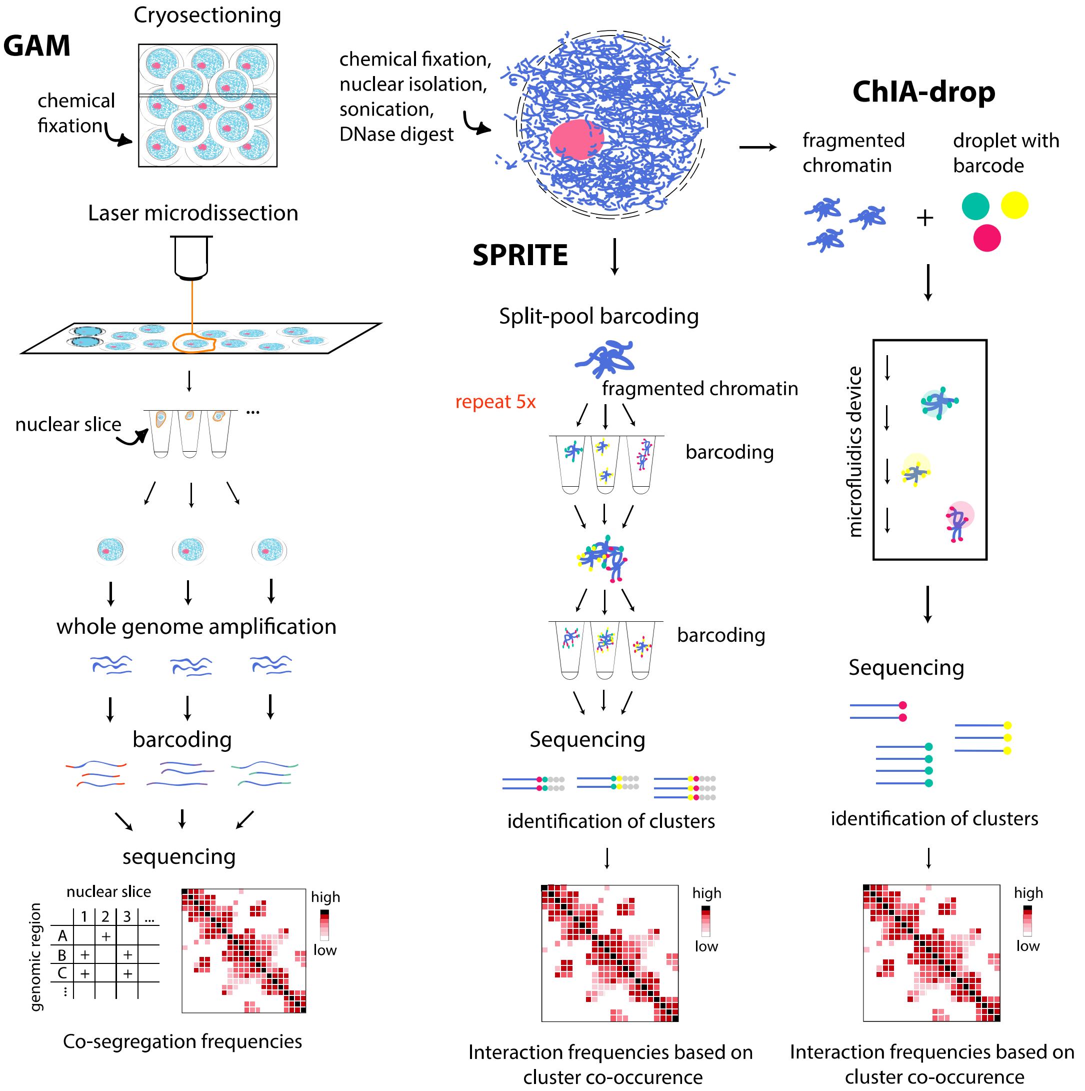


Figure 1.7: Ligation-free methods to map chromatin contacts genome-wide.
Genome Architecture Mapping (GAM) measures co-segregation frequencies of genomic regions by slicing the nucleus into thin nuclear sections and sequencing the DNA content of a large number of randomly collected slices. To obtain nuclear slices, cells are fixed and cryosectioned. Single nuclear slices are isolated from the cryosection using laser microdissection. DNA is extracted from each nuclear slice by whole genome amplification and sequenced. The sequence information is used to score the presence or absence of genomic loci in each slice. Spatial proximity of all pairs of loci in the genome is inferred from the frequency of their co-occurrence in the population of slices. Split-pool recognition of interactions by tag extension (SPRITE) detects chromatin interactions of multiple genomic regions by tagging single crosslinked chromatin complexes with unique combinations of identifiers before sequencing. Cells are fixed and the crosslinked chromatin is fragmented using sonication. The resulting chromatin complexes are split onto a 96-well plate, and DNA in each well is ligated to a unique adapter. All wells are pooled and split again into 96 wells, followed by adapter ligation. The process is repeated five times until each chromatin complex is labelled with a unique combination of adapter sequences. DNA is purified and sequenced, and the adapter combination of each sequenced DNA fragment is used to identify all genomic regions with the same combination of adapters that were initially crosslinked together, hence that were in spatial proximity. ChIA-Drop detects chromatin contact by barcoding crosslinked chromatin complexes, after cell fixation, lysis, and chromatin fragmentation. Barcodes are delivered in droplet containing a unique identifier, and reactions for adapter ligation and DNA amplification. Each chromatin complex is loaded onto a droplet in a microfluidics device and sequenced. Barcodes identify regions from the same droplets, meaning regions that were crosslinked due to spatial proximity. Figure adapted from Kempfer and Pombo (2019).

Table 1.1: Comparison of methods used to detect chromatin contacts.

3-C-based methods

Assay	Description	Number of contacts	Multiplicity of contacts	Single cell information	Number of cells	Detectable contacts
3C	proximity ligation and selection of target regions with primers, detection by quantitative PCR	one vs one	pairwise	no	100 million (Naumova et al., 2012)	protein mediated
4C	proximity ligation and enrichment for contacts with one bait region by inverse PCR, detection by sequencing	one vs all	pairwise	no	robust: 10 million (van de Werken et al., 2012a); low input: 340 K (Schwartzman et al., 2016)	protein mediated
5C	proximity ligation and enrichment for larger target region with primers, detection by sequencing	many vs many	pairwise	no	robust: 50-70 million (Dostie and Dekker, 2007); low input: 2 million (Kim et al., 2018)	protein mediated
Hi-C	proximity ligation and enrichment for all ligated contact pairs, detection by sequencing	all vs all	pairwise	no	robust: 2-5 million (Rao et al., 2014) low input: 100-500 K (Belaghzal et al., 2017; Lu et al., 2018)	protein mediated
TCC	tethered proximity ligation and enrichment for all ligated contact pairs, detection by sequencing	all vs all	pairwise	no	25 million (Kalhor et al., 2011)	protein mediated
PLAC-seq, ChIA-PET	proximity ligation and pulldown of specific protein-mediated contacts, detection by sequencing	many vs many	pairwise	no	robust: 100 million (Li et al., 2017a) low input: 500 K (Fang et al., 2016)	protein mediated (specific)
Capture-C, C-HiC	proximity ligation and target enrichment using probes for genomic regions of interest, detection by sequencing	many vs all	pairwise	no	robust: 100 K (Davies et al., 2016) low input: 10-20 K (Oudelaar et al., 2017)	protein mediated
single-cell Hi-C	proximity ligation and enrichment for all ligated contact pairs, detection by sequencing	all vs all	pairwise	yes	hundreds	protein mediated

Imaging-based approaches

Assay	Description	Number of contacts	Multiplicity of contacts	Single cell information	Number of cells	Detectable contacts
2D-FISH	fixation to flatten cells, hybridisation of fluorescent probes to target regions, measurement of 2D spatial distances	between 2 and 52 regions*	pairwise or more	yes	hundreds	all in spatial proximity
3D-FISH	fixation of cells, hybridisation of fluorescent probes for target regions, measurement of 3D spatial distances	between 2 and 52 regions*	pairwise or more	yes	hundreds	all in spatial proximity
cryo-FISH	fixation of cells, cryosectioning, hybridisation of fluorescent probes for target regions, measurement of 2D spatial distances	between 2 and 52 regions*	pairwise or more	yes	hundreds	all in spatial proximity
Live cell imaging	fluorescent labelling of genomic loci in living cells, measurement of spatial distances over time	between 2 and 12 regions	pairwise or more	yes	hundreds	all in spatial proximity

Ligation-free methods

GAM	cryosectioning of fixed cells, DNA extraction from nuclear sections and sequencing, inferring spatial distances from co-segregation of genomic regions in nuclear sections	all vs all	pairwise or more	yes	hundreds (Beagrie et al., 2017)	all in spatial proximity
SPRITE	fixation of cells, identification of crosslinked chromatin fragments by split pool barcoding and sequencing	all vs all	many	no	10 million (Quinodoz et al., 2018)	protein-mediated
ChIA-Drop	fixation of cells, identification of crosslinked chromatin fragments by droplet-based and barcode-linked sequencing	all vs all	many	no	10 million (Zheng et al., 2019)	protein-mediated

* Classical FISH experiments rarely distinguish between more than 2-5 differentially labelled regions simultaneously (Shimizu et al., 2015). Cycles of probe hybridisation can increase this number up to 52 (Müller et al., 2002).

1.11 Comparing approaches to map chromatin contacts

Fundamental differences exist between current approaches for mapping 3D genome folding, including how the chromatin is fixed and prepared, their power to detect multiple chromatin contacts, or contacts with different spatial distances and protein occupancy, and their ability to detect long-range contacts within the same or different chromosomes. These differences have sometimes led to observations that can be difficult to reconcile between the different approaches.

1.11.1 Fixation and chromatin preparation

With the exception of live-cell methods (such as DAM-based and CRISPR-based approaches), all chromatin folding techniques start by crosslinking DNA-protein complexes to stabilise nuclear structures. Chemical fixation using formaldehyde is the most common approach for crosslinking, but concentrations, buffers and fixation times vary widely; for example, 1% formaldehyde is typically used for C-methods, 4% for most DNA-FISH experiments in whole cells, and 8% for GAM or cryo-FISH in nuclear slices. Other fixatives include solvent-based precipitation using ethanol, methanol or acetone. A recent imaging study compared the effects of formaldehyde fixation and cryofixation on nuclear structure using partial wave spectroscopy (Li et al., 2017b). It revealed that weaker fixatives (such as 4% formaldehyde in PBS) introduce larger structural distortions than stronger fixatives (such as glutaraldehyde, often used for electron microscopy). However, the distinction between condensed and decondensed chromatin remains detectable at the population level (Li et al., 2017b), which is consistent with the ability of all current chromatin folding methods to successfully map euchromatin and heterochromatin. The effect of varying crosslinking conditions (from no fixation to 5% formaldehyde fixation) has been examined in Capture-C experiments; similar short range-interactions were detected under all conditions, but formaldehyde concentrations below 2% gave increased efficiency of detection (Oudelaar et al., 2017). Our own previous work showed that the organisation of the active form of RNA polymerase II, which marks transcription sites, can be highly disrupted with weaker fixatives, but not with the fixation regimen used for GAM or cryoFISH (Guillot et al., 2004). In FISH, denaturation of the DNA by heat and formamide induces fine structural changes in chromatin folding, such as slight distortions of the interchromatin space (Markaki et al., 2012; Solovei et al., 2002). However, 3D-FISH preserves the organisation of centromeres seen by imaging the same cells before and after hybridisation (Markaki et al., 2012), and cryoFISH retains the organisation of active RNA polymerase II sites (Xie et al., 2010). In ‘resolution after single-strand exonuclease resection’ (RASER-)FISH, heat denaturation of the DNA is avoided and DNA accessibility is achieved by exonuclease digestion, thereby reducing the effects of DNA denaturation (Brown et al., 2018).

1.11.2 Multiplicity of chromatin contacts

The dependency of 3C-based methods on DNA-end ligation results in a preferential detection of lower multiplicity contacts that involve fewer genomic regions (O'Sullivan et al., 2013). However, in every 3C library a proportion of all ligation events occurs between more than two DNA fragments. Current methods to capture these higher complexity ligation events include multi-contact 4C (MC-4C) (Allahyar et al., 2018), which uses long-read sequencing (such as Nanopore sequencing) of 4C libraries to capture 3-way contacts of a region of interest, and chromosomal walks (C-walks) (Olivares-Chauvet et al., 2016), which implement multiple ligation steps, followed by dilution and barcoding of the isolated ligation products. Alternatively, methods such as the concatemer ligation assay (COLA) (Darrow et al., 2016) and Tri-C (Oudelaar et al., 2018) generate 3C libraries with a restriction enzyme that cuts small DNA fragments, which increases the frequency of detecting multiple ligation events in one sequencing read. Estimates based on direct comparison of pairwise and multiway ligation events indicate that only 17 % of chromatin contacts in mouse ESCs are pairwise contacts and, therefore, the majority of the genome is involved in higher-order contacts between more than two genomic loci (Olivares-Chauvet et al., 2016). These observations are supported by a recent study using SPRITE that showed that classical ligation-dependent methods under-represent higher complexity contacts (Quinodoz et al., 2018). Assays that do not depend on ligation detect DNA fragments that are in spatial proximity regardless of the number of interacting genomic loci. For example, long-range multiway contacts between genomic regions harbouring super-enhancers were readily found in GAM (Beagrie et al., 2017), FISH (Beagrie et al., 2017) and SPRITE (Quinodoz et al., 2018) data, but had not previously been detected with Hi-C. Furthermore, analyses of triplet interactions between TADs in GAM showed that multiple interactions between super-enhancer regions and active genes are a common feature of genome conformation in mouse ESCs (Beagrie et al., 2017).

1.11.3 Spatial distance between contacting genomic regions

The spatial distance between genomic loci is thought to influence the probability of ligation irrespective of the frequency of contacts. Whereas cryo-FISH and SPRITE have readily detected abundant inter-chromosomal contacts in human, mouse and Drosophila cells (Branco et al., 2008; Branco and Pombo, 2006; Rosin et al., 2018; Tan et al., 2018a), C-based methodologies are more often used to explore specific contacts within chromosomes, with some exceptions (Loviglio et al., 2016; Monahan et al., 2019; Simonis et al., 2006; Spilianakis et al., 2005; Tan et al., 2018a). Recent CRISPR-Cas9 live-cell imaging of a small number of chromatin contacts within and between chromosomes showed that inter-chromosomal contacts display spatial distances in the range of ~280 nm, in contrast to distances of ~190 nm for intrachromosomal interactions (Maass et al., 2018a). Interestingly, only the intrachromosomal contacts could be observed in matching Hi-C data, indicating a dependency of close spatial distances for successful proximity ligation.

1.11.4 Protein-mediated interactions versus bystander contacts

GAM and all imaging-based techniques collect all possible spatial relationships between genomic regions, regardless of their involvement in a protein-mediated interaction, and allow sampling of the whole range of spatial distances within the interphase nucleus. Thus, contacts detected by these methods do not only reflect protein-mediated chromatin interactions, but also detect bystander contacts. However, it is possible to identify the most specific contacts through effective sampling to take into account all the behaviours of all genomic regions at all linear distances across the cell population. In this regard, GAM currently has more statistical power than FISH as it samples all possible combinations, whereas FISH remains limited to the analyses of a subset of regions or chromosomes.

1.11.5 Levels of concordance between different methods

The validation of results obtained by C-methods often entails the use of DNA-FISH on a few selected loci. Many examples show agreement between 3C interaction frequencies and spatial distances measured by FISH, especially at large genomic distances (Giorgetti and Heard, 2016; Hakim et al., 2011; Lieberman-Aiden et al., 2009; Rao et al., 2014; Tang et al., 2015; Wang et al., 2015). Loci in the same TAD are often closer in nuclear distance than loci in different TADs (Dixon et al., 2012; Nora et al., 2012), and interaction frequencies obtained from Hi-C correlate with spatial distances at and above the TAD level (Wang et al., 2015). A linear relationship between Hi-C contacts and FISH distances was found by investigating the physical distances between all TADs along a chromosome (Wang et al., 2015). An overall correlation between Hi-C interactions and median spatial distance measured by high-throughput FISH have recently been shown for 90 pairs of loci. However, the range of physical distances between genomic regions containing Hi-C interactors (with high-ligation frequency) and non-interactors (with low ligation frequency) overlap extensively, with about 20% of distances being closer between two non-interactors than two interactors (Finn et al., 2019). Thus, Hi-C captures spatial proximity but Hi-C interactions are not easily translated into physical distances. Other comparisons between FISH and C-methods have also found non-trivial relationships between physical distance distributions and population-average interaction frequencies (Giorgetti and Heard, 2016) and show that contact frequency is distinct from average spatial distance, both in polymer simulations and in experimental data (Fudenberg and Imakaev, 2017).

The use of FISH to validate Hi-C results has helped investigate false positives in Hi-C data, assuming FISH is correct, but is not a valid strategy for an unbiased search for contacts that are missed by Hi-C (that is, false negatives). Thus, any underrepresented contacts in Hi-C data have so far not been systematically studied. The development of orthogonal genome-wide ligation-free approaches, such as GAM and SPRITE, have been able to identify new aspects of 3D genome folding that had not been detected by Hi-C but which are fully validated by FISH (Beagrie et al., 2017; Quinodoz et al., 2018). The first and relatively small GAM dataset identified specific long-range contacts across linear genomic distances that span tens of megabases, which involve active and enhancer-rich genomic regions (Figure 1.8a). One promising outcome of the emergence of these orthogonal approaches is the development of analysis tools that use the information they generate about such long-range contacts to discover the same contacts in Hi-C data (Figure 1.8a). In this regard, it is interesting to note that CTCF depletion in human cells results in the detection by Hi-C of long-range contacts between super-enhancers (Rao et al., 2017), which raises the possibility that CTCF-mediated contacts may be preferentially detected by Hi-C in normal conditions, but once CTCF-dependent interactions are lost, other underlying folding patterns, including long-range contacts, become easier to detect.

The first SPRITE dataset has also highlighted novel aspects of 3D folding that are not readily captured by Hi-C (Quinodoz et al., 2018). By discriminating contacts according to their multiplicity, SPRITE shows a contact decay with genomic distance that is very similar to Hi-C when considering only low complexity SPRITE clusters (2-10 genomic regions per contact hub; Figure 1.8c). By contrast, SPRITE shows striking abundance of long-range contacts when considering also higher-order contacts, which confirms early theoretical predictions that ligation-based approaches are biased to the detection of more simple 3D chromatin contacts (O'Sullivan et al., 2013). Although GAM and SPRITE are orthogonal methodologies, their frequency of contacts relative to genomic distance are remarkably concordant (Figure 1.8c; the data represented was obtained from refs. (Beagrie et al., 2017; Quinodoz et al., 2018)).

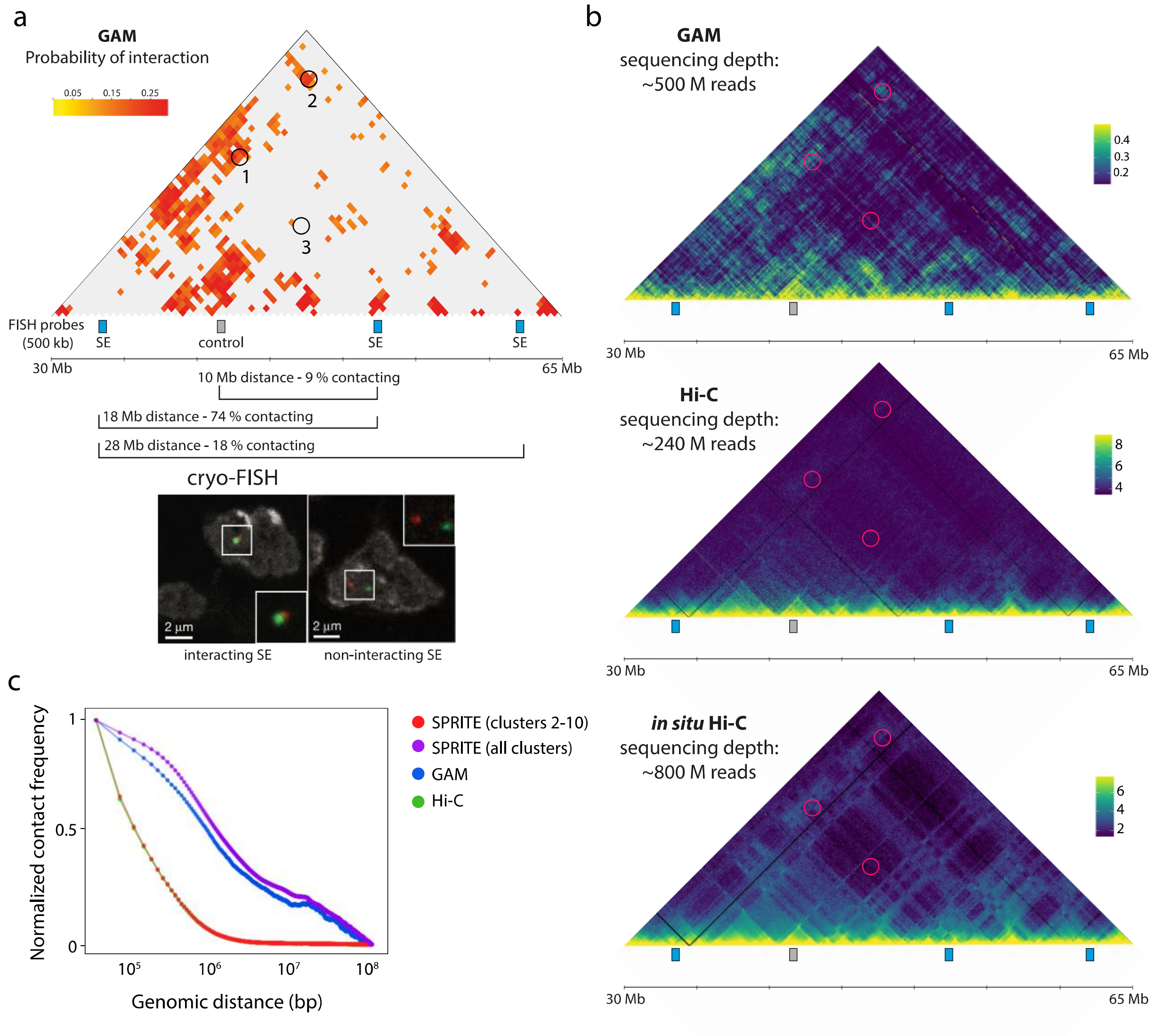


Figure 1.8:Comparison of long-range chromatin contacts across methods.

(a) Genome Architecture Mapping (GAM) detects significant interactions between super-enhancers (SE, circles 1 and 2), spanning large genomic distances (18 Mb and 28 Mb). The heatmap shows GAM interaction probabilities for chromosome 11 (region 30-65 Mb) in mouse embryonic stem cells (ESCs) at 500 kb resolution (Beagrie et al., 2017). Contacts between the super-enhancer regions also can be detected using cryo-fluorescence *in situ* hybridisation (cryo-FISH), which confirmed their interactions in a high percentage of cells in the population. While 74% of cells had the 18 Mb distant super-enhancer contact, and 18% of cells had the 28 Mb distant contact, only 9% of the cells revealed low spatial distances for a chosen 10 Mb distant control region with one of the super-enhancers (circle 3), which was not detected by GAM. The images show DAPI stained cryosections with interacting and non-interacting 500 kb super-enhancer regions (Beagrie et al., 2017). (b) Long-range super-enhancer contacts can also be found when looking at GAM contacts without filtering for the most significant interactions (~500 million reads, 40 kb resolution, mESCs, data, Beagrie et al., 2017), and although they are not readily detected in Hi-C data with average sequencing depth (~240 million reads, 40 kb resolution, mESCs, data, Dixon et al., 2012), they start to emerge in deep-sequenced *in situ* Hi-C data (~800 million reads, 50 kb resolution, mESCs Bonev et al., 2017). Heatmaps are visualised from the published, normalised matrix files. Scores are color-coded, where the color-code range (maximum and minimum cut-off) is determined by the mean value of the bin distances 1 to 20 and -50 to -30 from the diagonal, respectively. (c) The plot shows the distribution of contact frequencies detected by Hi-C, GAM and SPRITE (all clusters or clusters with 2-10 reads) along the linear genomic distance of chromosome 11 in mESCs, scaled to the maximum observed value in each dataset. The ligation-free methods GAM and SPRITE detect similar ranges of chromatin contacts, which can extend over large genomic distances. By contrast, Hi-C contacts typically extend over shorter genomic distances. However, SPRITE data can be sorted based on the number of interactions within one chromatin complex. When considering only small SPRITE clusters with fewer than 10 genomic regions in the same chromatin cluster, the range of detection between Hi-C and SPRITE is comparable, indicating that Hi-C favours less complex short-range contacts over long-range interactions involved in chromatin hubs with many interaction partners. Plot was generated by Christoph Thieme (laboratory of A. Pombo) using the same data used in panel b for GAM (Beagrie et al., 2017) and Hi-C (Dixon et al., 2012), and normalised SPRITE clusters for chr 11, according to Figure 3B of ref. (Quinodoz et al., 2018). Figure adapted from Kempfer and Pombo (2019).

1.11.6 Limitations and applications of different methodologies

The use of proximity ligation adds limitations because ligation has low efficiency, and is potentially affected by the local distance in the cluster between two ends of DNA or the topology of the two ends within the cluster. SPRITE also depends on ligation of a small oligo to each DNA end in a contact cluster; however, it is no longer dependent on the physical distance between two DNA fragments in the cluster, which allows mapping of all contacts within one chromatin complex. In C-methods and SPRITE, detection of contacts also depends on the efficiency of the fragmentation step to expose the DNA end. In GAM, there is no DNA restriction digest or ligation, and the detection of DNA depends on its extractability and sequencing depth. C-methods, GAM, SPRITE and FISH can be applied directly to cells, tissues or organisms, whereas CRISPR and Lac/Tet operator -based imaging and DAM-related methods require genetic engineering of cell lines or whole organisms, and will not be suitable for the analyses of most human biopsies.

Each of the assays discussed here have different limitations and applications, and thus contribute to our current understanding of 3D genome folding in different ways. 3C-based techniques have the advantage of providing enormous amounts of chromatin contact information in one comparably simple biochemical experiment, although they may require high-depth sequencing when aiming for high resolution. 3C-based methods, and in particular proximity ligation itself, also have important limitations that favour the detection of more simple contacts over higher-order chromatin contacts, which can lead to misunderstanding of the importance and abundance of certain interactions.

Imaging and ligation-free methods have the ability to detect chromatin contacts at all scales of chromosome folding, including contacts between chromosomes. GAM and SPRITE can be readily used for sequence-unbiased genome-wide explorations, whereas detection of contacts with DNA-FISH remains limited to pre-selected loci and is most often used to validate findings from genome-wide techniques. Imaging fluorescently labelled chromatin loci in live cells with CRISPR-based techniques will improve our understanding of possible artefacts resulting from chromatin preparation or fixation. Other developments based on cryo-focused ion beam (cryo-FIB) milling of intact, frozen cells (Mahamid et al., 2016), or cryolysis (Aitchison and Rout, 2015) also hold the potential of devising fixation-free versions of GAM and SPRITE that sample through fractionated frozen nuclei.

1.12 Aims of the thesis

The main goal of this thesis is to explore two major aspects of genome topology which have been understudied, mostly due to limitations of current methodologies, namely allele-specific chromatin folding and complex interactions that span large genomic distances in and between chromosomes. In short, using GAM and Hi-C in specific model systems with intrinsic controls, I aimed to advance our understanding of the functionality and importance of allelic differences and inter-chromosomal contacts in mammalian gene regulation. While Hi-C has been the method of choice for studies of 3D genome folding, the original version of GAM was not suitable for collection of high-resolution, larger datasets. Thus, another goal of the work presented in this thesis was the development of a high-throughput version of GAM that allows the collection of large datasets for subsequent high-resolution analysis of chromatin contacts and for determining allele-specific interactions.

Allele specificity of chromatin folding has been studied with 3C-based methodologies, which revealed few differences in the genome-wide folding of the parental alleles. The work presented in this thesis suggests that low efficiency in assigning contacts to their haplotypes makes it difficult to study allelic differences. I aimed to test whether phasing of parental genomes is more efficient in GAM, and further explore the possibility of gene regulatory functions of allele-specific genome topology.

I set out to gain insight into the functionality of long-range chromatin contacts by genome-wide mapping of DNA interactions in the context of congenital disease. Here, I chose to study the effects of a genomic deletion at the human 16p11.2 locus, associated with ASD, on gene expression in neuronal differentiation. With simultaneous mapping of chromosome topologies using Hi-C and collection of transcriptomic data using RNA-seq, I aimed to decipher mechanisms of gene regulation via long-range chromatin interactions in the context of the disease.

2. Materials and methods

2.1 Oligonucleotides

Primers and adapters were HPLC-purified, purchased from biomers.net.

Table 2. 1: Primer and adapter sequences.

Primers and adapters for Hi-C	Sequence
3C-Gapdh_for (ligation control)	5'TATCAAGGGTGCCCGTCACCTTCAGCTTTC
3C-Gapdh_rev (ligation control)	5'GGGCTTTTATAGCACGGTTATAAAGTGG
Gapdh-no-HindIII_for (digestion control)	5'AGCCATCAGCTATGCACGTA
Gapdh-no-HindIII_rev (digestion control)	5'GACTTGGAGGAGGTTTGCTG
Ncapd2-HindIII_for (digestion control)	5'CGCCAGTTTAGAAGCAGCTC
Ncapd2-HindIII_rev (digestion control)	5'TGTGCGATCTAACCTCATGG
Gapdh-HindIII_for (digestion control)	5'TGGAGGTTTCTTTCCTGTCC
Gapdh-HindIII_rev (digestion control)	5'TCCCCTTAGTTCGAGGGACT
P5 universal adapter (TruSeq)	5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7 index adapter 5 (TruSeq)	5' GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
P7 index adapter 6 (TruSeq)	5' GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG
P7 index adapter 12 (TruSeq)	5' GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG
PCR Primer 1.0 (P5) (TruSeq)	5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA
PCR Primer 2.0 (P7) (TruSeq)	5' CAAGCAGAAGACGGCATACGAGAT
Primers for GAM	Sequence
GAT-7N	5' GTG AGT GAT GGT TGA GGT AGT GTG GAG NNN NNN N
GAT-COM	5' GTG AGT GAT GGT TGA GGT AGT GTG GAG

2.2 Cell lines

- Mouse 46C ESCs (Ying et al., 2003) were kindly provided by Domingos Henrique (Institute of Molecular Medicine, Lisbon, Portugal).

- Mouse 16p ESC lines +/+*, df/+, df/dp* (Horev et al., 2011) were kindly provided by Alea Mills (Cold Spring Harbour Laboratory, Cold Spring Harbour, NY, USA). A diagram describing the deletion and duplication is presented in Chapter 5, Figure 5.1).
- Mouse F123 ESCs (a male, hybrid cell line, derived from S129/Jae and Cast) (Gribnau et al., 2003) were kindly provided by Bing Ren (University of California San Diego, San Diego, CA, USA).

2.3 Antibodies

2.3.1 Primary antibodies

- GATA-6 AF1700, goat polyclonal, IgG, (0.2 mg/ml), R&D Systems
- Nanog (14-5761), Lot: E04325-1653, Clone:eBioMLC-51, rabbit monoclonal, IgG, (0.5mg/ml), Invitrogen Antibodies, eBioscience
- Oct3/4 (c-10)x sc-5279, Lot: A2315, mouse monoclonal, IgG2b, (200µg/ 0.1ml), Santa Cruz
- Pan Histone (MAB3422), Clone: H11-4, mouse monoclonal, Merck

2.3.2 Secondary antibodies

- AlexaFluor488 (A21202), donkey anti-mouse IgG (H+L), Invitrogen
- AlexaFluor488 (A21208), donkey anti-rabbit IgG (H+L), Invitrogen
- AlexaFluor488 (A11055), donkey anti-goat IgG (H+L), Invitrogen
- AlexaFluor488 (A11001) goat anti-mouse IgG (H+L), Invitrogen

2.4 Kits

- KAPA Library Quantification Kit (Roche, 07960140001)
- TruSeq® Stranded Total RNA Library Prep Human/Mouse/Rat (Illumina, 20020596)
- PCR Mycoplasma test kit (PanReac AppliChem, A3744)
- Agilent RNA 6000 Nano Kit (Agilent Technologies, 5067-1511)
- WGA-4 GenomePlex® Single Cell Whole Genome Amplification Kit (Sigma Aldrich, 254-457-8)
- REPLI-g Mini Kit (Qiagen, 150023)
- MALBAC® Single Cell WGA Kit (Yikon Genomics, EK100101210)
- Ampli-1 WGA kit (Menarini Silicon Biosystems)
- MinElute PCR Purification Kit (Qiagen, 28004)

- MinElute 96 UF PCR Purification Kit (Qiagen, 28051)
- Quant-iT® PicoGreen dsDNA Assay Kit (ThermoFisher, P7589)
- Agencourt® AMPure® XP magnetic beads (Beckman Coulter, A63880)
- Qubit™ dsDNA HS Assay Kit (Life Technologies, Q32851)
- Agilent High Sensitivity DNA Kit (Agilent Technologies, 5067-4626)
- Illumina Nextera XT library preparation kit:

  Nextera® XT DNA Library Preparation Kit (Illumina, FC-131-1096)

  Nextera® XT Index Kit v2 Set A (Illumina, FC-131-2001)

  Nextera® XT Index Kit v2 Set B (Illumina, FC-131-2002)
- NextSeq 500/550 High Output v2 kit (75 cycles) (Illumina, TG-160-2005)
- NextSeq 500/550 High Output v2 kit (150 cycles) (Illumina, TG-160-2002)

2.5 Published datasets

Table 2. 2: Published datasets.

Data type	Specification	Cell type	Availability
DamID	Lamina associated domains identified with Lamin B1	Mouse embryonic stem cells (E14Tg2A)	(Peric-Hupkes et al., 2010) GEO: GSE17051
ATAC-seq	ATAC peaks	Mouse embryonic stem cells (E14)	ENCODE reference epigenome project GEO: GSM3109355
ATAC-seq	ATAC peaks	Mouse embryonic stem cells (F123)	(Juric et al., 2019) GEO: GSE119663
ChIP-seq	CTCF peaks	Mouse embryonic stem cells (F123)	(Juric et al., 2019) GEO: GSE119663
ChIP-seq	H3K27ac peaks	Mouse embryonic stem cells (F123)	(Juric et al., 2019) GEO: GSE119663
ChIP-seq	H3K4me1 peaks	Mouse embryonic stem cells (F123)	(Juric et al., 2019) GEO: GSE119663
ChIP-seq	H3K4me3 peaks	Mouse embryonic stem cells (F123)	(Juric et al., 2019) GEO: GSE119663
GAM	Sequencing data from nuclear profiles	Mouse embryonic stem cells (46C)	(Beagrie et al., 2017) GEO: GSE64881
Hi-C	Normalised Hi-C contact frequencies	Mouse embryonic stem cells (F123)	(Kubo et al., 2017) 4DN data portal: 4DNESKKSKG7Y
Hi-C	Raw Hi-C reads	Mouse embryonic stem cells (J1)	(Dixon et al., 2012) GEO: GSE35156
Hi-C	Normalised Hi-C contact frequencies	Mouse neuronal precursor cells (E14, 60 h differentiation)	(Bonev et al., 2017) 4DN data portal: 4DNFIDWM3HN5

2.6 Processing of published datasets

Single-end ChIP-seq reads were downloaded and aligned to the Mouse reference genome mm10 (Dec. 2011, GRCm38/mm10) using Bowtie2 v2.0.5 (Langmead and Salzberg, 2012). The reference genome was indexed, and the alignments were performed with default parameters. Replicated reads (i.e., identical reads, aligned to the same genomic location), occurring more often than the 95th percentile of the frequency distribution of each dataset, were removed. Peaks of enrichments were calculated using BCP (Xing et al., 2012). GAM data was processed to obtain normalised contact maps, exactly as described in Beagrie et al. (2017), with the difference that reads were mapped to the mm10 genome assembly instead of mm9, for comparison purposes with other datasets. Raw Hi-C data from Dixon et al. (2012) was processed as described in 2.18.1 Hi-C data analysis.

2.7 Molecular biology methods

If not described differently all classical molecular biological experiments were performed according to the handbook "Molecular Cloning: A Laboratory Manual"(Sambrook and Russell, 2001).

2.8 Cell culture

Table 2. 3: ESC culture reagents.

Reagent	Composition / supplier
Knockout serum replacement (KSR)	10828028, Invitrogen
Fetal bovine serum (FBS)	Heat-inactivated FBS, 30 min at 56°C (Gibco)
Trypsin-EDTA (0.05%), phenol red	25300054, Thermo Fisher
Accutase	A11105-01, Life Tech
Feeder medium	90 % DMEM (11995-065, Gibco), 10 % FBS
Freezing medium	90 % FBS, 10 % DMSO
F123 ESC medium	DMEM (11995-065, Gibco), supplemented with 15 % KSR, 1x Glutamax (35050, Gibco), 10 mM non-essential amino acids (11140-050, Gibco), 50µM beta- mercaptoethanol (31350010, Gibco), 1000 U / ml LIF (GFM200, Cell Guidance Systems)
46C ESC medium	GMEM Bhk21 (11710035, Gibco), supplemented with 10 mM non-essential amino acids (11140-050, Gibco), 0.1 mM beta-mercaptoethanol (31350010, Gibco), 1 mM sodium pyruvate (Gibco), 10 % FBS, 2000 U/ml LIF
Gelatine	0.1 % gelatine in 1x PBS
F123 gelatine	ESGRO Complete Gelatine (SF008, Merck)
ESGRO medium	ESGRO Complete Plus clonal grade (SF001, Merck)

2.8.1 Thawing cells

Cells were thawed in a water bath at 37°C and then directly added to 10 ml of media. Cells were pelleted by centrifugation at 190xg for 5 min, resuspended in fresh media, and seeded.

2.8.2 Splitting cells

Cells were washed in 1x PBS and trypsinised in Trypsin-EDTA for 5 min at 37°C. Reaction was stopped with two volumes of media. Cells were resuspended and trypsin was removed by centrifugation for 3 min at 190xg. The cell pellet was resuspended in fresh media and seeded, typically with a split radio of 1:6. F123 mouse ES cells were not passaged with Trypsin-EDTA, but using Accutase for 3-10 min at 37°C to gently lift the cells from the cell culture dish.

2.8.3 Freezing cells

For cryopreservation, cells were trypsinised at 70 – 80 % confluency. The cell pellet was resuspended in freezing media and frozen in 1-ml aliquots in cryo-vials. Vials were frozen in a cryo-box with isopropanol at -80°C for 24 h and then transferred to liquid nitrogen for storage.

2.8.4 Culturing 46C mESCs

ESCs were cultured directly on gelatine-coated culture dishes in media with a high concentration of Leukemia Inhibitory Factor (LIF) to maintain pluripotency of the cells and avoid spontaneous differentiation. Cells were split onto new gelatine-coated dishes every 48 h and media was changed every 24 h. Typically after one week in culture, cells were plated for harvest. After ~ 3 – 8 h ESCs settled and the media was changed to ESGRO media. Cells were grown in ESGRO media for ~ 48 h before harvest at 70 – 80 % confluency.

2.8.5 Culturing F123 mESC line

F123 ESCs were cultured on a layer of feeder murine embryonic fibroblasts (MEFs), that have been mitotically inactivated (GSC-6201G, Global Stem). Feeder cells were grown at 37°C in feeder media, and used up to 10 days after seeding. One day before culturing ESCs, dishes were coated with 0.1% gelatine and inactivated feeder cells were plated with a density of ~1500 cells per mm^2. After feeders have settled (~ 4 – 12 h after plating) ESCs were seeded onto the feeder layer and grown at 37°C in F123 ESC media. Cells were split onto new feeder-coated dishes every 48 h, and media was changed every 24 h. Typically, after two passages, feeder cells were removed from the ESC culture by splitting cells onto an uncoated dish for 30 mins. MEFs settle fast, while the ESCs remain in suspension. The cell suspension was transferred to a new uncoated plate for another 30 min to increase the efficiency of feeder removal, and afterwards cells were seeded on gelatine-coated dishes. The feeder-removal was repeated after 48 hours, and ESCs were subsequently plated for harvest. As feeder-removal results in reduced levels of LIF in the culture, the LIF concentration in the media was doubled when the cells were in feeder-free culture conditions. Cells were harvested after ~ 48 h at 70 – 80 % confluency.

2.8.6 Culturing 16p11.2 mESC lines

ESCs were cultured directly on gelatine-coated plates in M15 medium, supplemented with 15% FBS and 2000 U/ml LIF. Typically, cells were grown for one week before plating them for harvest or for subsequent differentiation. Cells were split every 48 h and media was changed every 24 h. Cell harvest was performed at 70 – 80 % confluency. Cell culture work with the 16p cell lines was conducted by Julietta Ramirez and Dr. Marta Slimak Mastrubuoni from our laboratory.

2.8.7 Neuronal differentiation of 16p ESC lines

Cell were differentiated according to Ferrai et al. (2017). Briefly, the differentiation of ESCs to dopaminergic neurons starts with differentiating ESCs into epiblast stem cells (EpiSCs) by growing them for 4 weeks in N2B27 basal medium containing Activin and FGF2. The EpiSCs were kept in culture for one week, followed by differentiation towards midbrain-specific dopamine neurons. The differentiation protocol includes blockade of FGF signalling, followed by subsequent enhancement of FGF signalling and simulation with Shh. After 5 days of differentiation neuronal progenitor cells were harvested, and after 16 days of differentiation premature dopaminergic neurons were harvested. The differentiation was conducted by Julietta Ramirez and Marta Slimak Mastrobuoni.

2.8.8 Mycoplasma test

For detection of Mycoplasma in cultured cells the PCR Mycoplasma test kit was used according to manufacturer's instructions.

2.9 RNA isolation

Cell lysis was performed in TRIzol Reagent (Invitrogen, 15596026) at room temperature (RT), and lysate was frozen in liquid nitrogen and stored at -80°C upon processing. Samples were incubated at RT for 5 min and homogenized with 200 ml chloroform per 1 ml of TRIzol, by shaking for 15 s and incubating 3 min at RT. After centrifugation at 12,000xg for 15 min at 4°C, the upper aqueous phase containing the RNA was transferred to a new tube and RNA was precipitated by adding 500 µl HPLC-grade isopropanol, incubating for 10 min at RT, and centrifugation at 12,000xg for 10 min at 4°C. Supernatant was removed and RNA pellet is washed with 75 % ethanol, air-dried for 10 min, resuspended in RNase-free water, and incubated at 55°C for 10 min. DNA was removed by DNase treatment using Turbo DNase (AM2238, Thermo Fisher) according to the manufacturer's instructions. Purified RNA was stored at -80°C.

2.10 RNA-seq

Purified RNA was analysed on the Bioanalyzer using the Agilent RNA 6000 Nano Kit to ensure intact, non-degraded RNA presence. RNA-seq libraries were generated from 1 µg of clean RNA using the TruSeq Stranded total RNA library preparation kit according to the manufacturer's (Sample prep guide 15031048). Libraries were analysed on the Bioanalyzer using the Agilent High Sensitivity DNA Kit and DNA concentrations were measured with the Qubit Quant IT kit according to manufacturer's specifications to estimate the molarity with the following formula.

$$Molarity\ (nM)\ =\ \frac{concentration\ (ng/\mu l)}{660\ g/mol\ x\ average\ fragment\ size\ (bp)}\ x\ 10^{6}$$

Samples were pooled and sequenced on the NextSeq500, paired end 75 bp, with the NextSeq 500/550 High Output v2 kit (150 cycles), following the manufacturer's instructions for denaturing, dilution, and sequencing of the libraries.

2.11 Hi-C

2.11.1 Preparation of Hi-C cell pellets

Cells were grown on a 100-mm cell culture dish to 70-80 % confluency at the day of harvest. A technical replicate was grown for counting cell numbers. Typically, 25 million cells were used for one Hi-C library. Cells in adherent cell culture were fixed in 14 ml fresh culture media, supplemented with 2% formaldehyde for 10 min at RT, while gently rocking the plate every 2 min. Crosslinking was stopped by adding 781 µl of 2.5 M glycine, and incubation at RT for 5 min, followed by 15 min on ice. Cells were scraped off the plate with a cell scraper and transferred into a 15 ml conical tube. Cells were spun down with 872xg for 10 min at 4°C, media was removed and the cell pellet was frozen in liquid nitrogen for storage at -80°C.

2.11.2 Chromatin preparation

Cells were resuspended in 500 µl fresh lysis buffer (5 µl 1 M Tris-HCl, pH 8.0, 50 µl 0.1 M NaCl, 5 µl 20 % Igepal CA-630, 440 µl H_2O, 2.5x protease inhibitor cocktail), incubated on ice for 15 min, followed by two rounds of disruption of the cells with 20 strokes using a Dounce homogenizer (pestle B) and 1 min incubation. After a total of 30 min in lysis buffer, cell lysate is transferred to a 1.7 ml tube and centrifuged with 200xg for 5 min at 4°C. The supernatant was discarded and the pellet was washed with 1x cold NEB 2 restriction buffer, followed by resuspension in 250 µl 1x NEB 2. The lysate was divided into 5 aliquots and 312 µl 1x NEB2 was added to each aliquot. 38 µl 1% SDS were added, aliquots were mixed by pipetting and incubated at 65°C for 10 min, shaking. Tubes were put on ice and 44 µl 10% Triton X-100 were added to each tube. Aliquots were incubated for 1 h at 37°C, shaking. Afterwards, 10 µl aliquot (undigested control) were stored at -20°C and 15 µl HindIII HF (100 U/µl = 1500 U) was added to each tube. Aliquots were incubated at 37°C while shaking overnight.

To determine the digestion efficiency, a 15 µl aliquot (digested control) was taken and DNA was extracted from both control samples (digested, undigested). DNA extraction from control samples was performed by adding TE buffer to 95 µl, then adding 2 µl RNaseA (1mg/ml) and incubating at 37°C for 1h. 5µl proteinase k (10mg/ml) were added and samples were incubated at 65°C for 2 h, shaking at 700rpm. Then, 100 µl phenol-chloroform were added, and samples were vortexed for 30 s, spun down at 21130xg for 10 min (RT) and the supernatant was transferred to a new 1.7 ml tube. An aliquot was taken and run on 1 % agarose gel to visually evaluate the digestion efficiency. Then, 10 µl sodium acetate, pH 5.2 were added and samples were mixed by pipetting. 250 µl 100 % ethanol (cold) were added, samples were inverted for mixing and incubate at -80°C for 1 h, spun down at 4°C, 9391 g for 20 min, and washed with 500 ml 70 % EtOH. Samples were resuspended in 25 µl H_2O. DNA concentration were measured with Nano Drop and 200 ng of each sample was run on a 1% agarose gel. The extracted DNA was normalised and qPCR analysis was performed using a primer pair within a HindIII restriction fragment (no HindIII site), and a primer pair spanning a HindIII restriction fragment. For the PCR reaction, 5 µl of SYBR PCR master mix were combined with 1 µl of each PCR primer (10 µM), 3 µl H_2O and 1 µl template, and the reaction was run on a qPCR cycler for 15 min at 95˚C, followed by 45 cycles of 10 s at 95˚C and 1 min at 60˚C. The digestion efficiency was calculated from the delta Ct-values of the two primer pairs from the digested and the undigested control using the following formula.

$$Digestion\ efficiency\ (\%)\ =\ 100\ -\ \langle\frac{100}{2^{(\Delta Ct\ (digested\ control)\ -\ \Delta Ct\ (undigested\ control))}}\rangle,\ \text{with}$$

$$\Delta Ct\ = Ct_{HindII\ site}\ -\ Ct_{no\ HindII\ site}$$

If needed, another 1000 units (10 µl) restriction enzyme were added to each tube for continued digestion at 37°C, shaking, o.n. Afterwards, digestion efficiency was determined again. The digestion efficiency was always above 50 % before proceeding to the subsequent steps.

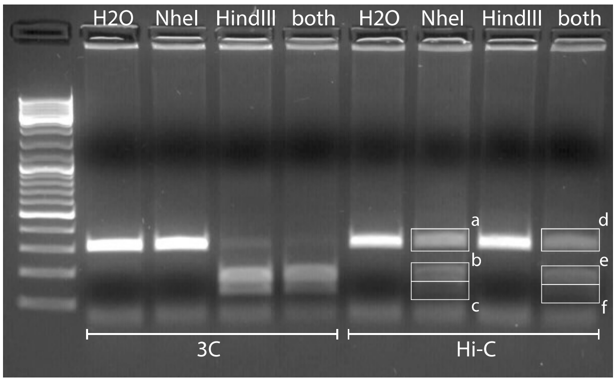
For biotin-labelling, samples were cooled down on ice. One aliquot was kept separately on ice as a control library to later determine the biotin-labelling efficiency. For biotin fill-in, 1.5 µl 10mM dATP, 1.5 µl 10mM dGTP, 1.5 µl 10mM dTTP, 37.5 µl 0.4mM biotin-14-dCTP (Invitrogen), and 10 µl 5U/µl Klenow (NEB) were added to each aliquot. Aliquots were incubated at 37°C for 45 min, with gentle mixing every 15 min, and placed on ice immediately afterwards. 86 µl 10 % SDS was added to all 5 tubes (including labelling control) and incubated for exactly 30 min at 65°C to inactivate the enzyme. Samples were placed on ice immediately and 575 ml lysate from each aliquot was transferred to a 15 ml conical tube for DNA ligation and 750µl 10x T4 DNA Ligase Reaction Buffer (500 mM Tris-HCl pH 7.5, 100 mM $MgCl_2$, 100 mM DTT), 750µl 10% Triton X-100, 10.5 µl 7.5% BSA, and 5345 µl H_2O were added. Samples were inverted 4x, the labelling control was incubated with 2µl T4 DNA Ligase (5 U/µl, Invitrogen) and the other aliquots with 20 µl T4 DNA Ligase at 16°C overnight (water bath, 4°C). 50 µl ATP (100 mM) were added to each tube and ligation was continued another 4 h.

2.11.3 DNA isolation and quality assessment

For reverse crosslinking 50 µl of 10 mg/ml proteinase K were added to each tube and samples were incubated over night at 65°C (water bath). Afterwards, another 50 µl proteinase k were added for 2 h of incubation at 65°C. Reactions were cooled down to room temperature and transferred into 50 ml conical tubes. 10 ml phenol-chloroform were added to each tube, samples were vortexed for 2 min, and then centrifuged for 10 min at 3000xg at RT. The aqueous phase (top) was transferred to a fresh 50 ml conical tube and the extraction was repeated using 8 ml chloroform. After centrifugation, the top phase was transferred to a 35 ml centrifuge bottle (SS34), the volume was brought to 10 ml with 1x TE buffer, 0.1 volumes (1 ml) of 3 M sodium acetate pH 5.2 was added, mix and samples were spun down briefly. 2.5 volumes (25 ml) of ice-cold 100 % ethanol were added and samples were mixed gently, and incubated at -80°C overnight. Samples were thawed and centrifuged at 4°C for 30 min at 20216xg, washed with 10 ml fresh 70 % ethanol 3x, with centrifugation at 4°C for 20 min at 20216xg. The supernatant was discarded, the DNA pellet was air-dried and resuspended in 450 µl 1x TE buffer and transfer to a 1.7 ml tube. Then, 500 ml phenol-chloroform were added, samples were vortexed for 30 s, and spun with 26915xg for 5 min at RT. Supernatant was transferred to a new tube and the extraction was repeated using 500 µl chloroform. 400 µl supernatant were transferred to a new tube and 0.1 volumes (40 µl) of 3M sodium acetate pH 5.2 were added, followed by 1 ml 100 % ethanol (cold), and incubation at -80°C for 2 h. Samples were spun down at 4°C with 26915xg for 20 min, supernatant was removed, and pellets were washed 4 times with 70 % EtOH. All EtOH was removed carefully and after drying DNA was resuspended in 25 µl 1x TE buffer with 1 µl RNase A (1 mg/ml). Samples were incubated for 15 min at 37°C, and all aliquots but the labelling control were pooled. DNA was quantified using the Quant-iT PicoGreen assay (Invitrogen). 8µl of a 1:10 dilution of each library was run on a 1 % agarose gel to visually assess the ligation efficiency. To test the formation of ligation products a PCR reaction was performed to amplify a ligation product formed by two directly adjacent restriction fragments. The resulting amplicon is digested with HindIII or NheI to assess the proportion of biotin-labelled library. For amplification a dilution series of the Hi-C library and the labelling control was made and amplified in a total of 5 reactions per sample by adding 5 µl 5x Hi-Fi buffer, 2 µl 50 mM $MgSO_4$, 2.5 µl 2 mM dNTPs, 0.125 µl 80 µM of oligonucleotides 3C-Gapdh_for and 3C-Gapdh_rev, 0.25 µl Phusion 2 U/µl, and 11 µl H_2O to 4 µl template and running the reaction in a PCR cycler for 30 s at 95˚C, followed by 35 cycles of 10 s at 98˚C, 30 s at 65˚C, and 30 s at 72˚C, followed by 1 x 7 min at 72˚C. 10 µl of each PCR product (300 bp) were run on a 1.2 % agarose gel to assess the amount of ligation product formed. The remaining reactions were pooled and digested with the following digestion reaction at 37°C for 1 h.

Reagent	H_2O	NheI	HindIII	Both
PCR Product	15 µl	15 µl	15 µl	15 µl
NEB2	5 µl	5 µl	5 µl	7.5 µl
H_2O	25 µl	25 µl	27.5 µl	45 µl
BSA (10mg/ml)	0	0.5 µl	0	0.75 µl
Enzyme	5 µl H2O	5 µl NheI	2.5 µl HindIII	5 µl NheI + 2.5 µl HindIII

The digested PCR products were run on a 2.0 % agarose gel and gel bands were quantified with ImageJ to determine the biotin-labelling efficiency with the following formula.



$$\text{Biotin labeling efficiency (\%)} = \frac{(b + c) / (a + b + c)}{(e + f) / (d + e + f)}$$

Figure 2. 1: Calculation of the biotin labelling efficiency of a Hi-C library. Agarose gel with the amplicon of the labelling control (3C) and the Hi-C library undigested (H2O), digested with NheI, HindII, and both enzymes. The efficiency is calculated by quantifying the intensities of the bands a – f.

2.11.4 Biotin removal, pull-down and library preparation

Up to 10 reactions, each with 5 µg of Hi-C library, were prepared for pull-down by removing biotinylated nucleotides from the ends of unligated restriction fragments. To 5 µg Hi-C library 1 µl 10 mg/ml BSA, 10 µl 10x NEB2, 1 µl 10 mM dATP, 1 µl 10 mM dGTP, 1.67 µl T4 DNA polymerase (3 U/µl) were added, and reaction was added up to 100 µl with H_2O and incubated for 4 h at 12°C. Then, 2 µl of 0.5M EDTA, pH 8.0, were added to each tube to stop the reaction, samples were pooled, and DNA was purified with a phenol-chloroform extraction, followed by DNA precipitation, as described above. DNA was resolved in 600 µl H_2O and divide in 3x 200 µl aliquots for sonication. DNA was sheared to a size of ˜300 to 500 bp using a Bioruptor Plus with the following setting. Settings: high, 30 s on / 30 s off, 15 cycles, at 4˚C. DNA was size selected for fragments between 300 and 500 bp with AmPure XP beads at RT according to the manufacturer's instructions using a 0.65x ratio for binding and removal of large fragments above 500 bp, followed by 0.85x for binding and selecting of fragments above 300 bp. DNA was eluted in 300 µl Qiagen elution buffer and quantified with the Qubit dsDNA HS assay. Samples with more than 2 µg DNA were used for subsequent pull-down.

For pull-down of ligation products, 150 µl Dynabeads MyOne Streptavidin C1 (Invitrogen) were washed with 400 µl Tween wash buffer (5 mM Tris, 0.5 mM EDTA, 1 M NaCl, 0.05 % Tween 20) for 3 min, then placed on a magnetic rack, supernatant was removed, and wash was repeated for a total of 3 times. Then beads were resuspended in 300 µl 2x binding buffer (10 mM Tris, 1 mM EDTA, 2 M NaCl), and the 300 µl Hi-C library were added. Sample was incubated on the beads at RT for 30 min rotating to allow all biotinylated DNA fragments to bind to the beads. Then, the beads were washed twice with 400 µl binding buffer, each time transferred to a new tube, and resuspended in 100 µl NEB ligase buffer. Then beads were resuspended in 10 µl 10x Ligase buffer (NEB), 4 µl 10 mM dNTPs, 5 µl T4 DNA polymerase (NEB, 3 U/µl), 5 µl T4 polynucleotide kinase (NEB, 10 U/µl), 1 µl large (Klenow) fragment (NEB, 5 U/µl), and 75 µl H_2O and incubated for 30 min at 20˚C. Beads were washed twice in 200 µl Tween Wash Buffer, and twice in 200 µl Qiagen elution buffer, before adding 5µl 10x NEB2, 10µl 1mM dATP, 3µl Klenow (exo-), and 32µl H2O (nuclease free) and incubating for 30 min at 37°C. Beads were washed twice in 200 µl Tween Wash Buffer, and twice in 200 µl Qiagen elution buffer, before adapter ligation. For adapter ligation sequencing adapters were prepared by annealing 15µl universal adapter P5 (15 µM), with 15 µl P7 indexing adapter (15 µM) in 70 µl H2O at 95°C for 5 min, followed by 70x decrease by 1°C, hold each for 1 min, and 1x 25°C for 30 min. Then, 10µl of 5x Ligase buffer (Invitrogen) were mixed with 5µl TruSeq adapters (15µM) and 32µl H_2O (nuclease-free), added to the beads with 3µl DNA ligase (Invitrogen) and incubated overnight at 16°C, shaking at 750 rpm. Beads were washed twice in 200 µl Tween Wash Buffer, and twice in 200 µl Qiagen elution buffer, and a test PCR was performed on the beads to determine the optimal number of cycles needed for amplification. The test PCR reaction was prepared by mixing 2.5 µl Hi-C library with 0.5 µl primer P5 (25 mM), 0.5 µl primer P7 (25 mM), 0.625 µl 10 mM dNTPs, 5 µl 5x Herculase buffer, 1 µl DMSO, 0.5 µl Herculase polymerase, and 14.375 µl H_2O, and running the reaction for 2 min at 98°C, followed by several cycles (tests included 5, 7, 10, 15 cycles) of 15 s at 98°C, 30 s at 62°C, 1 min at 72°C, and 1 cycles of 5 min at 72°C. PCR products were run on a 1.5% agarose gel to determine the number of cycles, at which a visible DNA smear appears, which were between 6-8 cycles for all prepared samples. All remaining material was amplified with the determined number of cycles, pooled afterwards, and purified twice with 1.8x Ampure XP beads according to manufacturer's instructions. Clean libraries were resuspended in Qiagen elution buffer and DNA was quantified with the Qubit dsDNA HS

Assay Kit and run on Bioanalyzer using the Agilent High Sensitivity DNA Kit to estimate the molarity with the following formula.

$$Molarity\ (nM) \ = \frac{concentration\ (ng/\mu l)}{660\ g/mol\ x\ average\ fragment\ size\ (bp)}\ x\ 10^6$$

Samples were sequenced on the NextSeq500, paired-end 75 bp, with the NextSeq 500/550 High Output v2 kit (150 cycles), following the manufacturer's instructions for denaturing, dilution, and sequencing of the libraries.

2.12 Sample preparation for cryosectioning

Ultrathin nuclear cryosections can be produced in the absence of resin-embedding, by the Tokuyasu method (Tokuyasu, 1973). Fixation of cells was performed as described previously (Branco and Pombo, 2006). Briefly, cells were grown on a 10mm cell culture dish to 70 % confluency, media was removed, and cells were fixed in 4% and 8% paraformaldehyde in 250 mM HEPES-NaOH (pH 7.6; 10 min and 2 h, respectively), gently scrapped, and softly pelleted and embedded in saturated 2.1 M sucrose in PBS and frozen in liquid nitrogen on copper sample holders.

2.13 Cryosectioning

Ultrathin cryosections were cut with a glass knife using an ultracryomicrotome (Leica Biosystems, EM UC7) at ~230 nm thickness, captured on sucrose-PBS drops and transferred either to autoclaved glass coverslips for immunofluorescence or to PEN membrane steel frame slides 4.0 µm (Leica Microsystems, 11600289) for laser microdissection.

2.14 Immunofluorescence

For immunolabelling, cryosections were washed in 1x PBS (3x 10 min each) and 20 mM glycine in PBS (30 min), permeabilised with 0.5 % Triton X-100 in PBS (v/v), for 10 min, and blocked for 1 h with PBS+ (1% BSA, 0.1% casein, 0.2% fish skin gelatin, in PBS, pH 7.6). All washes and antibody dilutions were done with PBS+. For imaging, primary antibodies were diluted 1:100, secondary antibodies 1:1000. For laser microdissection, the primary and secondary antibody dilutions were 1:50 and 1:500, respectively. Primary antibodies were incubated overnight at 4˚C, followed by a 1-h wash, and secondary antibodies were incubated for 1 h, followed by a 30-45 min wash. For imaging cryosections on glass, cryosections were washed 3x in PBS, and mounted in DAPI Vectashield. Coverslips were sealed with nail polish and imaged immediately after the nail polished had dried. For laser microdissection, cryosections were washed 2 times with PBS and 3 times with H_2O, then liquid was removed, the sample was air-dried for 10 min and immediately taken for laser microdissection (LMD).

2.15 Genome architecture mapping (GAM)

2.15.1 Staining of nuclear profiles (NPs)

Before collecting NPs under the laser microdissection microscope (LMD), cryosections were washed in sterile-filtered (0.2 µm syringe filter) 1x PBS (3 times, 5 min each), washed with sterile-filtered H_2O (3 times, 5 min each), and then taken to the LMD, or stained either with cresyl violet, or were immuno-labelled for using a pan histone antibody (see 2.14). Cresyl violet staining was performed with sterile-filtered 1 % (w/v) cresyl violet (Sigma-Aldrich**, C5042)** in H_2O for 10 min, followed by 2 washes with H_2O (30 s each). The same protocol was applied to different DNA and cellular dyes that have been tested in GAM, and that can be found in Table 3.2.

2.15.2 Collecting nuclear profiles

Individual NPs were isolated from the cryosection by laser microdissection using a laser microdissection microscope (Leica Microsystems, LMD7000). Unstained or cresyl-violet stained cells were identified under bright-field imaging and the laser was used to cut the PEN membrane surrounding each cell. Cut NPs were collected in a PCR Cap Strip filled with opaque adhesive material (Carl Zeiss Microscopy, 415190-9161-000). For each collection day, 1 or 2 caps were left empty and taken through the whole genome amplification (WGA) and sequencing process as a negative control for quality control purposes.

2.15.3 Whole genome amplification

Several different protocols were used to perform whole genome amplification (WGA). First, the WGA-4 GenomePlex® Single Cell Whole Genome Amplification Kit (Sigma Aldrich) was used to amplify DNA from NPs. For test purposes, also the REPLI-g Mini Kit (Qiagen), the MALBAC® Single Cell WGA Kit (Yikon Genomics), and Ampli-1 WGA kit (Menarini silicon biosystems) were applied to NPs, according to the manufacturer’s specifications. Most of the data produced during my PhD were produced using a WGA protocol developed by Dr.

Enric-Espel Mesferrer (University of Barcelona, Spain), Dr. Alexander Kukalev (our laboratory), and myself.

The WGA4 kit (Sigma) was used as previously described (Beagrie et al., 2017). Briefly, H_2O (13 µl) was added to each LMD caps containing a NP. Fragmentation master mix (8 µl proteinase K solution, 128 µl 10x single cell lysis and fragmentation buffer) was added to each lid (1.4 µl/lid), and 1 µl of human genomic DNA was added to a single lid without a NP to act as a positive control. The lids were pressed into a 96 well PCR plate and incubated upside down at 50 °C for 4 h. After incubation, the PCR plate was left to cool at room temperature for 5 min, before it was inverted and centrifuged at 800 xg for 3 min. The plate was heat-inactivated at 99°C for 4 min in a PCR machine and cooled on ice for 2 min. 2.9 µl 1x single cell library preparation buffer and 1.4 µl library stabilisation solution were added to each well and the plate was incubated at 95°C for 4 min, before cooling on ice for 2 min. 1.4 µl of library preparation enzyme was added to each reaction, then the plate was incubated on a PCR machine at 16 °C for 20 min, 24°C for 20 min, 37 °C for 20 min and finally 75 °C for 5 min. After WGA library preparation, the PCR plate was centrifuged at 800 xg for 3 min. 10x amplification master mix (10.8 µl), water (69.8 µl) and WGA DNA Polymerase (7.2 µl) were added to each well and the sample was PCR amplified using the program provided by the WGA4 kit supplier.

The WGA we developed in the laboratory is based on the original Multiple Annealing and Looping Based Amplification Cycles (MALBAC) protocol for WGA (Zong et al., 2012). In this technique, random hexamers bind to the genome in multiple annealing steps and amplify genomic DNA. The amplification products contain a common primer sequence, which was attached to the random hexamers, and are then further amplified with primers annealing to the common sequence. MALBAC has been used for genome but also transcriptome amplification in single cells, which had been published (Chapman et al., 2015). In this thesis, different modifications were applied to the protocol to achieve optimal DNA extraction and WGA performance for heavily cross-linked samples used for GAM (presented in supplementary Table 8.2) Here, I describe the final version of the WGA reaction. First, NPs were treated with lysis buffer and protease for cell lysis and protein removal. To each NP, 7 µl lysis buffer (30mM Tris Cl, 2mM EDTA, 5% Tween 20, 0.5% Triton X-100, 800 mM Guanidinium HCl pH 8.0), and 3 µl Qiagen protease (Qiagen, 3.175 U/ml) were added, and incubated overnight at 60˚C. Then, the protease was heat inactivated by incubating at 75˚C for 30 min. Samples

were placed on ice to cool down. Then, 0.8 µl 10 mM dNTPs, 0.8 µl 100 mM MgSO4, 2 µl GAT-7N (10µM), 4 µl 10x Thermo Pol buffer (NEB), 1.2 µl DeepVent polymerase (exo-) (NEB, 2U/µl), and 21.2 µl H_2O were added to each sample and DNA was amplified with the following PCR cycler conditions. First, 1x 3 min at 95˚C, followed by 11x 50 s at 20˚C, 50 s at 30˚C, 45 s at 40˚C, 45 s at 50˚C, 4 min at 65˚C, and 20 s at 95˚C, followed by 1x 5 min at 72˚C. Then, the adapter sequence from the GAT-7N primer, which is common to all amplified products from the first PCR reaction, is amplified with a common primer (GAT-COM) by adding 1.2 µl 10 mM dNTPs
0.4 µl 100 mM $MgSO_4$, 2.4 µl GAT-COM (10µM), 2 µl 10x Thermo Pol buffer (NEB), 0.6 µl DeepVent polymerase (exo-) (NEB, 2U/µl), and 13.4 µl H_2O. Amplification is done with 1x 3 min at 95˚C, followed by 24x 20 s at 95˚C, 30 s at 58˚C, 3 min at 72˚C, followed by 1x 3 min at 72˚C.

2.15.4 Preparation of sequencing libraries

For some samples collected in mouse 46C ESCs, WGA-amplified DNA was purified using either the Qiagen MinElute PCR Purification Kit or the MinElute 96 UF PCR Purification Kit and eluted in 50 µl of the manufacturer's elution buffer. For the majority of samples, including all F123 data, the WGA reaction was purified with Ampure XP beads (Ampure XP, Agencourt), or SPRI beads (DeAngelis et al., 1995). The concentration of each sample was determined by Quant-iT PicoGreen quantification. Sequencing libraries were then made using either the Illumina TruSeq Nano DNA HT kit or the Nextera XT library preparation kit, according to the manufacturer's recommendations. The Nextera XT library preparation was further successfully tested for the generation of high-quality sequencing libraries with only 1/5 of the volumes recommended by the manufacturer. Library concentrations were estimated with the Quant-iT PicoGreen assay and libraries were pooled together in batches of 96 or 192 libraries. The DNA concentration of the library pool was measured with the Qubit dsDNA HS assay and run on Bioanalyzer using the Agilent High Sensitivity DNA Kit to estimate the molarity with the following formula.

$$Molarity\ (nM)\ =\ \frac{concentration\ (ng/\mu l)}{660\ g/mol\ x\ average\ fragment\ size\ (bp)}\ x\ 10^6$$

As described in Beagrie et al. (2017), libraries produced with the TruSeq approach were sequenced in batches of maximum 96 samples in single end 100 bp rapid-run mode on two lanes of an Illumina HiSeq machine. Each library has 30 bp WGA adaptors at both ends, so the flow cell was not imaged for the first 30bp of each run (these are known as “dark cycles”). Libraries produced with the Nextera XT kit are randomly fragmented by Tn5 enzyme during the library prep, which means that sequencing adapters are inserted at random positions inside the WGA products. Thus, the WGA adapters are no longer present at the first position of the library insert DNA, which makes the HiSeq dark cycle run redundant. Nextera libraries were sequenced single end 75 bp with the NextSeq500/550 High Output v2 kit (75 cycles) on the NextSeq500 sequencer following the manufacturer’s instructions for denaturing, dilution, and sequencing of the libraries.

2.16 Microscopy

Images were acquired on a confocal laser-scanning microscope (Leica TCS SP8; 63x oil objective, NA 1.4), using pinhole equivalent to 1 Airy disk. Images from different channels were collected sequentially to prevent fluorescence bleed-through. Raw images (TIFF files) were merged in ImageJ and contrast stretched without thresholding. Image acquisition was done randomly based on the DAPI staining, to avoid bias.

2.17 Calculating the average nuclear radius from cryosections

Cryosections of F123 ESCs were prepared, stained with DAPI and imaged on a confocal laser-scanning microscope as described above. The radius of a nucleus can be estimated from the average of radii of nuclear slices, collected from random positions in a population of nuclear sections. The average radii of nuclear slices were calculated from 2 technical replicates (2 cryosections) for F123 mESCs by measuring the area of randomly chosen nuclear slices (Figure 2.2a-b) and using Area = (π * R^2). The incremental average of all measured slice radii per cryosection is plotted to determine whether the collection of measurements was saturated (Figure 2.2c). From the average slice radius R = 3.64 ± 0.30 µm we estimated a nuclear radius R_N = 4.63 µm, using $R_N = (4/\pi) * R$.

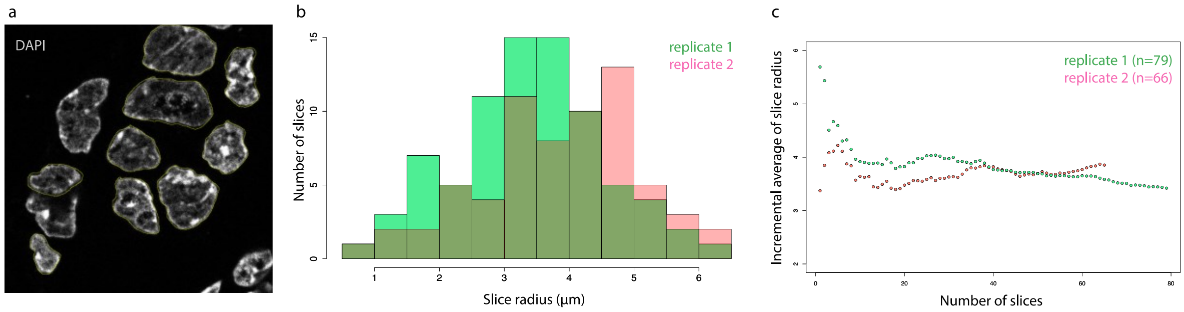


Figure 2. 2: Nuclear radius of F123 mESCs.
(a) Example cryosection stained with DAPI used for measuring nuclear slice radii. (b) Histogram showing the distribution of slice radii collected from 2 technical replicates. (c) Incremental average of slice radii for both replicates.

2.18 RNA-seq data analysis

For the 16p11.2 cell lines, sequencing reads were mapped to mm9 with STAR (Dobin et al., 2013), transcripts were quantified with RSEM (Li and Dewey, 2011), and analysis of count data was performed with DEseq2 (Love et al., 2014). To obtain differentially expressed genes between experiments, count values of genes were compared that had a value of transcripts per million (TPM) > 1 in at least one of the biological replicates. Genes were considered as differentially expressed between two experiments, when their adjusted p-values (Benjamini and Hochberg) were below 0.05. Analysis was performed by Dominik Szabo, with the guidance of Dr. Christoph Thieme from our laboratory.

RNAseq data from F123 was processed for standard and allele-specific gene expression analysis. The quality of the paired-end RNA sequencing reads was verified using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). No reads needed to be trimmed or removed due to quality concerns. The paired-end reads derived from RNA sequencing were mapped to the most recent mouse reference genome assembly mm10 (GRCm38.p6) using STAR (version 2.7.2c) (Dobin et al., 2013) under consideration of the current mm10 annotation (downloaded from ensemble: ftp://ftp.ensembl.org/pub/current_gtf/mus_musculus /Mus_musculus.GRCm38.98.gtf.gz) and available information of genomic variants in the mm10 F123 genome (described in 2.20.8). Following recommendations about best practices for data processing in allelic expression analysis (Castel et al., 2015), duplicate reads were removed from the data using Picard MarkDuplicates (version 2.21.1: https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.3.0/picard_sam_markduplicates_MarkDuplicates.php). Default options were used with the exception of REMOVE_DUPLICATES = TRUE. To quantify the overall expression of genes, mapped reads overlapping exons were assigned to the respective genes and summarized as gene specific count values using HTSeq-count (Anders et al., 2015). The use of HTSeq-counts to generate gene level read count values is recommended by the gold standard tool used for differential gene expression analysis DESeq2 (Love et al., 2014). Options were set to count reads overlapping exons of genes, accounting for the paired end nature of reads, only considering primary alignments and the default minimal alignment quality of 10. The same annotation file was used as described before in the read mapping step. Subsequently TPM values were calculated normalising count values for gene length and library size.

To differentiate between the expression of genes located on the two parental alleles, reads overlapping heterozygous genomic variants were counted in an allele-specific manner. Reads overlapping those heterozygous variants located within exons of genes were counted using GATK ASEReadCounter (The Genome Analysis Toolkit (GATK) version 4.1.3.0: https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.3.0/org_broadinstitute_hellbender_tools_walkers_rnaseq_ASEReadCounter.php). Subsequently only genomic variants within regions of high mappability and with a minimum total coverage of 20 reads were considered to reduce the risk of introduced biases. In case multiple genomic variants were present within the same gene, the counts were aggregated over the gene in an allele specific manner using the available haplotype information described above in the read mapping step. Aggregated counts were tested for significant allele specific expression differences (binomialtest vs 0.5) and the false discovery rate was controlled for by correcting resulting p-values for multiple testing using the Benjamini & Hochberg method. Subsequently ASE ratio $(\frac{readcounts\ supporting\ CAST\ allele}{total\ readcounts})$ and Log2foldchange $log2\left(\frac{readcounts\ supporting\ CAST\ allele}{readcounts\ supporting\ J129\ allele}\right)$ were calculated. Analysis was performed by Julia Markowski (Roland Schwarz laboratory, Max Delbrück Centrum, Berlin).

2.19 Hi-C data analysis

Hi-C data mapping, normalisation, and quality controls were performed by myself. Further Hi-C analyses were performed by Dr. Ehsan Irani and Dr. Christoph Thieme from our laboratory.

2.19.1 Generation of Hi-C contact maps

Published mESC Hi-C data from Dixon et al. (2012) was mapped to the mm9 reference genome and corrected using the iterative correction pipeline (Imakaev et al., 2012), as it was done in the original publication, by Markus Schueler from our laboratory. All newly produced Hi-C data was analysed by myself, and mapped using the TADbit pipeline from Serra et al. (2017). Reads were filtered for unligated fragments (dangling ends), self-ligation, random DNA breaks, and PCR duplicates. Raw Hi-C contact frequencies were normalised using the iterative correction pipeline (Imakaev et al., 2012) for one iteration.

2.19.2 Correlation analysis of Hi-C datasets

The pairwise similarity of intrachromosomal Hi-C matrices was assessed at 50kb resolution using HiCRep (v. 0.99.6) (Yang et al., 2017). For data correlations, the stratum-adjusted correlation coefficient (scc) for all pairs of chromosomes was calculated by using a maximum distance threshold of 20 Mb and an estimated optimal smoothing of zero. Results values were evaluated separately as well as aggregated mean values from all chromosomes.

2.19.3 Viewpoint-based Hi-C analysis

Z-score normalised interaction matrices were generated by normalising Hi-C matrices (after log transformation) for genomic distance.

$$z_{i,k+i}^{\log} = \frac{H_{i,\,k+i}^{\log} - \left\langle H_{j,\,k+j}^{\log} \right\rangle_j}{\operatorname{std}\left(H_{j,\,k+j}^{\log}\right)}$$

From the z-scores, normalised contact frequencies were filtered for a viewpoint-based analysis, by selecting only contacts of one genomic region with the remaining genome.

2.19.4 Differential Hi-C contact analysis

In order to study differential Hi-C contacts, different datasets were first normalised using z-score normalisation: the contact frequency between all loci with the same genomic distance was normalised to have the zero average and the standard deviation of one. Then gained and lost contacts between different datasets were identified after subtracting the generated z-score matrices as respectively positive and negative values.

2.19.5 Defining A/B compartments in Hi-C data

A/B compartments were identified at previously described (Liebermann-Aiden et al., 2009). Briefly, for each chromosome the observed over expected matrix was generated,

$$H'_{i,k+i} = \frac{H_{i,k+i}}{\langle H_{j,k+j} \rangle_j}$$

Then, a correlation matrix, $C_{i,j} = \mathrm{corr}(H'_i, H'_j)$ was generated, where the Pearson correlation between loci i and j is given by $C_{i,j}$. Principal component analysis (PCA) was performed on the correlation matrix C. Usually the first eigenvector stores the information about compartments. To indicate which values corresponds to A or B compartments, the first eigenvector was correlated with the GC content, knowing that A correlates with more GC rich regions.

2.19.6 Analysis of topologically associating domains

Insulation scores for each TAD boundary were calculated as previously described (Crane et al., 2015). Normalised contact frequencies were measured in a window box 10 times larger than the resolution of the dataset, moved at an offset of two windows from the diagonal of the matrix. This measure quantifies the contacts that span a particular genomic region. At TAD boundaries this value is lower than within TADs, resulting in a local drop of the insulation score of the boundary. This local drop was called as a boundary.

2.20 GAM data analysis

If not stated otherwise, GAM data analysis was performed by Dr. Alexander Kukalev and Dr. Ibai Irastorza Azcarate from our laboratory.

2.20.1 Mapping of nuclear profiles

Nuclear profiles from 46C mESCs were mapped to the mm10 reference genome. Nuclear profiles from F123 mESCs were mapped to an N-masked mm10 genome assembly, where all SNPs in the F123 genome were substituted with N, to allow unbiased mapping of maternal and paternal reads. Mapping was done with bowtie2, as previously described (Beagrie and Schueler, 2017).

2.20.2 Calling positive windows

For NPs collected with the Sigma WGA4 kit, which were processed with the TruSeq library preparation kit (Illumina), positive windows were identified using a curve fitting approach, as previously described (Beagrie et al., 2017). Briefly, the genome was split into equal-sized windows, and the number of reads was calculated from each nuclear profile overlapping each genomic bin. Two distributions were fitted to the histogram of the number of reads per window in each NP. A negative binomial distribution represents noise, and was used to determine a threshold number of reads, where the probability of observing more than this number of reads mapping to a single genomic window by chance was less than 0.001. This threshold was independently determined for each NP, and windows were called positive if the number of reads was greater than the determined threshold. To obtain a robust estimate of the noise, we fit a log-normal distribution representing signal simultaneously with the negative binomial, however the parameters of the log-normal distribution are not used in determining the threshold.

For all other NPs we used a nucleotide coverage-based approach to identify positive windows. Similar to the old window calling procedure, the genome was first split into equal size bins and the number of reads in each bin and number of nucleotides sequenced in each bin were calculated for each NP sample with bedtools (Quinlan and Hall, 2010). Next, we calculated lowest percentile of nucleotide coverage distribution that only allow positive windows with 3 reads and more for each sample in the dataset, and use the mean value of all percentiles as optimal percentile for the entire dataset. To obtain individual threshold for each NP in nucleotides, we then apply the optimal percentile to individual nucleotide coverage distributions for each NP. Windows were called positive if the number of nucleotides sequenced in each bin was greater than the determined individual threshold.

2.20.3 Checking for cross-contamination between samples

To compare the genomic content of nuclear profiles on each plate and exclude potentially cross-contaminated samples we use Jaccard similarity index. Briefly, we represented genomic content of each NP as a vector of zeros and ones, corresponding to the negative and positive genomic windows aligned according to the linear genome. We calculated Jaccard similarity index between all vectors corresponding to the samples that were processed together on the same 96 well plate. Samples with a Jaccard similarity index > 0.4 were excluded from the data analysis as potentially contaminated. The pipeline was developed by Thomas M. Sparks, our laboratory.

2.20.4 Quality metrics of NPs

Low-quality samples due for example to failed capture of NPs at the time or low DNA extraction were removed from the final set of GAM libraries and all subsequent analyses. A number of quality metrics were measured for each sequencing libraries produced from single NP, 3xNPs or water controls. First, the percentage of mapped reads and percentage of PCR duplicate reads were measured. For assessing contamination of nuclear profiles with DNA from other species, 100,000 randomly selected reads from each sample were mapped against common species in the laboratory (human, *E. coli*, yeast) using Fastq-screen (bioinformatics.babraham.ac.uk/projects/fastq_screen).

NPs are expected to cover a proportion of the whole genome, organized in consecutive stretches of genomic DNA being present or absent from the NP, reflecting chromatin looping in and out of a thin nuclear slice. Depending on the radial position of the slice relative to the whole nucleus, varying numbers of chromosomes are expected to be present in a single NP. Therefore, for every NP, we assessed the total number of positive windows as a measure of genome coverage, the number of positive windows immediately adjacent to another, reflecting the proportion of orphaned windows in the genome, and the number of positive chromosomes for each sample. In order to effectively count the chromosomes, present in each NP, an algorithm was applied that checks for stretches of positive windows in the genome, as a proxy for presence or absence of a chromosome. In detail, we checked the number of 50 kb positive windows in all Mb regions (20 positive windows in each Mb) and calculated the 4th quartile for all the chromosomes. If a certain chromosome's average number of 50 kb positive windows in a Mb is the same or higher than the 4th quartile of all the chromosomes, this specific chromosome is called as present in the NP. Finally, we used these metrics, together with the number of uniquely mapped reads to evaluate NP quality and decide on quality thresholds for the GAM datasets.

2.20.5 Determining resolution of pairwise co-segregation matrices

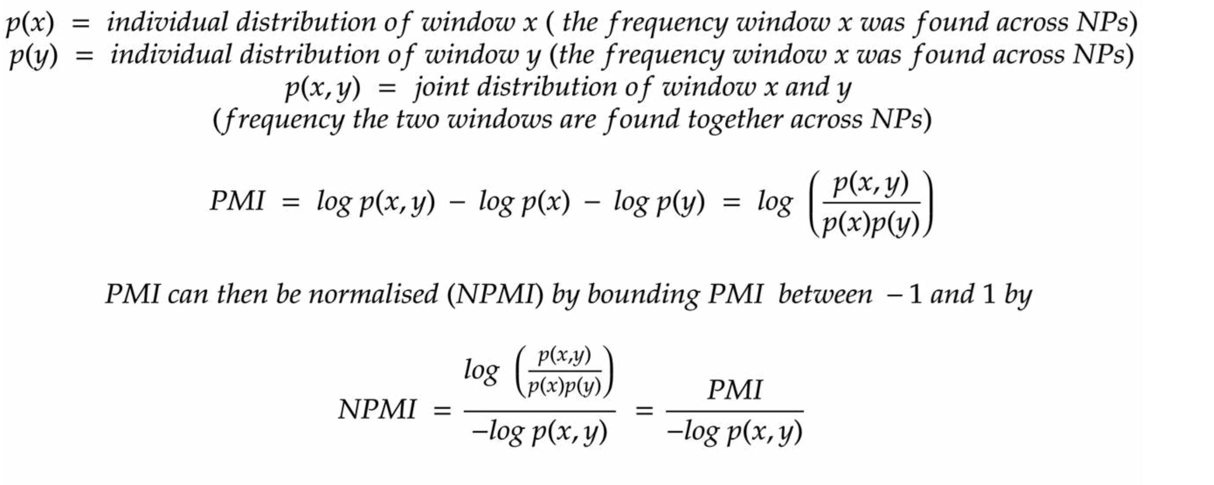
Each NP contains a very limited amount of genomic material. For good quality genome sampling we expect each pair of genomic loci to be found together in one of the NPs at-least once. To determine highest resolution that can be used for pairwise chromatin contact matrices, we conducted a data erosion by removing NPs from the dataset, and asked how many NPs are sufficient to sample entire genome at multiple resolutions. This analysis was performed for different genomic resolutions (10 to 100 kb) to estimate the highest resolution that was deemed robust to study genome organisation. The highest resolution where we observed good quality of genome sampling was used in the further analysis.

2.20.6 DNA detection efficiency

In order to assess the quality of genome sampling, the detection efficiency in the F123 GAM dataset was calculated according to Beagrie et al., 2017. The detection efficiency is 1, when the probability to find a genomic region is the same as the actual detection frequency of that region. The efficiency is less than 1, when the probability exceeds the detection frequency. For simplicity, the same probability was assumed for all genomic loci. The probability of a single locus to be detected in a nuclear profile was calculated based on the following criteria; the genomic size of the locus (bin size in kb), the size of the nucleus (nuclear radius estimation in 2.17), the thickness of the nuclear slice (0.22 µm), and the total genome size (mm10 genome length 5.46 Gb). This analysis was performed by Luca Fiorillo under the supervision of Mario Nicodemi (University of Naples, Italy).

2.20.7 Normalisation

The use of different normalisations of the GAM was explored by Thomas M. Sparks and Christoph Thieme (our laboratory). Pointwise mutual information (PMI) is a measure of association used in information theory and statistics. It is frequently used in linguistics to infer the co-association of words between documents (Watford et al., 2018). The PMI of genomic windows x and y describes the difference between the probability of both windows being found in the same NP (i.e. their joint distribution) and their individual distributions across all NPs. PMI assumes that finding one window is independent of finding the second window.



$p(x)$ = *individual distribution of window x (the frequency window x was found across NPs)*

$p(y)$ = *individual distribution of window y (the frequency window x was found across NPs)*

$p(x,y)$ = *joint distribution of window x and y*
(frequency the two windows are found together across NPs)

$$PMI = log\, p(x,y) - log\, p(x) - log\, p(y) = log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

PMI can then be normalised (NPMI) by bounding PMI between −1 and 1 by

$$NPMI = \frac{log\left(\frac{p(x,y)}{p(x)p(y)}\right)}{-log\, p(x,y)} = \frac{PMI}{-log\, p(x,y)}$$

NPMI normalisation was implemented for GAM by Thomas M. Sparks from our laboratory.

2.20.8 Phasing of nuclear profiles

The high SNP density of the F123 genome was used to phase the reads from NPs to their haplotypes. For generating genotype calls for the hybrid F123 (CAST×S129) cells, we downloaded parental genome sequencing data from publicly available databases. For *Mus musculus castaneus*, we downloaded the genome sequence from the European Nucleotide

Archive (accession number ERP000042). S129/SvJae genome sequencing data was downloaded from the Sequence Read Archive (accession number SRX037820). We performed reads trimming using Cutadapt (https://cutadapt.readthedocs.io/en/stable/) and mapped the reads to mm10 genome assembly using Burrows-Wheeler Aligner (http://bio-bwa.sourceforge.net/). SNP location and sequence were identified using bcftools (http://samtools.github.io/bcftools/bcftools.html). SNPs that have less than 5 reads per SNP and quality below 30 were excluded from the analysis. High quality SNPs were masked with N in the reference genome assembly and NPs were mapped to the N-masked genome using bowtie2. The reads mapped to N-masked genome were checked for the presence or absence of a SNP, and sorted to the corresponding haplotype using SNPsplit (Krueger and Andrews, 2016). The phased reads we used to assign positive windows to their parental alleles, as the majority of them contained phased reads from either the maternal or the paternal allele. A minimum coverage of 75 nucleotides (corresponding to 2 phased reads) was determined to assign a positive window to one haplotype. Positive windows with more than 2 reads from both alleles, were assigned to both haplotypes. Phasing of nuclear profiles was developed by A. Kukalev.

2.20.9 Defining A and B compartments

We calculated A and B compartments for GAM as described above (2.19.5), with the only difference that we did not always extract the principal component 1, but the first component that explain the most variance.

2.20.10 Detection of differential contacts

To identify differential contact between the alleles, allele-specific contact matrices were normalised, normalised contact intensities were subtracted, and, ultimately, contacts seen strong in both or at significantly different levels by either of the two alleles were extracted. For the analysis of allele-specific contacts in F123 mESCs we subjected NPMI normalised contact matrices at 50 kb resolution for all autosomal chromosomes for pairwise comparisons. In order to avoid amplification of spurious contacts from noise and zero inflation, we limited our analysis to 4 Mb genomic distance. Next, we used GAM window detection frequency to flag potentially oversampled or undersampled genomic regions in the GAM data (applied range > 0.01; < 0.07) and excluded those regions from the comparisons. To adjust for intrinsic differences of value distributions a normalisation step is considered essential. From a number of tested normalisation approaches, we identified z-scores

$$z = x_d - \langle x \rangle \frac{\{\backslash bar\{x_d\}\}}{S_d}$$

for values of same bin-distance (d) of a chromosome to be best suited for the data. We confirmed that contact frequencies of each normalised chromosomal matrix can be parameterized by a Normal distribution with very good fit and assigned p-values for each contact accordingly. Next, we extracted all contacts with p-value<0.05 or p-value>0.95 as maternal-specific or paternal-specific contacts, respectively. In contrast to these contacts which are either strong on the maternal or the paternal allele, we defined a set of contacts which are observed by both methods at similarly high value intensities. We defined these strong and common contacts to be the top 10% of contacts with a z-score delta between -1 and 1 ranked by the minimum of z-score normalised scores in both datasets.

2.21 TF enrichment analysis

For finding enriched TF motifs in F123 ATAC-seq regions Regulatory Genomics Toolbox Motif (www.regulatory-genomics.org/rgt) matching was used with the motif database HOCOMOCO11 MOUSE MOTIFS (www.regulatory-genomics.org/motif-anaysis). Identified motifs were filtered for expressed TFs in F123 mESCs using the TPM-values from the RNA-seq data. Only motifs of expressed TFs (TPM>1) were used in the analysis. Then, ATAC-seq peaks were mapped to genomic windows (50 kb) with common or allele-specific contacts and the coverage of each motif was calculated in each dataset (considering only regions identified by ATAC-seq). The 32 motifs with the highest coverage (percentage of windows containing the motif) were selected for subsequent motif pair analysis. Motif pairs were identified by their occurrence in contacting windows, and ranked based on their coverage (percentage of contacting windows containing a motif pair). TF enrichment analyses in differential contacting pairs of windows was developed and performed by Yingnan Zhang, Catherine Baugner, and Dr. Lonnie Welch (Ohio University, Athens, OH, USA)

2.22 Gene Ontology enrichment

Gene Ontology (GO) enrichment analysis of genes with allele-specific expression was performed using GO-Elite version 1.2.5 (Gladstone Institutes; http://genmapp.org/go_elite). Selected results (to decrease term redundancy) are reported. Default parameters were used as filters: z-score threshold > 1.96, permutation-derived p-value < 0.05, number of genes changed > 2. Over-representation analysis was performed with the "permute p-value" option, 2000 permutations. UCSC Known Gene IDs were converted into the correspondent Ensembl Gene IDs (using the UCSC KnownToEnsembl table, downloaded from the UCSC Table browser http://genome.ucsc.edu/cgi-bin/hgTables) before performing the GO enrichment analyses. The group of genes used as background was chosen based on expression (TPM>1) and the presence of SNPs (only phased genes with at least 1 SNP were included). Analysis was performed together with A. Kukalev.

3. GAM technology development

Notes to author contributions:

The experimental work presented in this chapter was performed by myself, in collaboration with Dr. Alexander Kukalev, Prof. Dr. Enric Espel-Mesferrer (University of Barcelona, Spain) and Robert A. Beagrie. A. Kukalev often performed the WGA or the library production. R.A. Beagrie advised and helped with collecting nuclear profiles for 46C ESC work, and further taught me how to perform GAM. E. Espel-Mesferrer developed the in-house WGA protocol in collaboration with A. Kukalev using samples that I collected and processed for library preparation. Further, Gesa Loof, Izabela Harabula, Leonid Serebreni, and Dr. Warren Winick-Ng helped generating samples for some of the tested variables. I. Harabula performed the experiment presented in Figure 3.25. The majority of all computational analysis presented here was performed by A. Kukalev, with contribution of R.A. Beagrie, and Dr. Ibai Irastorza Azcarate. The analysis presented in Figure 3.3 was performed by R. Beagrie, Dr. Antonio Scialdone (Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Germany), and Carlo Annunziatella (Laboratory of M. Nicodemi, University of Naples, Italy). The analysis presented in Figure 3.19 was done by Luca Fiorillo (laboratory of M. Nicodemi, University of Naples, Italy). Data normalisation and testing for bias (Figure 3.26) was done by Thomas M. Sparks. If not noted otherwise, all mentioned authors are from the laboratory of Prof. Ana Pombo, Max Delbrück Center, Berlin.

3.1 Objective and research motivation

Our laboratory developed Genome Architecture Mapping (GAM), a technique to map chromatin contacts via accessing co-segregation frequencies of genomic regions in nuclear slices. The main objective of my PhD was to apply GAM to mouse ESCs, to obtain a large dataset with high statistical power, that would allow a deep exploration of chromatin topology with high genomic resolution and with allelic information. To achieve this long-term goal, I started by optimising the GAM protocol to be able to produce samples in high-throughput, suitable for the collection of large datasets. The optimisations had several major goals that would help overcome several of the bottlenecks of the original experimental design used in Beagrie et al. (2017). The bigger challenge was first and utmost the efficiency of DNA extraction from crosslinked nuclear slices using whole genome amplification (WGA). In 2015, when I joined the laboratory, GAM was performed using the WGA4 kit from Sigma, with successful DNA extraction, as shown in Beagrie et al. (2017). The challenge at the time was to keep the efficiency of the experiment, after moving the laboratory of A. Pombo from London to Berlin. In 2017 the Sigma WGA4 kit no longer yielded effective extraction of DNA from GAM nuclear slices, so that unplanned efforts to test other WGA kit providers and to devise an in-house WGA approach became necessary. Several implementations were included in the protocol to improve the efficiency of DNA extraction, ultimately leading to the development of a WGA protocol independent from commercially available kits. The original GAM protocol was complex, time-consuming and expensive. This chapter presents an overview of experimental optimisations and their outcome, followed by explorations of different quality metrics of a GAM dataset collected with the optimised GAM pipeline, and comparisons to the previously produced dataset published in Beagrie et al. (2017).

3.2 Reproducibility of GAM

To test the reproducibility of the GAM technique, I first applied the original GAM protocol (Beagrie et al., 2017) on the same mESC line (46C; Figure 3.1). Briefly, nuclear profiles were isolated from 230 nm thin cryosections using a Leica laser microdissection microscope. DNA was extracted from single nuclear profiles and amplified by whole genome amplification using the WGA4 kit from Sigma. The extracted DNA was purified with column purification and prepared for Next Generation Sequencing using the TruSeq DNA library preparation kit from Illumina. Sequencing libraries were pooled and sequenced on the HiSeq2000 system. With the standard GAM protocol, I was able to extract DNA equally well as in the GAM dataset that was previously produced using a Zeiss laser microdissection microscope in A. Pombo’s lab in London (Beagrie et al., 2017), showing that GAM can be used independent of the user, laser microdissection instrument, or laboratory installations (Figure 3.1).

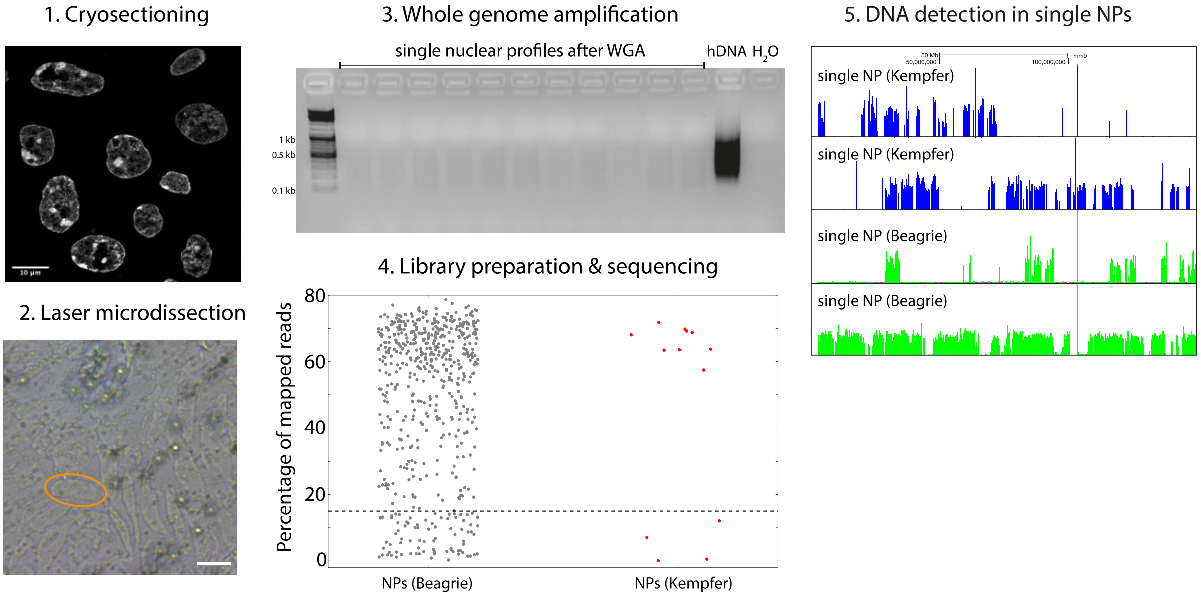


Figure 3. 1: Reproducibility of the GAM pipeline.
(**1-4**) The GAM protocol comprises cryosectioning (1), laser microdissection of single nuclear profiles (NPs) from the unstained cryosection (2), DNA extraction by whole genome amplification (WGA; 3), library preparation and sequencing (4). Applying the original GAM protocol (Beagrie et al., 2017) to 46C mouse ESCs reproduces quality metrics of single nuclear profiles. **(3)** After WGA, DNA from NPs is visible as faint DNA smear on a 1.2 % agarose gel, compared to WGA material from 5 ng of human DNA, and a clean water control. DNA ladder: Gene Ruler 1kb plus. **(4)** After sequencing, the percentage of reads mapping to the mouse genome indicates the success of DNA extraction from a NP. Typically, a NP contains consecutive stretches of the genome, interrupted by genomic regions that are not present, which corresponds to DNA folding in and out of a thin cryosection, here shown by typical sequencing tracks of NPs mapping to the mouse chromosome 6 (mm9).

3.3 Optimising the GAM protocol for high-throughput sample production

To reduce the hands-on time and the costs of production per nuclear profile, while retaining high efficiency of DNA extraction, several variants of different steps of the GAM protocol were tested. The sum of the modifications implemented led to the development of a whole new GAM pipeline (Figure 3.2), which drastically reduces time and cost per nuclear profile. Each modification was tested in side-by-side comparisons with the variant of the step of the previously used protocol. For steps before and during WGA, when the DNA sample cannot be divided in two, the quality of each method variation was tested against the quality metrics of sequencing libraries of nuclear profiles produced with the previous version of the protocol. For all protocol changes in library preparation and sequencing, tests were performed on the same DNA input material after WGA.

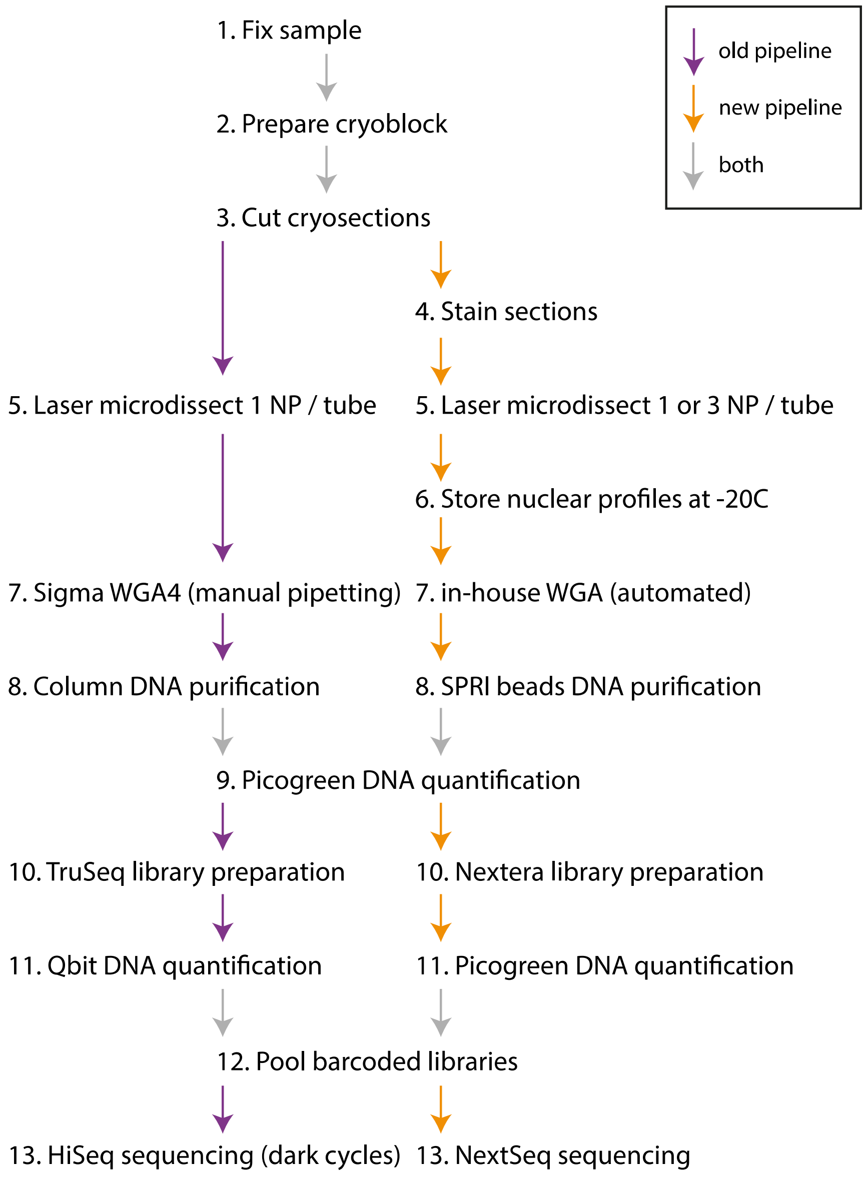

1. Fix sample
2. Prepare cryoblock
3. Cut cryosections
4. Stain sections
5. Laser microdissect 1 NP / tube
5. Laser microdissect 1 or 3 NP / tube
6. Store nuclear profiles at -20C
7. Sigma WGA4 (manual pipetting)
7. in-house WGA (automated)
8. Column DNA purification
8. SPRI beads DNA purification
9. Picogreen DNA quantification
10. TruSeq library preparation
10. Nextera library preparation
11. Qbit DNA quantification
11. Picogreen DNA quantification
12. Pool barcoded libraries
13. HiSeq sequencing (dark cycles)
13. NextSeq sequencing
old pipeline
new pipeline
both

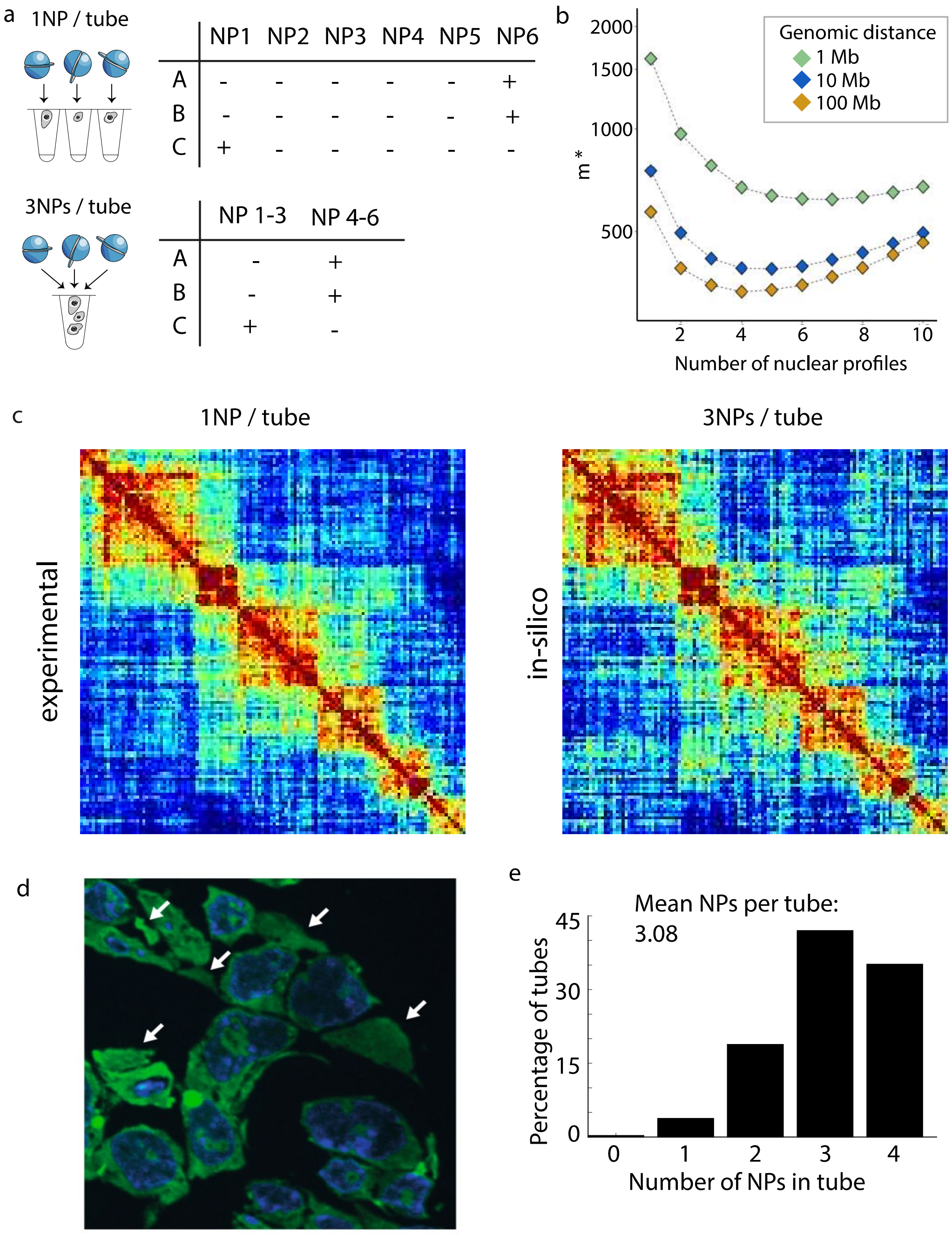
Figure 3. 2: The new GAM pipeline was optimised to reduce experimental time and costs.
Protocol steps of the first GAM pipeline (purple arrows; Beagrie et al. 2017), and the new pipeline (orange arrows). Common steps are indicated with grey arrows.

3.3.1 Multiplex GAM

One formulation of GAM, developed in collaboration with the laboratory of Mario Nicodemi, is multiplex GAM (Beagrie et al. in preparation), where instead of having only one nuclear profile per GAM sample (or tube), 2-6 nuclear profiles are extracted jointly. This is possible because each single nuclear profile contains ~8 % of the diploid genome, and each locus is detected in only ~1 out of 10 nuclear profiles. This implementation can reduce the number of tubes that are required to collect a GAM dataset und thereby reduce the workload and costs of the experiment.

To explore the possibility of collecting multiplexed NPs, a statistical model, termed SLICE (previously designed for 1NP GAM data (Beagrie et al., 2017)) was further developed to estimate a theoretically optimal number of tubes required to obtain interactions between loci separated by different genomic distances (analysis done by Carlo Annunziatella, laboratory of M. Nicodemi, University of Naples; and Antonio Scialdone, Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Germany). When combining NPs within one tube, the minimal number of samples (m* in Fig. 3.3b) required to map contacts at a given resolution (here we chose 30 kb) goes down with increasing number of NPs per tube, leading to an optimum of 3 - 4 NPs per sample (Figure 3.3b). When combining more than 4 NPs in a single sample, the power to resolve contacts decreases, especially at larger genomic distances, and more samples start to be required for the generation of the dataset. We decided to explore the use of 3NPs per sample for collecting multiplex GAM data (Figure 3.3a). This collection mode was first tested in silico, where we generated a 3-NP dataset out of 480 single NPs by combining the sequencing information from 3 single NPs into one file, and then computed co-segregation matrices from 160 in-silico 3-NPs (analyses performed by Robert Beagrie; Figure 3.3c). The datasets show high similarity, especially at shorter genomic distances below 4 Mb (r = 0.8 – 0.9), which slightly decreases at larger genomic distances (r = 0.7).

However, sections collected without nuclear staining do not always contain a nuclear profile, but might contain only cytoplasm. To be able to quantify the amount of sections with and without nuclear content, cryosections were stained with a cytoplasmic marker and a nuclear marker (analyses performed by A. Kukalev; Figure 3.3d). In 230 nm thick cryosections, approximately one out of four cell slices does not contain a nuclear profile, which means that 4 unstained NPs per sample corresponds to a mean number of 3.08 NPs (Figure 3.3e). In the subsequent sections of this thesis, I use both standard GAM with 1NP per sample and multiplex GAM with 3NPs per sample. For every test, the collection mode (1NP or 3NP) is indicated in the figure legend.



	NP1	NP2	NP3	NP4	NP5	NP6
A	-	-	-	-	-	+
B	-	-	-	-	-	+
C	+	-	-	-	-	-

	NP 1-3	NP 4-6
A	-	+
B	-	+
C	+	-

Figure 3. 3: Multiplex-GAM reduces the number of tubes per dataset.

(a) Segregation tables for 3 NPs per tube would be expected to have similar co-segregation frequencies to 1 NPs. (b) Minimal number of tubes necessary (m*) to map contacts at 30 kb resolution separated by linear genomic distances of 1 Mb to 100 Mb, calculated for mESCs (46C), with a SLICE thickness of 220 nm (analysis by C. Annunziatella, A. Scialdone) (c) HoxA locus in 46C mESCs in 480 x 1NPs (experimental data) and the in-silico merged 3NP data from the same 480 NPs (160 x 3 NPs). (d) Cryosection from 46C mESCs stained with DAPI (DNA, blue), and SytoRNA Select (RNA, green). (e) Percentage of tubes that contain 1 – 4 nuclear profiles (identified with DAPI) when collecting 4 cellular profiles (identified with SytoRNA Select). Four cellular profiles contain on average 3.08 nuclear profiles.

3.3.2 Quality assessment criteria of GAM libraries during optimisations

For assessing the quality of GAM libraries, DNA samples extracted from single NPs were sequenced and mapped to their reference genome, and to several other genomes (human, *E. coli*, yeast, others) to probe for contamination. For mouse ESCs, the percentage of reads mapping to the mouse genome is an essential mark of quality assessment, as it proves extraction of DNA from a nuclear profile, while also indicating potential contamination (represented by reads mapping to other genomes), or failures in the WGA (often represented by a high percentage of unmappable reads). Any GAM sequencing dataset that passes this first quality control is further analysed by calling positive windows at 50 kb resolution and assessing the percentage of whole genome coverage in each nuclear profile. Nuclear profiles can have different amounts of the genome covered depending on the position of the profile in the nucleus, as apical sections are smaller than equatorial sections and therefore have a lower genome coverage. To include representation from NPs with different DNA content, 48 nuclear profiles were produced per condition for most experiments to help compare the range of genome coverage of samples obtained using different experimental conditions. Further, we reasoned that a good quality nuclear profile would not only have a high percentage of reads mapping to the mouse genome, but also contain entire stretches of DNA, represented as consecutive positive windows. Thus, in a good quality nuclear profile, a relatively small proportion of positive windows are expected to be found without an immediate neighbouring window; we call this metric ‘percentage of orphan windows’. Although a larger number of metrics were extensively explored, here I use the percentage of mapped reads, genome coverage, and the percentage of orphan windows to describe the quality of a nuclear profile.

3.3.3 Cell lines used for optimising GAM

For the optimisation of GAM, two different mouse ESC lines were used, 46C and F123. 46C was used for most of the early experiments, to allow direct comparisons with the published GAM dataset (Beagrie et al., 2017). Subsequent experiments were conducted in F123 mESCs, the cell line chosen for the production of a large GAM dataset with thousands of nuclear profiles. A comparison of GAM datasets produced from either cell line shows a similar percentage of mapped reads, genome coverage, or orphan windows between F123 and 46C, and that the genome sequence differences between the lines do not seem to affect mapping efficiency to the reference genome (Figure 3.4). These analyses suggest that parameters optimised for one ESC line, are equally suitable for the other line. For the following results shown in this chapter, the cell line (46C or F123) will be indicated in the figure legend.

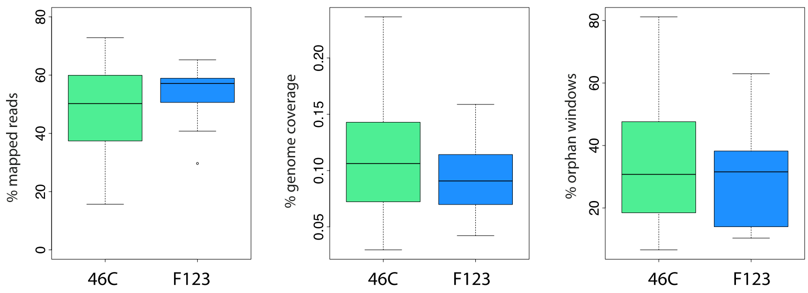


Figure 3. 4: Nuclear profiles (NPs) from the mouse ESC lines 46C and F123 show similar quality metrics. Percentage of mapped reads, the genome coverage and orphan windows in 46C and F123 NPs. Collection of multiplexed NPs with the WGA4 kit from Sigma, library prep: Nextera, 46C: 120x 3NPs, F123: 46x 3NPs.

3.3.4 Protocol parameters were optimised with different purposes

The GAM protocol can be divided in different experimental stages; cryosectioning, laser microdissection (LMD), whole genome amplification (WGA), DNA purification, library preparation, and sequencing. The efforts to improve the efficiency, hands-on time, and costs of these experimental steps and the different parameters that were addressed in each step are listed in full in the appendix (Tables 8.1 and 8.2), but for the sake of brevity, only a subset of the optimisations is discussed in detail.

3.3.5 Implementing save stopping points

First, we implemented several save stopping points for the user to be able to collect larger batches of nuclear profiles simultaneously. With the option to incubate cryosections overnight (Figure 3.5), sectioning and LMD collection can be separated, allowing the collection of larger sample numbers, and with the option to freeze cryosections, large numbers of sections can be prepared and processed at the same time if needed. With the possibility to freeze nuclear profiles after collection at the LMD microscope, several rounds of collection can be accumulated and processed together in the following WGA and library preparation steps, allowing the generation of up to 192 samples in one batch.

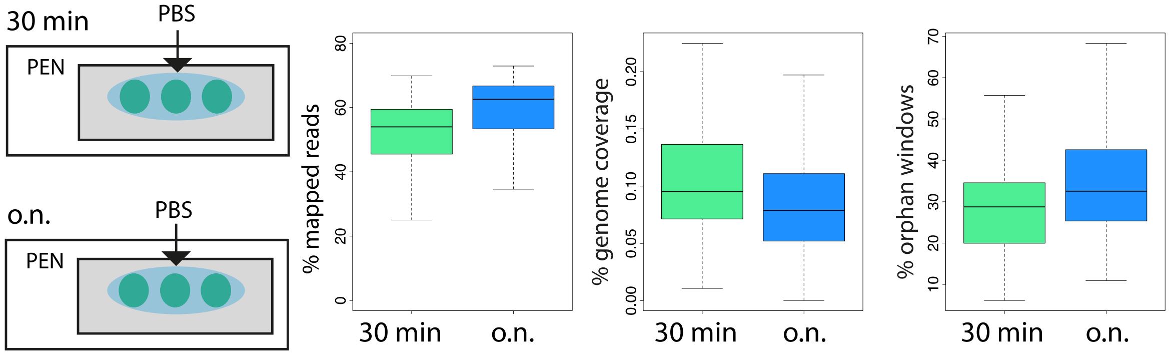


Figure 3. 5: Overnight storage of samples does not affect quality metrics of GAM samples. Cryosections can be stored in PBS overnight (o.n.) without any significant influence on the quality of nuclear profiles, as shown for the percentage of mapped reads, genome coverage, and orphan windows for the default PBS wash of 30 minutes and the overnight incubation. Cell line: F123, Collection: 3NPs, WGA: in-house, 125 x 3NPs with 30 min incubation, 264x 3NPs with overnight incubation.

3.3.6 Reducing the hands-on time per nuclear profile

To reduce the hands-on time per nuclear profile, the library preparation time was reduced, which comprised three days per 96-well plate of NPs with the TruSeq library prep kit, plus one day of DNA purifications using column purifications on the library input material from the WGA. To reduce the time for DNA purification, several methods were tested, which are suitable for large sample numbers. The original GAM protocol used a column-based DNA purification (Qiagen MinElute columns), such that each sample is processed in Eppendorf tube format. To process samples in 96-well plates, we tested a vacuum-based DNA purification (Qiagen MinElute 96 UF Plate), and solid phase reversible immobilization (SPRI) magnetic beads, which both result in good quality NPs, while reducing the hands-on time required to clean up a 96-well plate of NPs from one day to one hour. Next, we tested the compatibility of GAM with the Nextera library preparation kit. By changing from TruSeq to Nextera library preparation kits, the time required to process a 96-well plate of NPs decreases from 3 days to 1 day. The implementation of automated pipetting of the WGA reaction and of the library preparation reduced the time per NP further, while retaining library quality. In total, the time-optimised GAM protocol for a dataset of nuclear profiles (4x 96 NPs) requires 7 days of hands-on time, and 2 days of sequencing, which is a more than 4-fold reduction of time compared to the original protocol (28 days hands-on time, 12 days sequencing; Table 3.1).

Table 3. 1: Days required for the production of four 96-multiwell plates of GAM samples with the old and the new GAM pipeline.

Protocol step	Pipeline (Beagrie et al., 2017) (days)	New pipeline (days)
Collection of NPs	8	2
DNA extraction by WGA	4	2
DNA purification	4	1
Library preparation	12	2
Total hands-on time	28	7
Sequencing	12	2
Total	40	9

3.3.7 Reducing the cost per nuclear profile

With the goal to reduce the cost per nuclear profile, the number of kit-based reagents was minimised, starting from less critical steps in the protocol, such as DNA purification and library preparation. First, we adopted the use of SPRI magnetic beads for all DNA purification steps. We tested the performance of SPRI beads (produced in-house by A. Kukalev) using a published protocol (DeAngelis et al., 1995), in comparison to commercially available SPRI beads (AMPure XP, Agencourt). DNA fragment distribution and yield were identical between commercially available beads and the SPRI beads produced in-house when processing the same input of whole genome amplified DNA produced from nuclear profiles with either type of beads (Figure 3.6a). The in-house produced SPRI beads were implemented into the revised GAM pipeline.

To further reduce the cost of library preparation, we scaled down the volumes recommended by the manufacturer of the Nextera library preparation kit, according to published recommendations (Lamble et al., 2013). We prepared libraries from the same WGA-material extracted from nuclear profiles using either the recommended volume or reduced volumes to 20% of the original. Samples were sequenced and the distribution of reads in sequencing tracks were found identical between the two conditions (Figure 3.6b). To extend this exploration, we divided the genome in 50 kb windows and called positive signal finding excellent overlap of positive windows genome-wide (Figure 3.6c). The similarity between the full volumes and the reduced volumes produced at the same time exceeded the similarity between technical replicates (both full volumes) produced at different days. Thus, we decided to implement the reduced volume library preparation in the revised in-house GAM pipeline, which reduced the costs per NP immensely.

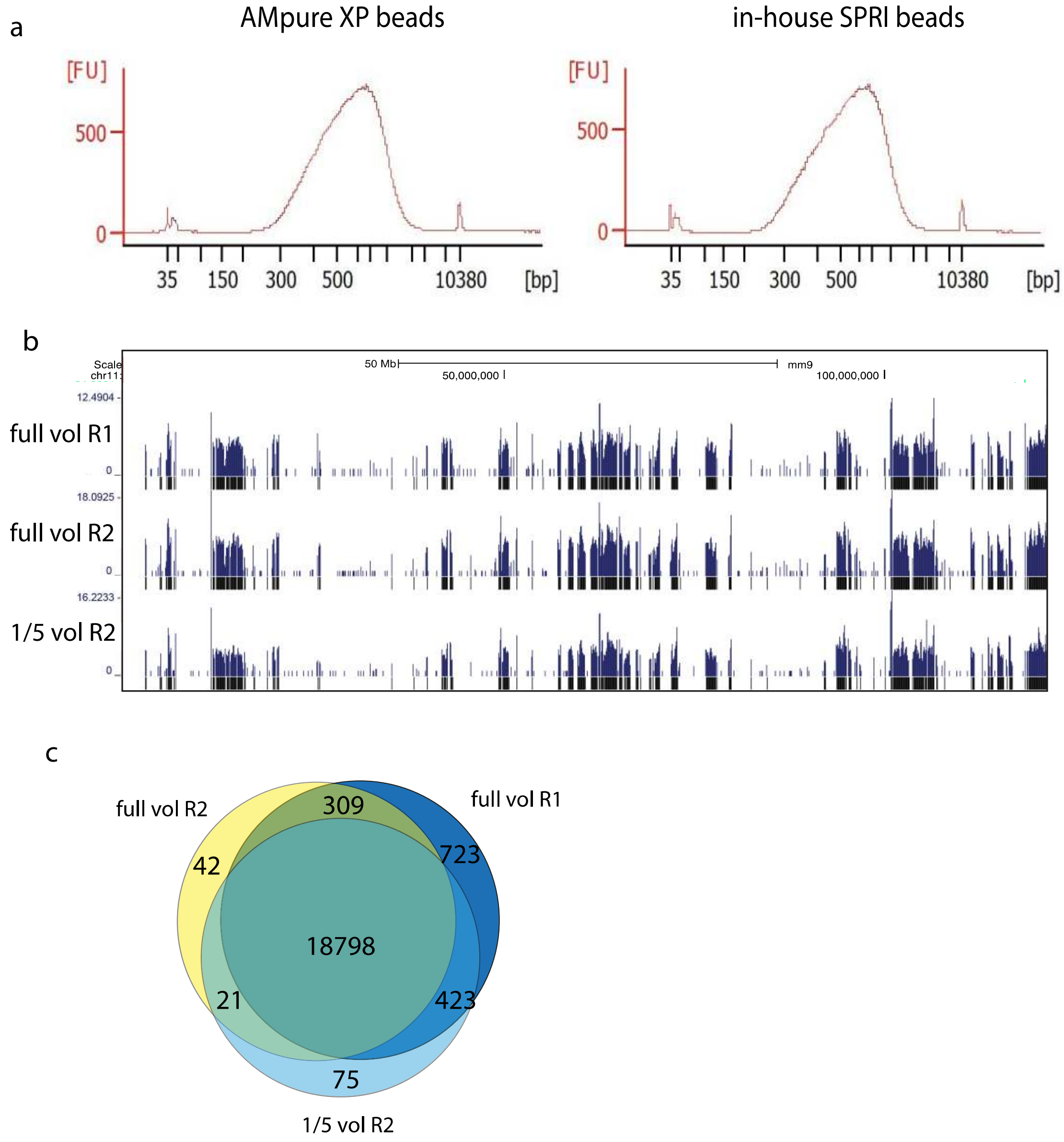


Figure 3. 6: Cost reduction per nuclear profile.

(a) Electrogram of amplified and purified DNA from F123 nuclear profiles after WGA. DNA purification with AMpure XP beads and in-house SPRI beads results in identical DNA fragment length distributions. FU, fluorescence units (b) Sequencing tracks mapping to mouse chromosome 11 for two technical replicates of library preparation with full volumes, and a replicate using only 1/5 of the reaction volume. Positive 50 kb windows (black) are displayed under each track. (c) Venn plot displays the common and different positive windows for technical replicates of library preparations.

3.3.8 Visualisation of nuclear profiles

In the original GAM protocol, NPs were collected on the LMD under transmitted light illumination that helped visualise sections from cells, but does not allow choosing of cellular slices that contain nuclei (Beagrie et al. 2017). For this reason, several cellular slices did not produce useful data. To improve the efficiency of NP collection, we aimed to find a nuclear or cellular marker that would allow visualisation of cellular slices containing nuclei under the LMD microscope.

First, we tested different DNA dyes that could potentially stain nuclear profiles. Most commonly used DNA dyes work well in cryosections for imaging NPs in standard fluorescence microscopy. However, they are more challenging to apply in LMD microscopy, due to the high autofluorescence of the membrane slides used for LMD, which excluded some potential dyes from being suitable for GAM (Supplementary table 8.2). To be useable in the GAM protocol, the staining also has to be compatible with the subsequent DNA extraction by WGA. All DNA dyes tested gave no or little DNA extraction, suggesting that the binding of the fluorescent dye to the DNA directly interferes with successful DNA extraction and amplification (Figure 3.7a, b). This led to the exploration of histological stains, which do not specifically label nuclei, but that can be used to more robustly find cellular slices. Here, I found several dyes that stain cryosections (Supplementary table 8.2) and could potentially enable easier and more reliable collection of nuclear profiles, with one of them being also compatible with subsequent DNA extraction by WGA (Figure 3.7c, d).

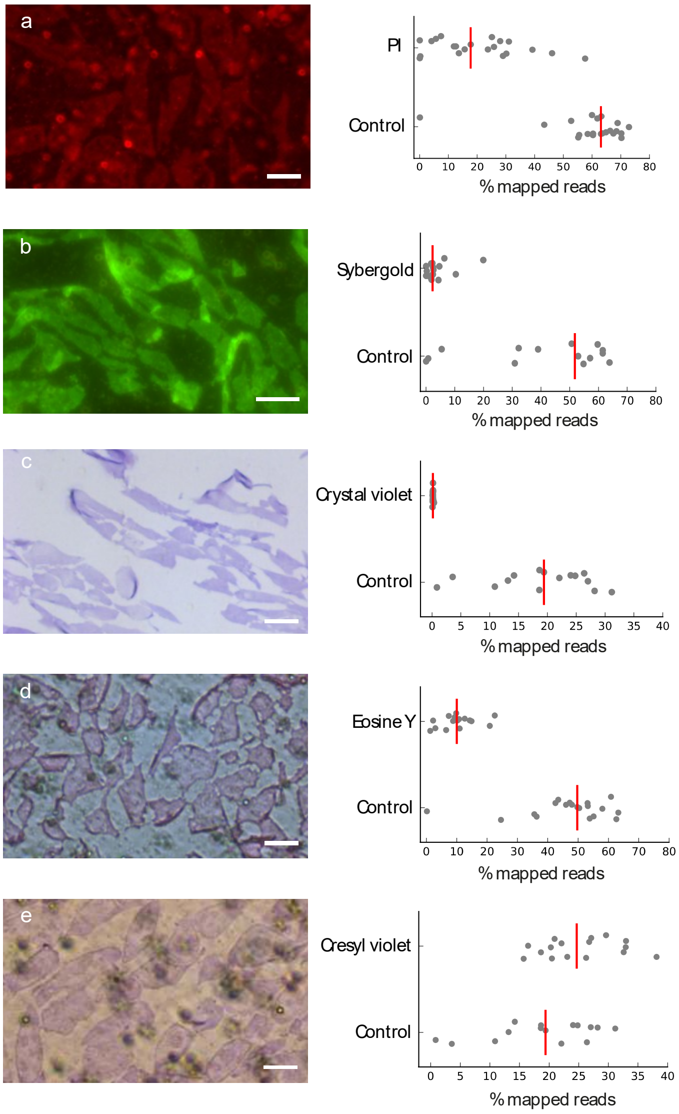


Figure 3. 7: Visualisation of nuclear profiles in GAM.
Most stains tested to label nuclear profiles in cryosections gave very low percentage of mapped reads per NP. Control: unstained NPs. (a) Propidium iodide (PI) staining on LMD slides is visible under the LMD microscope but interferes with DNA extraction by WGA. (b) SybrGold staining is visible under the LMD microscope but interferes with DNA extraction by WGA. (c) Crystal violet staining is visible under the LMD microscope but interferes with DNA extraction by WGA. (d) Eosine Y staining is visible under the LMD microscope but interferes with DNA extraction by WGA. (e) Cresyl violet staining is visible under the LMD microscope and allows DNA extraction by WGA. Cell line: 46C, collection: 3NPs, WGA: Sigma WGA4.

Cresyl violet qualified as a cellular marker to effectively collect nuclear profiles, however it still had the drawback of not visualising nuclei, which means that a proportion of the nuclear profiles would potentially contain a cellular slice without genomic DNA. Thus, I explored using immunofluorescence of DNA binding proteins. I showed that nuclei can be visualised under the LMD microscope after staining with a pan-histone antibody (Figure 3.8a), and most importantly, the immunofluorescence protocol had no negative effect on the subsequent DNA extraction. Compared with cresyl-stained samples, the cryosections stained by immunofluorescence have improved quality of DNA extraction, as shown by an increase in the percentage of mapped reads, genome coverage, and a decrease in percentage of orphan windows (Figure 3.8c, d).

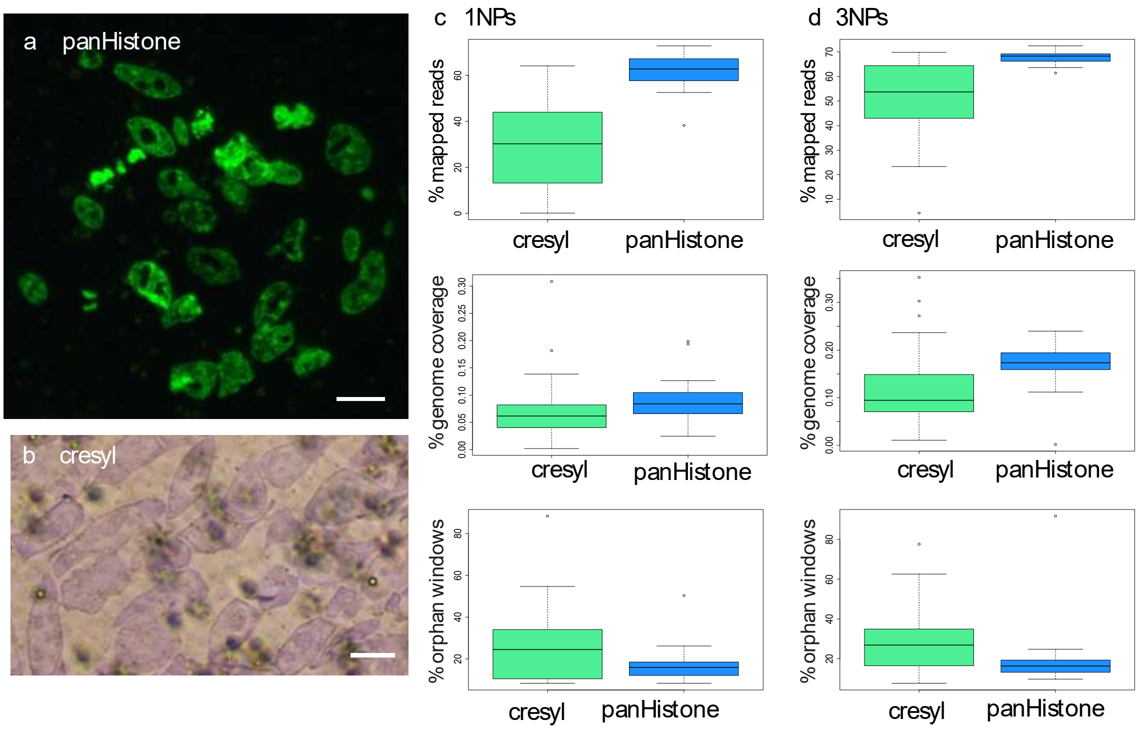


Figure 3. 8: GAM can be combined with immunofluorescence.
(a) Anti-pan-histone staining under the LMD microscope. Staining of histones using a pan-histone antibody raised in mouse, followed by indirect immunodetection with fluorochrome-Alexa anti-mouse antibodies. (b) Cresyl violet staining under the LMD microscope. (c) Percentage of mapped reads, genome coverage, and orphan windows in single nuclear profiles for cresyl-stained samples and samples undergoing an immunofluorescence staining against pan-histone. N = 31 NPs (cresyl), 24 NPs (pan-histone) (d) Percentage of mapped reads, genome coverage, and orphan windows in multiplexed nuclear profiles for cresyl-stained samples and samples undergoing an immunofluorescence staining against pan-histone. N = 54 x 3 NPs (cresyl), 23 x 3 NPs (pan-histone). Cell line: F123, WGA: in-house.

3.3.9 Laser-microdissection

Next, I tested several parameters related with the LMD collection of nuclear profiles, such as the material of the membrane slide, or the orientation of the sample in the LMD microscope (Figure 3.9a). Interestingly, while the orientation of the sample in the LMD has no effect on NP quality (Figure 3.9b), different membrane types strongly influence the extraction efficiency, resulting in samples with strongly reduced percentages of mapped reads (Figure 3.9c). For the collection of NPs, we use opaque-filled caps (Leica), specially designed for laser microdissection, which are thought to reduce statics during dissection. To test the effect of the opaque-filled caps on subsequent DNA extraction, I used standard PCR caps to collect NPs and compared them with the LMD caps. In fact, NPs cut into PCR caps are of slightly reduced quality in comparison to the LMD cap, noticeable mostly in a reduced percentage of mapped reads (Figure 3.9d). Thus, although PCR caps have the advantage of being less expensive than LMD caps, they are not recommendable for GAM. I also found that Pol and

PPS membranes are not suitable. Therefore, no changes of the materials used for laser microdissection were implemented in comparison to the original protocol.

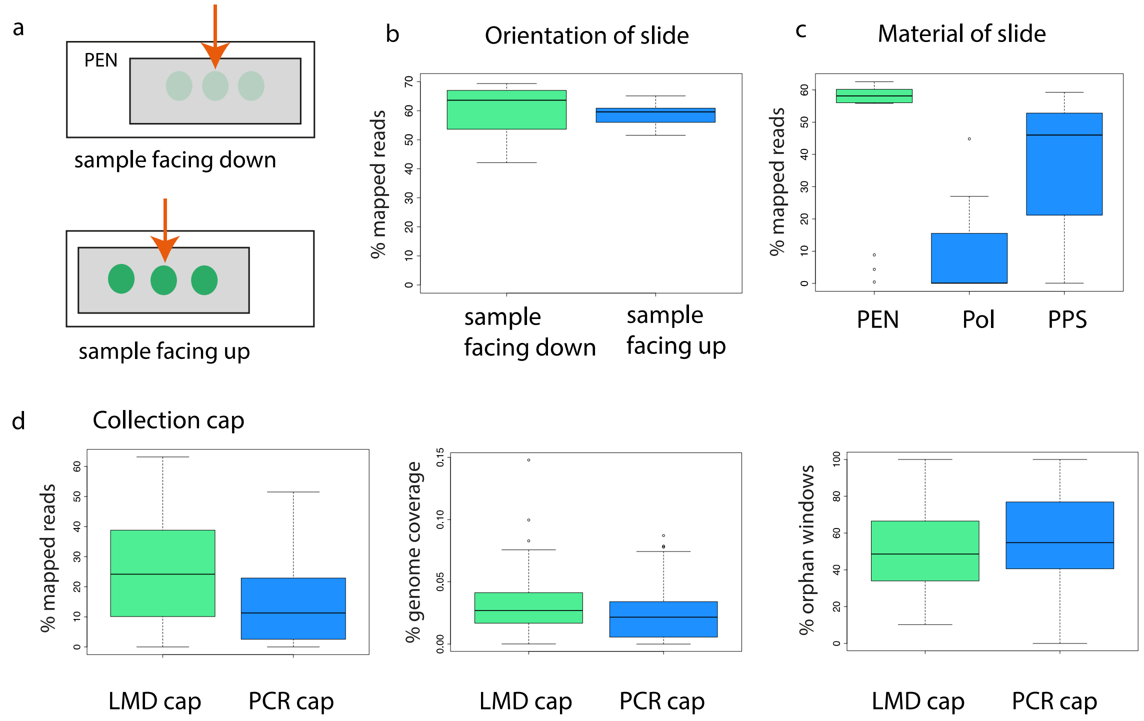


Figure 3. 9: Testing different parameters for laser microdissection of nuclear profiles. (a) Different orientation of the membrane slide in the LMD microscope may result in different exposure of the sample to the laser beam. (b) Orientation of the membrane slide in the LMD microscope does not affect DNA extraction from nuclear profiles, as shown by the percentage of mapped reads per NP. N = 14 x 3NPs facing up, 14 x 3NPs facing down. Cell line 46C, WGA: Sigma WGA4. (c) Different membranes influence the efficiency of DNA extraction from nuclear profiles, as shown by the percentage of mapped reads per NP. N = 16 x 3NPs (PEN), 16 x 3NPs (Pol), 8 x 3NPs (PPS). Cell line 46C, WGA: Sigma WGA4. (d) Nuclear profiles fall into a collection cap after LMD. Percentage of mapped reads, genome coverage, and percentage of orphan widows per NP show slightly reduced quality for PCR caps compared to LMD caps. N = 79 NPs (LMD cap), 112 NPs (PCR cap). Cell line: F123, WGA: in-house.

3.3.10 Whole Genome Amplification using different methodologies

(a) DNA extraction with the WGA4 kit from Sigma

The efficiency of the WGA reaction in the extraction and amplification of DNA from single nuclear profiles is an essential step for the successful analysis of 3D genome topology with GAM. We tested various conditions in the WGA reaction, first by optimising the performance of the WGA4 kit from Sigma Aldrich, a kit designed to extract and amplify DNA from single cells. The WGA reaction relies on lysis and DNA fragmentation, annealing of universal primers to the fragmented DNA, and amplification. Although the DNA content of the GAM nuclear slices is 20-50x lower that a single cell, the WGA4 kit was successfully used in

Beagrie et al. (2017) to extract DNA from nuclear profiles, reaching 83 % efficiency in the extraction of 30 kb windows (Beagrie et al., 2017).

In 2017, while testing different experimental parameters in the WGA4 kit (Table 3), we noticed a sudden drop in the quality of GAM sequencing data from nuclear profiles. We noticed that all WGA4 kits received after January 2017 failed to extract DNA from GAM nuclear profiles, although they could still amplify control genomic DNA at low concentrations equivalent to a single cell. To better understand the cause of the decreased quality of GAM datasets found in three consecutive batches of GAM collection, we quantified the percentage of mapped reads, the genome coverage and percentage of orphan windows in data collected with batches received before and after January 2017 (Figure 3.10a). Consistently, the whole batch of NPs collected with the new WGA4 kits showed a low percentage of mapped reads, low genomic coverage and a high fraction of orphan reads, indicating inefficient DNA extraction. The low quality of these datasets was confirmed by examining the fragment size distribution of DNA extracted from NPs with the WGA4 kit before and after January 2017 on an agarose gel (Figure 3.10b). NPs from the affected batches displayed bulks of small DNA fragments in almost every sample. These low molecular weight fragments are likely to be dimers of WGA primers that ligate in the absence of extracted DNA from NPs. Visualisation of the sequencing and positive window tracks of poor-quality NPs processed for GAM before or after January 2017 also highlighted strong discrepancies between detection of raw reads and ability to call positive windows from the later datasets (Figure 3.10c). The window calling approach previously devised for GAM data to identify signal and noise in each sample was based on fitting a binomial curve on the reads distribution, before setting a sample-specific threshold based on the fitting of the curves (Beagrie et al., 2017) (Figure 3.10d). A sample-specific threshold is required to call windows in GAM data due to the fact that each dataset contains varied amounts of DNA depending on whether the nuclear slices are equatorial or apical, resulting in different coverage and number of reads per genomic region (Beagrie et al., 2017).

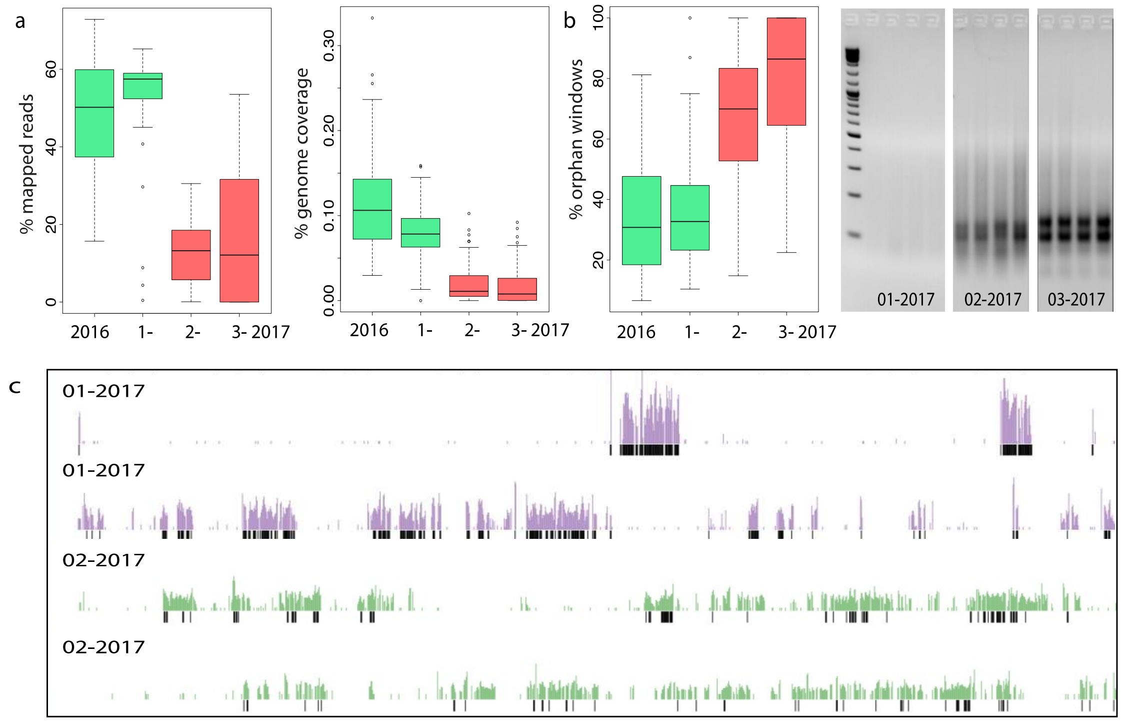


Figure 3. 10: Efficiency of DNA extraction with the WGA4 kit from Sigma dropped unexpectedly. (a) A sudden drop in NP quality appeared in all GAM batches after January 2017, noticeable by a low percentage of mapped reads, low genome coverage, and a high percentage of orphan windows. (b) Low quality metrics appeared simultaneously with bulks of small DNA fragments after WGA, visible when running DNA from nuclear profiles on a 1.2% agarose gel. DNA ladder: Gene Ruler 1kb plus. (c) Sequencing tracks mapping to mouse chromosome 16 from the January and February batch (2017) of the WGA4 kits. Tracks of positive windows underneath the sequencing tracks show less efficient window calling in the February samples, which explains the low genome coverage in these samples. Cell lines: F123 and 46C. N = 120 x 3 NPs (2016), 47 x 3 NPs (01-2017), 94 x 3 NPs (02-2017), 166 x 3 NPs (03-2017), 15 x 3 NPs (10-2016).

Visual inspection of the curve fitting graphs showed that the original curve fitting failed to identify the noise, setting the threshold of positive windows too high. To overcome this problem, we tested alternative approaches to call windows, and developed one that sets the threshold in each sample by defining a percentile rank for the minimum nucleotide coverage required for calling a positive window. The percentile itself is defined for a set of NPs (e.g. a full GAM dataset, or a batch of NPs) and represents the noise in this particular set of samples (typically this percentile is between 77 to 81 for the current GAM datasets, depending on the window size). This approach allowed recovery of positive windows from nuclear profiles in all batches, including those from the batch from January 2017, suggesting the revised approach to be more robust to the quality of sequences from GAM datasets. When applying the coverage-based window calling approach, both genome coverage and the percentage of orphan windows improved (Figure 3.11a, b).

Next, we tested whether the revised coverage-based window calling improved the detection of windows in the batches of nuclear profiles from February 2017 onwards. Importantly, most of the collected NPs had a low percentage of mapped reads (<15 %), which drastically reduces the genomic information that one NP can provide. When taking into consideration only the samples that pass this first quality criterion of >15% mapped reads, and applied the coverage-based window calling, we could obtain the expected range of genome coverage and percentage of orphan windows in the WGA4 batch of 02-2017, but the improvements introduced were not sufficient to reach the range of good quality values in the 03-2017 or subsequent batches produced with the latest batches of the Sigma WGA4 kit (Figure 3.11c). Therefore, it was necessary to find alternative WGA kits or to develop our own.

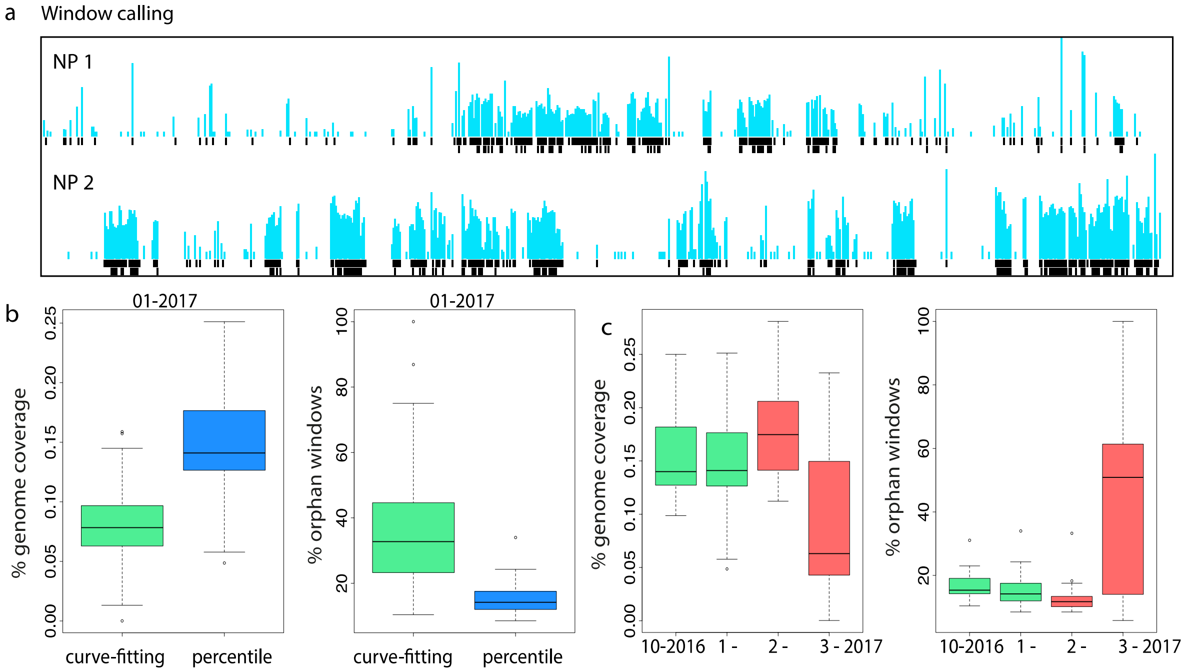


Figure 3. 11: A new window calling approach improves the recovery of positive windows from NPs. (a) Sequencing track of two nuclear profiles and positive windows called with the percentile-based approach (top), and the previously used curve-fitting approach (bottom). (b) Genome overage and percentage of orphan windows for the January 2017 batch of nuclear profiles processed with the curve-fitting and the percentile-based windows caller. (c) Genome coverage and percentage of orphan windows in NPs with more than 15% mapped reads (current quality control criteria for good quality NPs) with the coverage-based window calling. Cell lines: F123 and 46C. N = 120 x 3 NPs (2016), 47 x 3 NPs (01-2017), 94 x 3 NPs (02-2017), 166 x 3 NPs (03-2017), 15 x 3 NPs (10-2016).

(b) Testing other WGA methodologies for GAM

To identify an alternative to the Sigma WGA4 kit, we first tested different commercially available WGA kits from other suppliers, and evaluated their performance on NPs. Those were in particular, the Ampli-1 kit from Menarini Silicon Biosystems, the Multiple Annealing and Looping Based Amplification (MALBAC) kit from Yikon Genomics, and the Repli-G kit from Qiagen. The challenge with these alternative commercially available kits is that they have been optimised for extracting DNA from single cells without crosslinking. GAM requires the extraction of DNA from a small fraction of the cell (~230 nm thin nuclear slices), corresponding to approximately ~1/40 of an mESC nucleus, which contain highly crosslinked chromatin (fixation in 8% formaldehyde for 2h). The only kit that gave promising results was the MALBAC kit from Yikon Genomics, which produced GAM samples with good percentage of mapped reads, but lower coverage and high percentage of orphan windows, compared with those produced with WGA4 kits obtained prior to 2017 (Figure 3.12a, b). Although we were able to improve the performance of the MALBAC kit, the GAM datasets produced with this kit had unreliable quality, often with high percentage of unmappable reads, or *E. coli* contamination (Figure 3.13), making it unsuitable for high-throughput production of NPs.

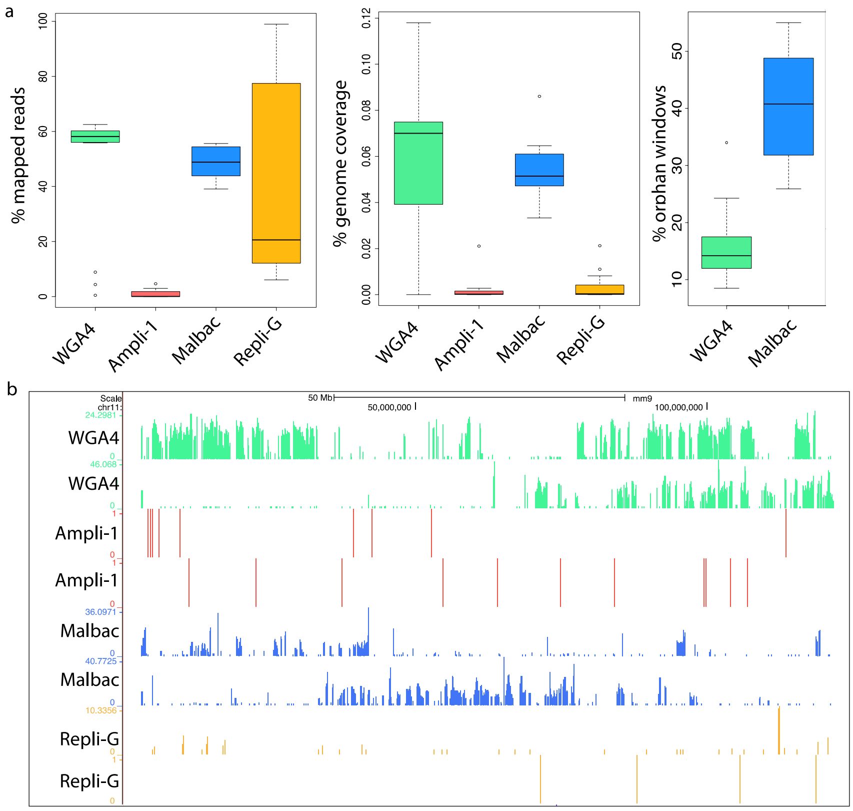


Figure 3. 12: Testing different WGA reactions for GAM.
(a) Percentage of mapped reads, genome coverage, and orphan windows for different WGA reactions. Cell line: F123: Sigma n = 32 x 3NPs, Ampli-1 n = 8 x 3NPs; MALBAC n = 8 x 3NPs. (b) DNA from multiplexed nuclear profiles (3 NPs) mapping to mouse chromosome 11 (mm9) for different WGA reactions.

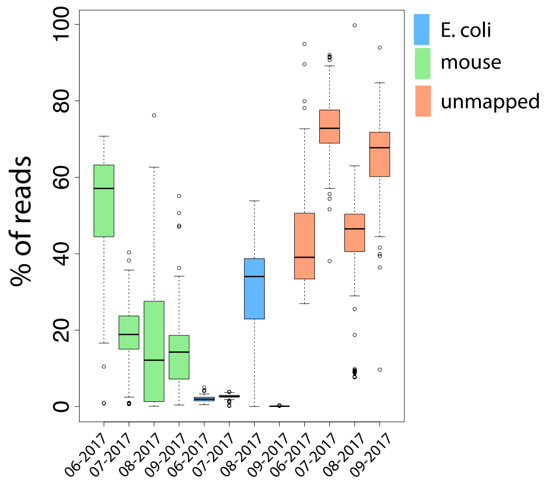


Figure 3. 13: Variability in quality of the MALBAC kit from Yikon.
Percentage of reads mapping to mouse and *E. coli*, and unmappable reads in NPs processed with different lots of the MALBAC kit. Numbers of NPs from left to right: 40 x 3NPs, 96 x 3NPs, 112 x 3NPs, 96 x 3NPs. Cells: F123 mESCs collected by myself, and other murine cells collected by I. Harabula and A. Kukalev.

(c) Whole Genome Amplification independent of commercially available kits

To overcome the limitations of commercially available kits, the only alternative option was to develop an in-house protocol for WGA. We took inspiration from the MALBAC protocol, in which random hexamers bind genomic DNA in multiple annealing steps before amplification. The amplification products contain a common primer sequence, which was attached to the random hexamers, and are then further amplified with primers annealing to the common sequence. MALBAC has been used for DNA and RNA amplification in single cells (Chapman et al., 2015; Zong et al., 2012). Here, we show that a modified version of MALBAC can be used to amplify DNA from crosslinked nuclear profiles, with similar performance to the previously used WGA4 kit from Sigma (Figure 3.14).

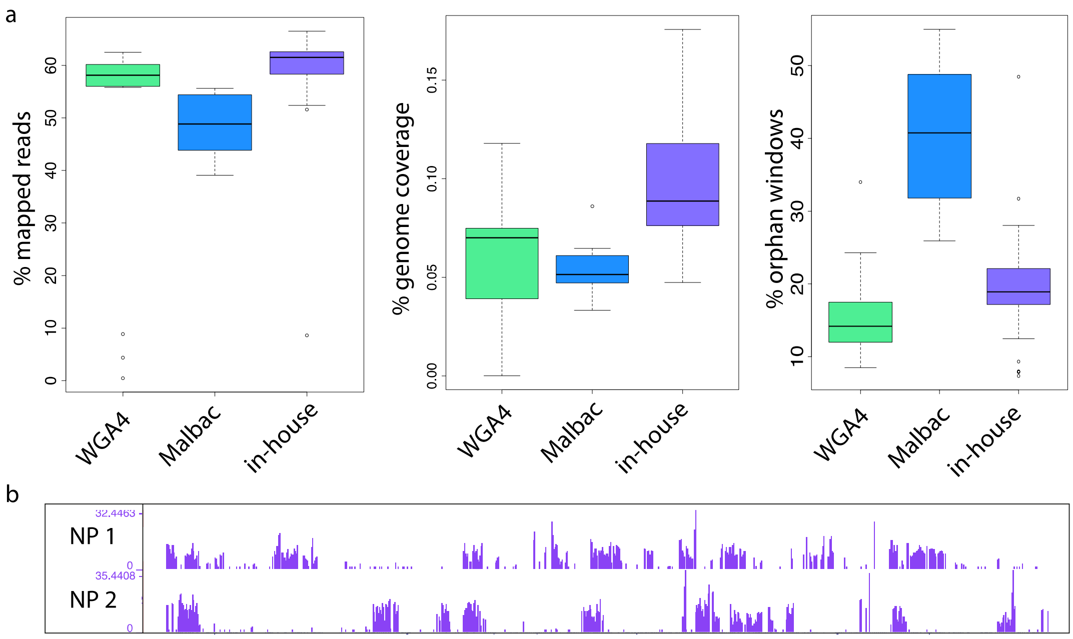


Figure 3. 14: The in-house version of MALBAC extracts good quality DNA from NPs.
(a) Percentage of mapped reads, genome coverage, and orphan windows for different WGA reactions. Cell line: F123: Sigma n = 32 x 3NPs; MALBAC n = 8 x 3NPs; in-house n = 15 x 3NPs. (b) DNA from multiplexed nuclear profiles (3 NPs) mapping to mouse chromosome 11 (mm9) for the in-house WGA reaction.

To improve the WGA protocol, we started by testing several parameters to enhance the efficiency of DNA extraction. A critical condition of the WGA of GAM NPs is the treatment with protease to reverse the crosslinking and lyse the nuclear slice. Thus, we tested the influence of incubation time and concentration of protease during the incubation on DNA extraction efficiency. We found that increasing the incubation time has little effect on the quality of nuclear profiles (Figure 3.15a), whereas protease concentration is critical for successful DNA extraction (Figure 3.15b). To ensure optimal DNA amplification, we tested several parameters in the PCR reaction, such as concentration of nucleotides and magnesium (Figure 3.15d), volume of the WGA reaction (Figure 3.15e), annealing time of the hexamers and PCR elongation times (Figure 3.15c), or PCR primer annealing temperature (Figure 3.15f). While most of these changes had a negative effect on DNA amplification and were not implemented in the protocol, prolonged annealing and elongation times appear to have a positive influence on DNA extraction and could potentially be implemented in the protocol (Figure 3.15c).

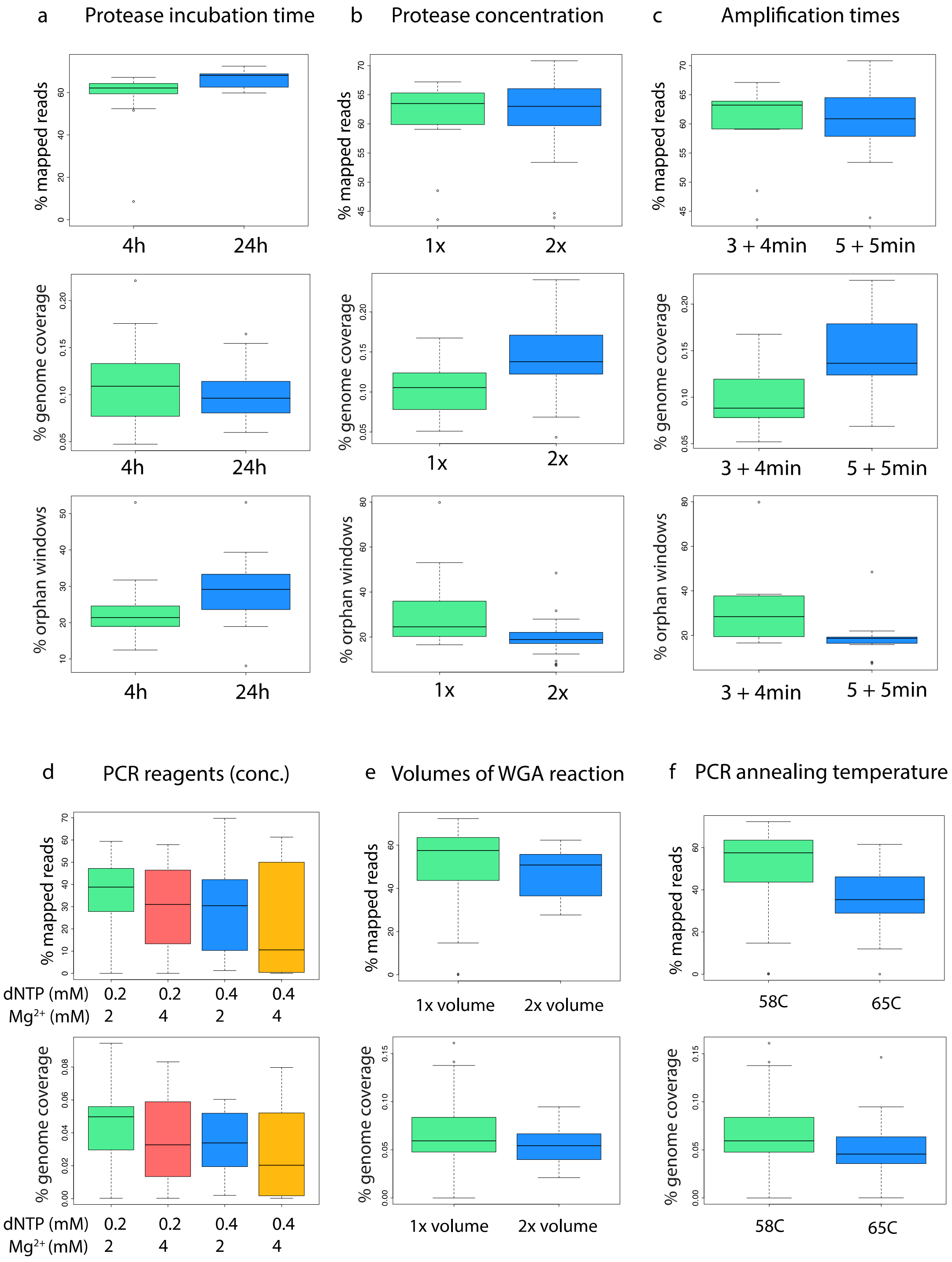


Figure 3. 15: Improving the performance of the in-house WGA reaction.
Percentage of mapped reads, genome coverage, and orphan windows for different WGA conditions. Collection: 3 NPs, cell line: F123, WGA: in-house, n varies between plots. Protease incubation time n = 23 x 3NPs each; Protease concentration n = 23x 3NPs (1x protease), 47x 3 NPs (2x protease); Amplification times n = 16x 3NPs (each); Concentrations of PCR reagents n = 16x 3NPs (each) ; Volumes of WGA reaction n = 63x 3NPs (1x vol), 15x 3NPs (2x vol), PCR annealing temperature n = 63x 3NPs (58°C), 38x 3NPs (65°C).

3.4 Final decisions for data collection of nuclear profiles in F123 mESCs

Considering all the tested conditions, the following protocol was finally chosen; Cells are cryosectioned (230 nm) and transferred to a PEN-membrane slide with a 2.1M sucrose in PBS drop, washed with PBS for 30 min, and stained with cresyl violet. Then 1 or 4 cresyl-stained NPs are laser-microdissected into a single 8-well LMD collection cap. The DNA is extracted and amplified using the in-house WGA protocol based on MALBAC, with a 2x concentration of Qiagen protease, at 60°C, for 4 h. DNA is amplified with 3-min primer annealing, and 4 min elongation time, at 58°C annealing temperature, and with 11 cycles in the preamplification step, followed by 26 cycles in the second amplification of the WGA. After WGA, DNA is purified using 1.7x (for 1NP) or 0.8x (for 3NPs) SPRI beads (in-house production), quantified with PicoGreen, then 1 ng per sample is used as input for the Nextera XT library preparation kit. The library reaction is scaled down to 1/5 of the recommended volumes from the manufacturer to save costs. Libraries are purified with 1.7x SPRI beads (in-house production) and library DNA quantified using PicoGreen. Next, the same amount of DNA is pooled from each library into a sequencing pool containing 192 GAM libraries, and sequenced on the NextSeq 500 sequencing device with 75 bp single-end sequencing, to an average sequencing depth of 2.7 million reads per library.

While optimising the GAM pipeline, many successful GAM samples were collected from 1NP and 3NPs in the F123 cell line. Samples that passed general quality metrics (a high proportion of reads mapping to the mouse genome) were included in the F123 dataset. However, the majority of data collection for a deep GAM dataset in F123 was produced subsequently with the above-mentioned final conditions. The samples produced during optimisation had the same distribution of quality metrics when compared with the batch of GAM samples collected subsequently.

3.5 Exploration of datasets produced with different GAM protocols

When working with smaller numbers of NPs during the protocol optimisations, only basic quality assessments as described above are feasible, due to the nature of the GAM method, which works by sampling nuclear slices in random orientations and with varied DNA content, and results in contacts maps of complete genomes only after combining the information from a few hundred nuclear profiles. To fully understand the efficiency of DNA detection in nuclear profiles collected with different GAM protocols, deeper explorations of the entire collection of GAM datasets is necessary. To help assess the quality of GAM datasets produced with the new GAM pipeline using F123 mESCs, I used the previous published

GAM dataset produced in 46C mESCs. The 46C dataset comprises 471 single NPs (of which 408 passed quality controls) and 5 negative controls, collected in 12 batches (Beagrie et al. 2017). The F123 dataset collected in 1NP mode was used for this comparison, and consists of 1281 samples and 56 controls produced after laser-microdissecting LMD membrane without cells, collected in 22 batches, all with the in-house WGA protocol. Before evaluating NP quality, cross-contamination between NPs from the same batch was investigated by measuring the similarity of positive windows between all wells in a multiwell plates (described in 2.20.3). Thirteen samples were excluded from the F123 dataset leaving 1268 NPs before applying quality cut-offs.

3.5.1 Quality metrics of nuclear profiles collected with different GAM pipelines

First, I evaluated the consistency of the percentage of mapped reads (to mm10) over the period of data collection (Figure 3.16). In both datasets, 46C and F123, the majority of samples passed the quality control introduced previously in Beagrie et al. (2017) of >15 % mapped reads. However, there are noticeable differences between the datasets. Most of the 46C batches tend to have larger proportions of NPs with a high percentage of mapped reads (more than 50 % mapped reads), whereas NPs in many batches in F123 show a widespread distribution of % mapped reads (about 20 – 70 % mapped reads) per dataset.

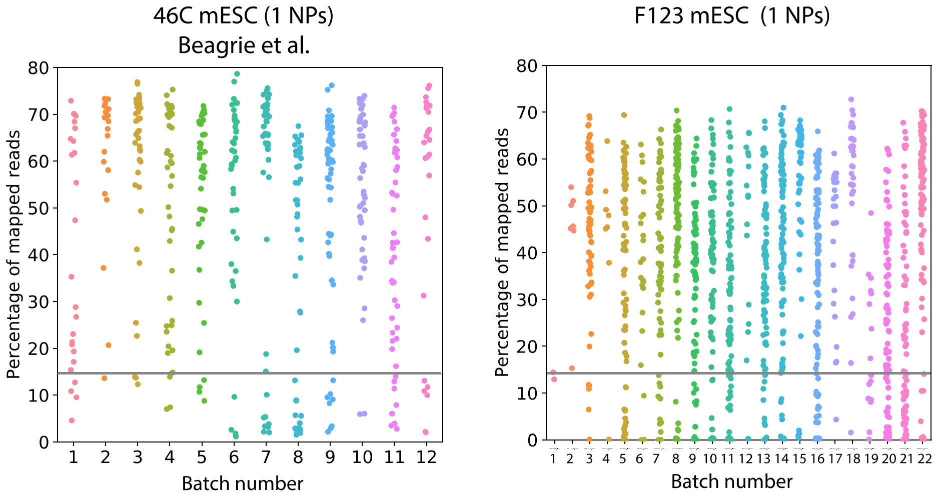


Figure 3. 16: Nuclear profiles show consistent quality metrics between batches.
Each dot represents individual 1NP GAM samples from mESCs 46C and F123 clones. The percentage of mapped reads is shown for each batch of NPs, in chronological order of collection, for 471 single NPs from 46C collected with the WGA4 kit from Sigma, and 1281 single NPs from F123 collected with the in-house WGA protocol.

A number of additional features and quality metrics were assessed in both datasets (Table 3.2). Previous work showed that an average sequencing depth of 1 million reads was sufficient to saturate the detection of 30 kb positive windows in single GAM datasets (Beagrie et al., 2017). We decided to sequence F123 nuclear profiles to a higher average sequencing depth of 2.7 million reads per NP, in contrast with 1.1 million reads per NP in 46C. As expected, the increase in sequencing depth in F123 is accompanied by a proportional increase of uniquely mapped reads per NP (Figure 3.17). Another hallmark of NP quality is potential contamination with DNA from other species. Here, the majority of NPs from the F123 dataset has less contamination with other genomes than the 46C dataset (Figure 3.17), probably due to the LMD collection of F123, which took place in a separate room, whereas for the 46C data collection, the LMD equipment was located in a shared room.

Table 3. 2: Quality metrics of NPs.

Metric	Description
Total reads	Number of sequenced reads (sequencing depth)
Percentage of mapped reads	Percentage of reads mapping to the mouse genome
Mapping to human	Percentage of reads mapping to the human genome
Mapping to *E. coli*	Percentage of reads mapping to the *E. coli* genome
Mapping to *S.saccharomyces*	Percentage of reads mapping to the *S.saccharomyces* genome
Uniquely mapped reads	Number of reads mapping uniquely to the mouse genome (after removal of PCR duplicates)
Genome coverage	Number of positive windows (50 kb) in the genome divided by the total number of windows in the genome.
Orphan windows	Percentage of positive windows (50 kb) that are surrounded by negative windows for each chromosome. Then, the average of all chromosomes is calculated.
Number of chromosomes	To count the number of chromosomes, the fourth quartile of the number of positive windows in one megabase is calculated. Afterwards, the chromosomes that contain at least one megabase window with more positive windows than the fourth quartile are counted as present chromosomes.

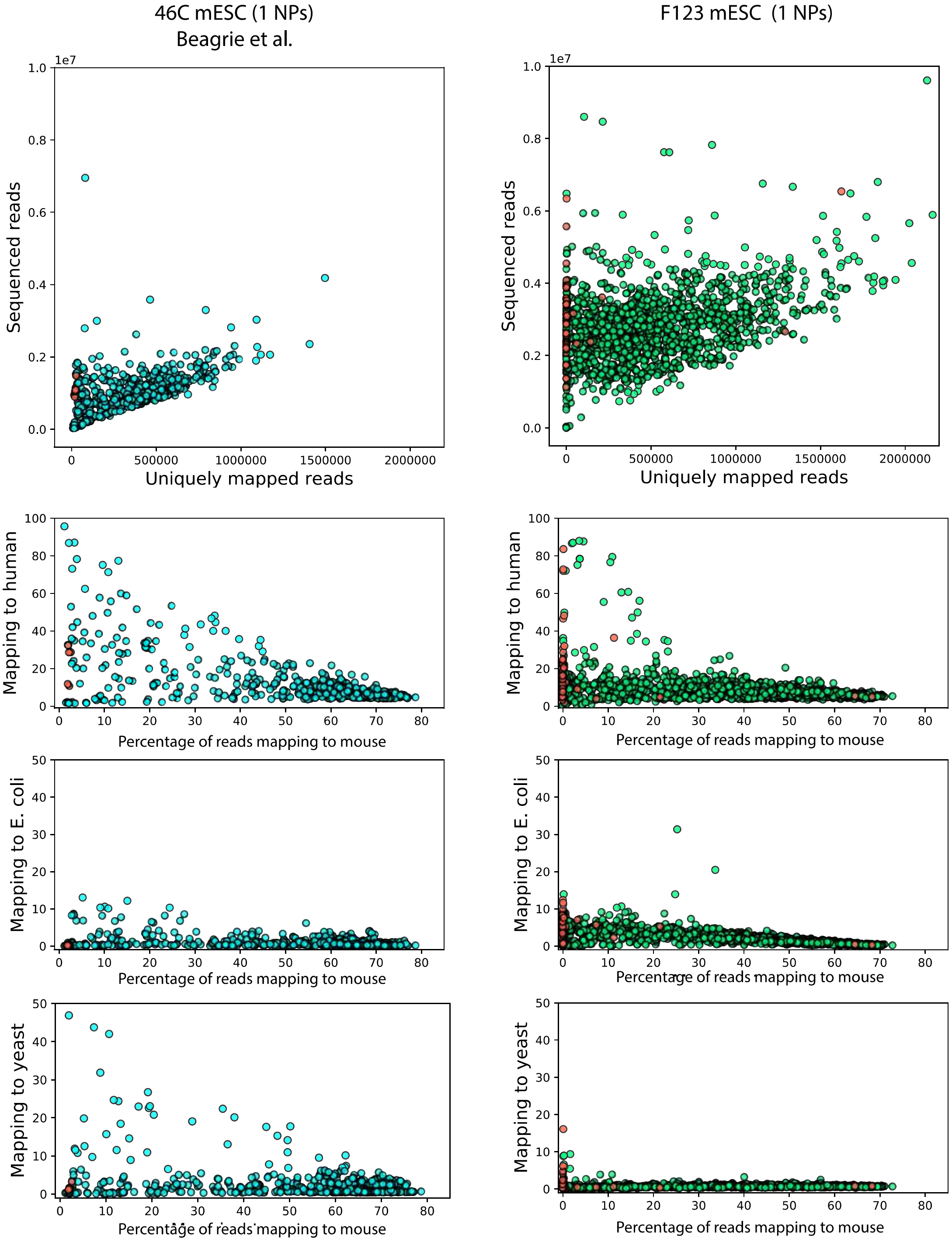


Figure 3. 17: Sequencing reads of 46C (blue) and F123 (green) single NPs map predominantly to the mouse genome.

From top to bottom: Number of sequenced reads vs. uniquely mapped reads to mouse; percentage of reads mapping to the human genome vs reads mapping to mouse; percentage of reads mapping to the *E. coli* genome vs. reads mapping to mouse; percentage of reads mapping to the yeast genome vs. reads mapping to mouse. Negative controls are shown in red.

Next, I examined four quality metrics that best describe the quality of 1NP GAM data, namely the number of uniquely mapped reads, genome coverage, percentage of orphan windows, and number of chromosomes. Both 46C and F123 datasets follow a similar general trend, where having more uniquely mapped reads results in higher genome coverage, accompanied by a small tendency to cover a larger number of chromosomes, and a lower percentage of orphan windows (Figure 3.18). This trend is less pronounced for the F123 NPs, where many uniquely mapped reads does not result in a proportional increase in genome coverage as observed in 46C (Figure 3.18). Samples with higher number of chromosomes do not necessarily have a high genome coverage, suggesting that parts of the genome were not recovered out of the NP (Figure 3.18). Most strikingly, F123 NPs often have higher percentages of orphan windows, even when larger numbers of chromosomes are detected (Figure 3.18). Taken together, these results suggest that the F123 data collected had less efficient extraction of DNA.

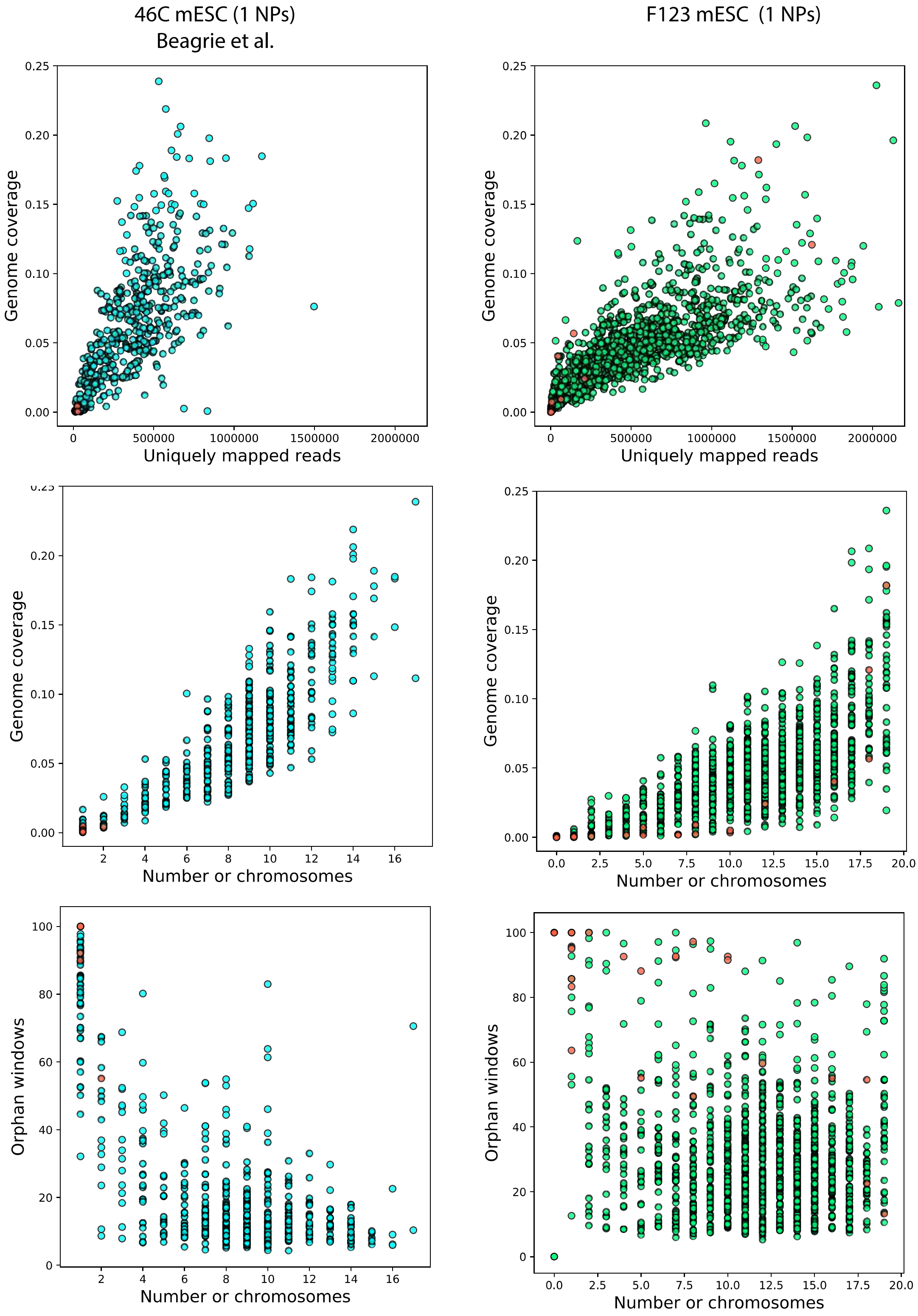


Figure 3. 18: NPs in the F123 dataset cover a lower percentage of the genome with the same or more uniquely mapped reads than 46C.

Quality metrics of 46C (blue) and F123 (green) from top to bottom: Percentage of genome coverage vs. number of uniquely mapped reads; percentage of genome coverage vs number of chromosomes; percentage of orphan windows vs number of chromosomes. Negative controls are shown in red.

To test whether the samples with lower extraction in F123 belonged to specific batches of NPs, I next studied batch effects in the genome coverage and the percentage of orphan windows, the two criteria that were most different between the two datasets (Figure 3.17). For most batches, the genome coverage per NP in both 46C and F123 is very consistent, with only mild differences in the sample distribution between the two datasets. A similar trend is observed for orphan windows, but, while the majority of NPs from 46C has less than 30% orphan windows, NPs from F123 are more heterogeneous, with many samples having up to 50-60 % of orphan windows. These analyses showed that the quality criterium of >15 % mapped reads used in Beagrie et al. (2017) was not sufficient for removing lower quality samples from the F123 dataset.

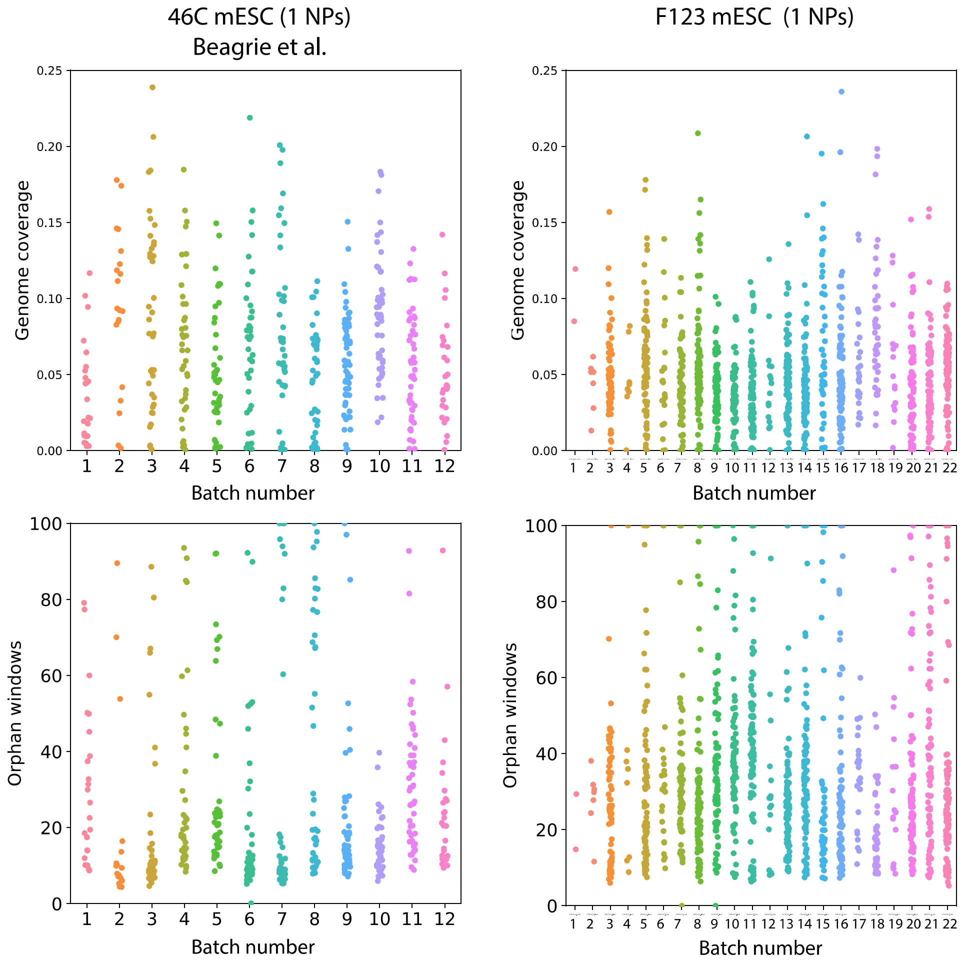


Figure 3. 19: Nuclear profiles show consistent quality metrics between batches.
The percentage genome coverage (top) and the percentage of orphan windows (bottom) is shown for each batch of NPs, in chronological order of collection, for 471 single NPs from 46C, and 1281 single NPs from F123.

3.5.2 Defining quality control criteria for the GAM datasets from F123 mESCs

The parameter that displayed the most striking difference between 46C and F123 data was the percentage of orphan windows. We considered to use it as an additional parameter to %

mapped reads, to help identify lower quality GAM samples in the F123 dataset. Noting that GAM samples from 46C rarely had more than 60 % of orphan windows, and the majority of F123 NPs was also below this mark, while negative controls showed the opposite behaviour, the value of 60 % orphan windows was chosen as an appropriate criterium (Figure 3.20). Twelve additional samples with very low number of uniquely mapped reads (<20,000) were also excluded as they may correspond to samples that were under-sequenced or were contaminated by sequences from other genomes. In total, considering both the % orphan windows and the number of uniquely mapped reads, led to the exclusion of 145 NPs from the dataset, leaving 1123 good quality NPs for generating segregation tables from (Table 3.3, Figure 3.20). This cut-off removes 11% of lower quality samples, which is an improvement comparing with 63/471 NPs in the first GAM dataset (Beagrie et al. 2017) (Table 3.3, Figure 3.20).

Table 3. 3: Number of NPs in F123 and 46C 1NP datasets before and after quality control.

F123 1NP dataset	**Number of NPs**
collected	1281
cross-contaminated	13
> 60 % orphan windows	133
< 20 000 uniquely mapped reads	12
passed quality control	1123
46C 1NP dataset (Beagrie et al.)	**Number of NPs**
collected	471
< 15 % mapped reads	63
passed quality control	408

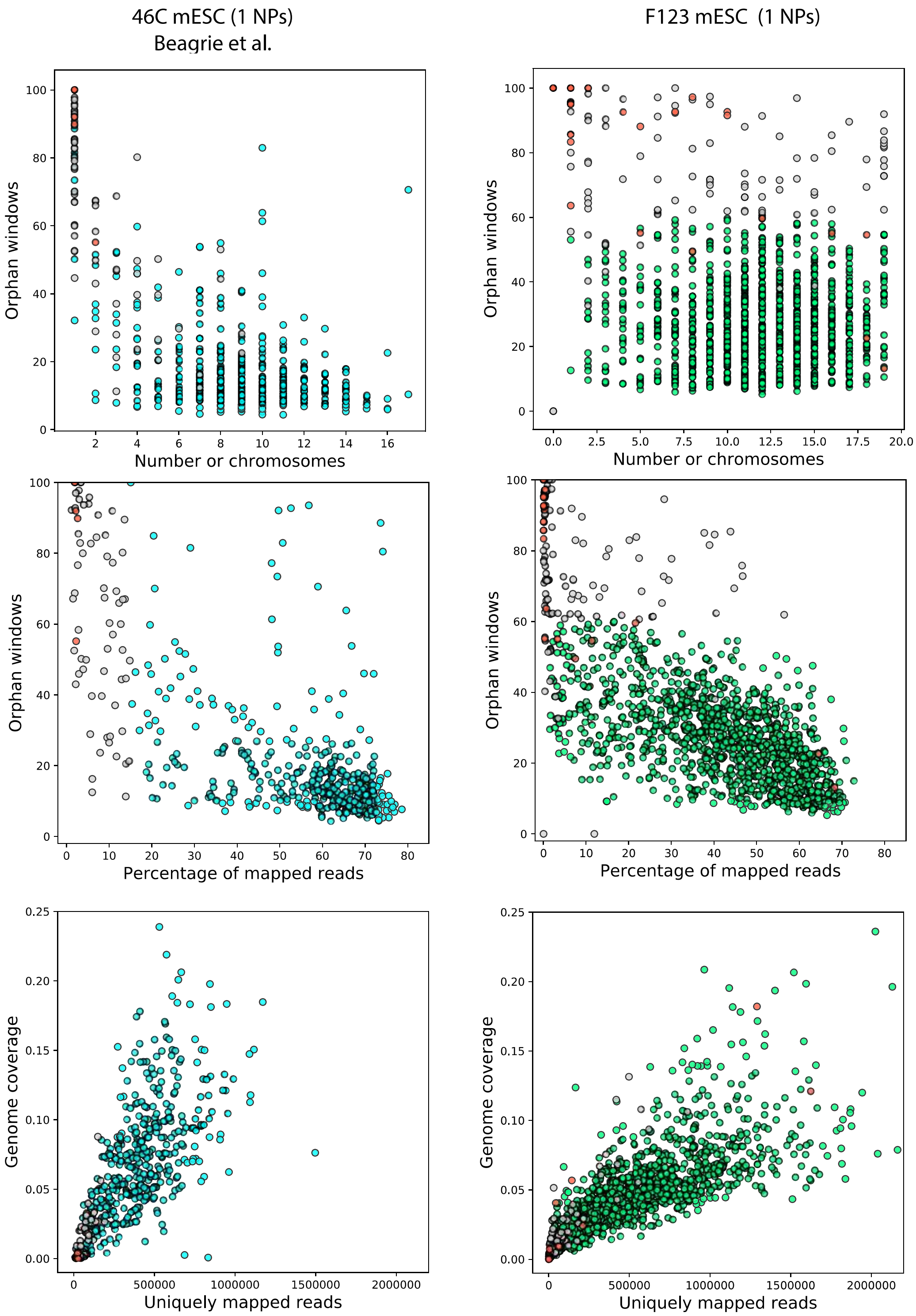


Figure 3. 20: Different quality control criteria are applied on 46C and F123 to identify low quality NPs. In 46C (blue), samples with less than 15 % mapped reads (grey) are removed from the dataset. In F123 (green), samples with less than 20,000 uniquely mapped reads and more than 60 % orphan windows are removed from the dataset. Negative controls obtained from laser microdissection of membrane not containing NPs are shown in red.

3.5.3 Differences in detection efficiency between the two GAM pipelines

A GAM dataset consists of a number of NPs, each with different parts of the genome present. The information obtained from each NP can be summarised in a segregation table, which shows in which NPs any genomic locus was detected (at a given resolution). With the knowledge of a few parameters (the nuclear radius of the cell, the slice thickness, and the genome length) the expected number of NPs that contain a single genomic locus can be calculated (Beagrie et al., 2017). This expected locus detection frequency can be compared with the actual segregation frequency of a locus, which gives an estimate of the efficiency of DNA detection. The efficiency in GAM describes the probability of a locus being detected and can inform about the extent of false negative loci that were present in a NP but not extracted or detected (Figure 3.21b). In 46C, the median detection frequency of a genomic locus at 30 kb resolution in the collection of 408 NPs is 6.2 % (Figure 3.21a), which results in a detection efficiency of 83 % (Beagrie et al., 2017). We suspected that the reduced genome coverage observed in F123 NPs would impact on the detection frequency of genomic loci and thereby also the overall detection efficiency. We find that the median detection frequency in F123 at 30 kb is 4.2 % (Figure 3.21a), and the detection efficiency is 55 % at the same resolution (Figure 3.21c, analysis by Luca Fiorillo, laboratory of M. Nicodemi, University of Naples).

The successful extraction of all genomic material from a nuclear profile is a crucial parameter for the quality of GAM datasets and the resolution that they can be studied at. To investigate the consequences of this reduced efficiency in F123, we explored the consequences of different detection efficiencies by modelling GAM in silico (analysis by Luca Fiorillo and Francesco Musella, Nicodemi lab, University of Naples). With the model, the theoretical number of NPs that is required to obtain saturated contact maps at a given resolution (40kb) was determined. As a metric for saturation of contacts, the model considers convergence of replicates (that is 0.9 Pearson correlation between replicates). At 100 % efficiency, a GAM dataset needs a theoretical minimum of 1000 NPs to achieve convergence (Figure 3.21d). Importantly for our subsequent analyses, this number remains stable until 40 % efficiency showing that GAM data has a good tolerance for efficiency variations. Below 40 %, efficiencies become crucial, as the number of NPs that are required for convergence between replicates increases to impractical numbers of samples of ten thousand or more.

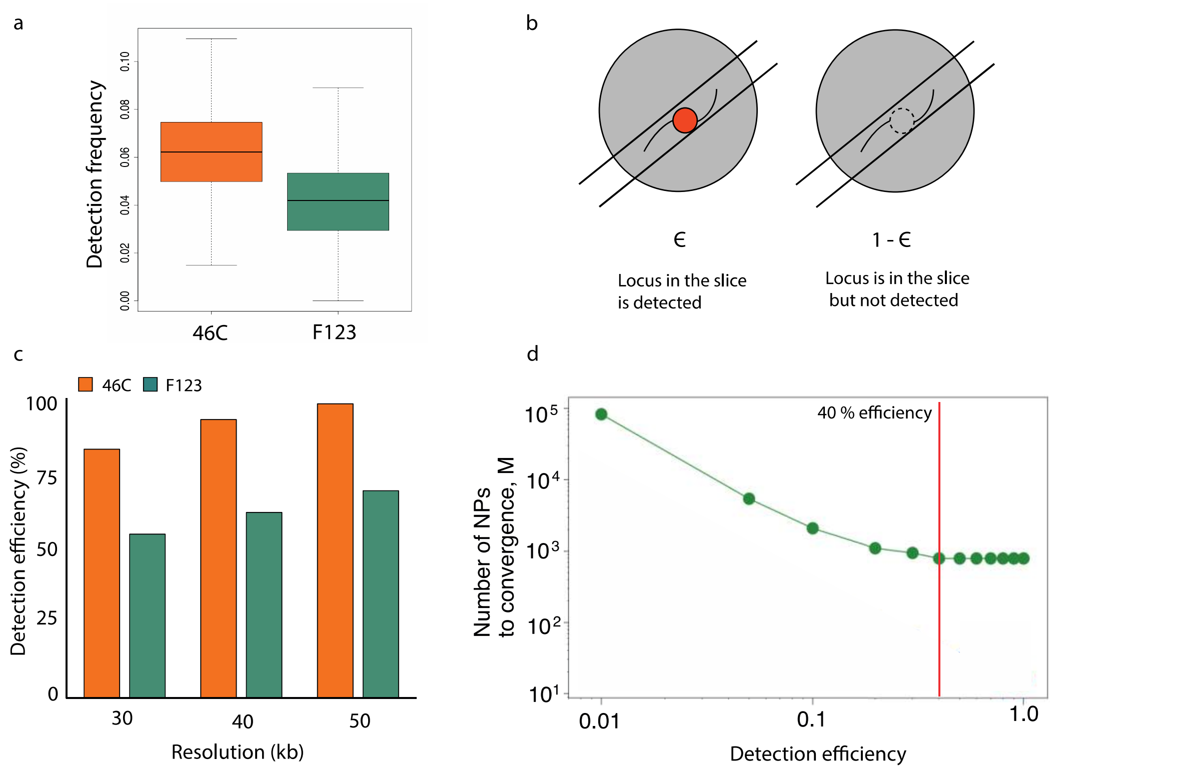


Figure 3. 21: Detection efficiency of genomic loci is reduced in the F123 dataset compared to 46C. (a) Mean detection frequencies for all genomic windows at 30 kb in 46C and F123. (b) Illustration of the effect of DNA extraction efficiency (ε). (c) Genome-wide detection efficiency at different resolutions in 46C and F123. (d) The impact of efficiency on capturing chromatin contacts with GAM. Number of NPs required to achieve convergence of replicates in a theoretical GAM experiment with different efficiencies. Material in c and d by L. Fiorillo.

3.5.4 Extraction efficiencies in F123 are not uniform throughout the genome

Next, we asked whether the difference in extraction efficiency in F123 had a preference for specific regions of the genome namely between open (eu-) chromatin and closed (hetero-) chromatin. First, we compared the detection frequencies per genomic window and found that window detection frequencies do not correlate well between in 46C and F123 (Figure 3.22). For some genomic windows, the lower detection frequency in F123 is apparent, while for others this difference is less pronounced.

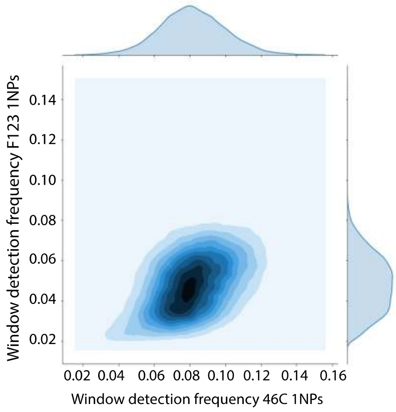


Figure 3. 22: Genomic windows are differentially detected in the 46C and F123 datasets.
Density plot of window detection frequencies in the 46C and the F123 dataset. Histograms of window detection are shown above (46C) and on the right (F123), and show a wider distribution in F123 consistent with two or more behaviours in window detection.

Next, we tested whether the under-detected windows in F123 share common features that could explain why we see them less often than expected. We reasoned that the feature of DNA to most likely interfere with extraction, is the compaction of chromatin. More condensed chromatin with a high protein occupancy might suffer stronger from reduced efficiencies. To test this hypothesis, we first used published ATAC-seq data for mESCs (E14) to identify euchromatic DNA regions, by using the density of ATAC peaks as a proxy for the openness of chromatin. We grouped genomic windows (50 kb) based on the number of ATAC peaks per window, and compared the window detection frequencies of each group (Figure 3.23a). In both, 46C and F123, the detection frequency is higher in windows with higher ATAC peak density, as expected, since the decondensed state of open chromatin takes up larger areas of the nucleus and is sampled slightly more frequently (Beagrie et al., 2017). In the F123 dataset, the difference in detection frequency between ATAC-rich and ATAC-poor genomic regions is very pronounced (Figure 3.23b). In 46C, windows across all detection frequencies can have high numbers of ATAC peaks, which results in a more uniform detection of windows with and without ATAC peaks. In F123, the majority of windows with high numbers of ATAC peaks have higher detection frequencies than windows with low numbers of ATAC peaks, which results in the under-detection of closed chromatin in the F123 dataset.

While ATAC-peak occupancy can function as a proxy for euchromatin, the absence of ATAC-seq signal does not necessarily identify a genomic region as heterochromatic. Heterochromatin contains condensed DNA, which can be found predominantly at the periphery of the nucleus, in proximity to the nuclear lamina (Peric-Hupkes et al., 2010). Thus, to mark heterochromatin, we classified genomic windows according to their membership to LADs using published DamID data in mESCs for Lamin B1, which identifies DNA interacting with Lamin B1. In 46C, windows covering LADs have a similar range of detection frequencies as windows outside of LADs (Figure 3.23c). In F123, windows covering LADs have lower detection frequencies than windows outside of LADs (Figure 3.23c). The difference in LAD detection can be seen in more detail when comparing detection frequencies for each genomic window between the datasets, and colouring them based on their location in or outside LADs (Figure 3.23d). To explore this trend, we compared the average nucleotide coverage and the average number of reads per 50 kb window in LADs and non-LADs in 46C and F123 GAM datasets. While windows in 46C have similar reads and coverage between LADs and non-LADs (Figure 3.23e), windows in F123 show a clear difference, displaying lower nucleotide coverage and less reads in LAD windows (Figure 3.23f). These results point towards a general problem of DNA extraction in LADs.

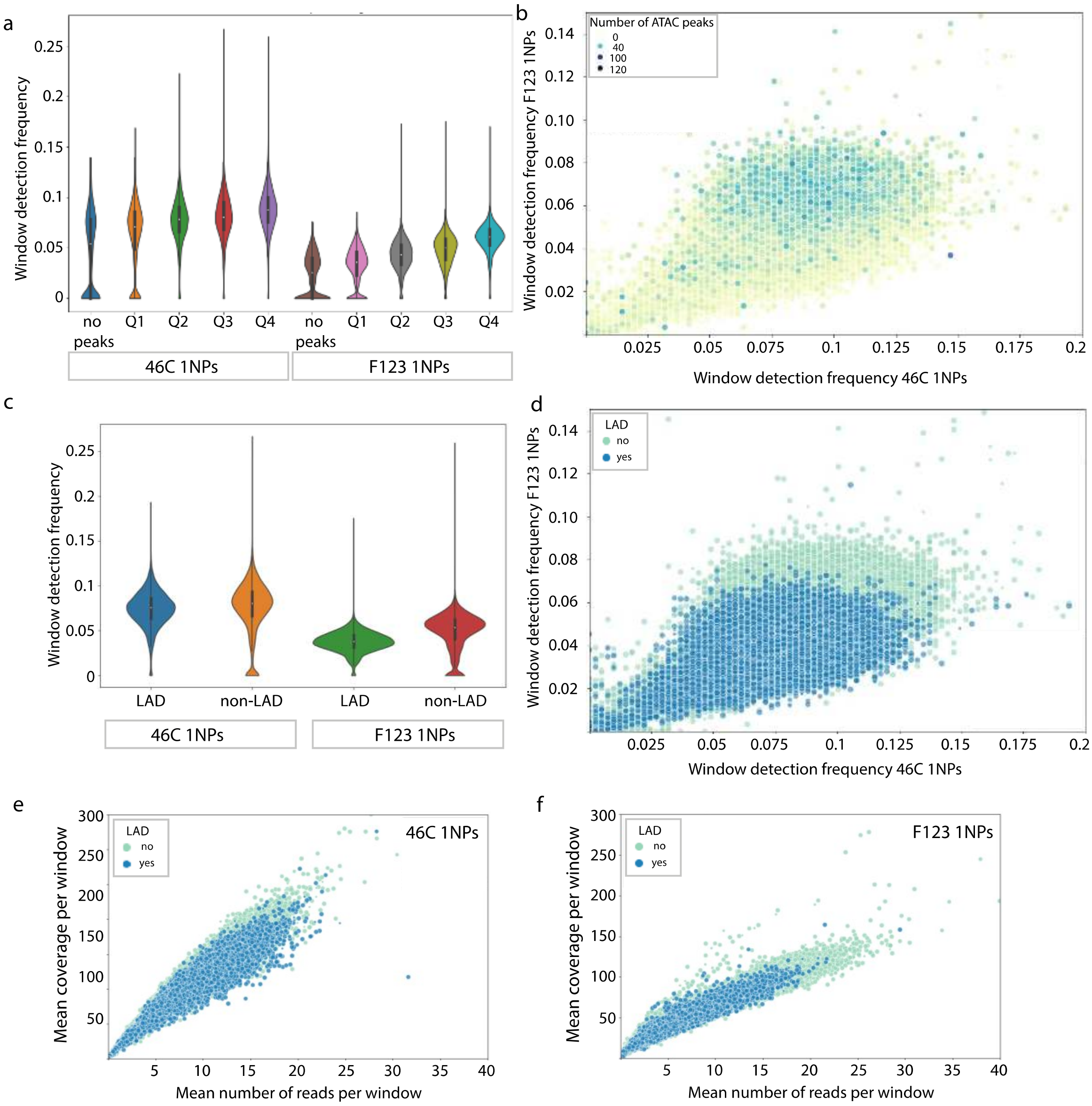


Figure 3. 23: Window detection frequencies in the F123 dataset are not uniform, with reduced frequencies in lamina associated domains (LADs).

(a) Detection frequencies in 46C and F123 of genomic windows (50 kb), grouped based on the number of ATAC peaks per window. ATAC-seq data from mESCs (E14). (b)Window detection frequencies of 46C and F123 coloured by the number of ATAC-seq peaks per window (50 kb). (c) Detection frequencies of 50 kb windows, grouped according to their overlap with a LAD (yes / no) in 46C and F123. (d) Detection frequencies of 50 kb windows coloured according to their overlap with a LAD (yes = blue, no = green) in 46C and F123. LAD windows are plotted on top. (e) Mean coverage and mean number of reads per window in 46C coloured according to lamina-association. LADs (blue) are plotted on top of non-LADs (green). (f) Mean coverage and mean number of reads per window in F123 coloured according to lamina-association. LADs (blue) are plotted on top of non-LADs (green).

3.5.5 Exploring biases in the F123 data and their origin

The previous observation that LADs are detected less often than non-LADs in the F123 dataset leads to two main questions. First, what is the cause of this bias in the F123 dataset? Second, can it be corrected experimentally and/or computationally? To address the first question, we asked whether the LAD under-sampling is a feature of the sum of all NPs or if this phenomenon occurs already in single NPs. The first scenario would point towards a biased collection of NPs, presumably a preferential collection of larger, more equatorial slices of the nucleus, while missing small apical slices that are enriched for lamina and the associated large proportions of heterochromatin (Figure 3.24a, b). The second scenario would point towards problems in extracting DNA from heterochromatic regions, independent of other features of the NP. To distinguish these scenarios, we calculated the ratio of non-LAD to LAD windows (at 50 kb) that are present in the mESC linear genome (ratio: 1.53), and compared this number with the ratio of non-LAD to LAD windows in every GAM sample from single NPs. In 46C, NPs distribute symmetrically around the genome-wide ratio (Figure 3.24c). In samples with lower genome coverage, NPs tend to cover either LADs or non-LADs, which appears natural for NPs with less content of the nucleus present. With increasing genome coverage, the ratio of LADs to non-LADs becomes more balanced and, in many NPs, close to the genome-wide ratio. In F123, the distribution of NPs follows the same trend as in 46C (Figure 3.24d), indicating that the F123 dataset contains equatorial as well as apical slices, which excludes a biased collection towards more equatorial NPs. However, the ratio of non-LADs to LADs is increased in almost every nuclear profile (Figure 3.24d), suggesting insufficient extraction of DNA in LADs.

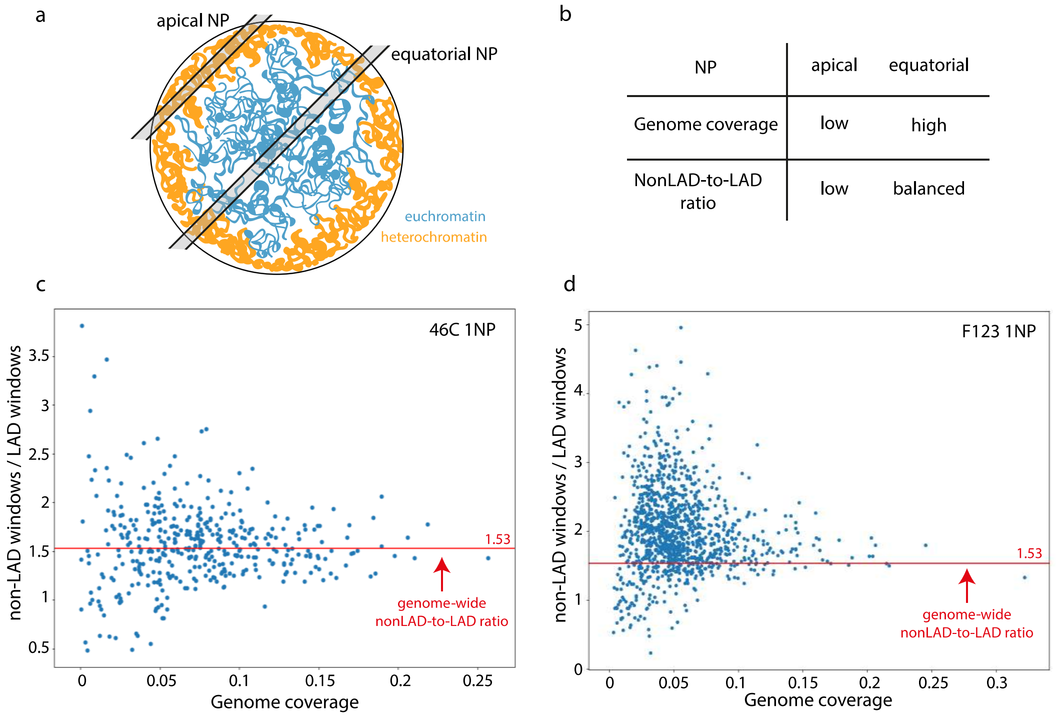


Figure 3. 24:The majority of all F123 nuclear profiles contains fewer positive windows overlapping with LADs.

(a) Illustration of the genomic content of apical and equatorial NPs. (b) Expected genome coverage and proportion of LADs in apical and equatorial NPs. (c) Ratio of non-LADs to LADs vs. genome coverage in single NPs of the 46C dataset. (d) Ratio of non-LADs to LADs vs. genome coverage in single NPs of the F123 dataset.

The lower detection of LADs raised the possibility that the lower extraction could result from the peripheral position of the DNA from LADs which is closer to the laser beam that microdissects each NP from the cryosection. In this scenario, DNA from LADs could potentially be exposed to the laser beam. To test this hypothesis, we made use of the GAM datasets of 3NPs that were also collected in F123, but not presented here so far. Here, 4 cellular slices are microdissected together (Figure 3.25a), corresponding to 3 NPs since approximately 25% of cellular slices in ESCs do not contain nuclear material (details see Figure 3.3d, e). However, when comparing the detection frequencies in LADs and non-LADs, the multiplexed version of GAM clearly displays the same bias as the 1NP dataset (Figure 3.25b), showing that destruction of DNA during laser microdissection is an unlikely cause for the under-detection of LAD/heterochromatic DNA in the F123 datasets. This result is further confirmed at the level of nucleotide coverage and reads per windows, which are clearly reduced in LAD windows compared to non-LAD windows in both the 1NP and 3NP datasets (Figure 3.25c).

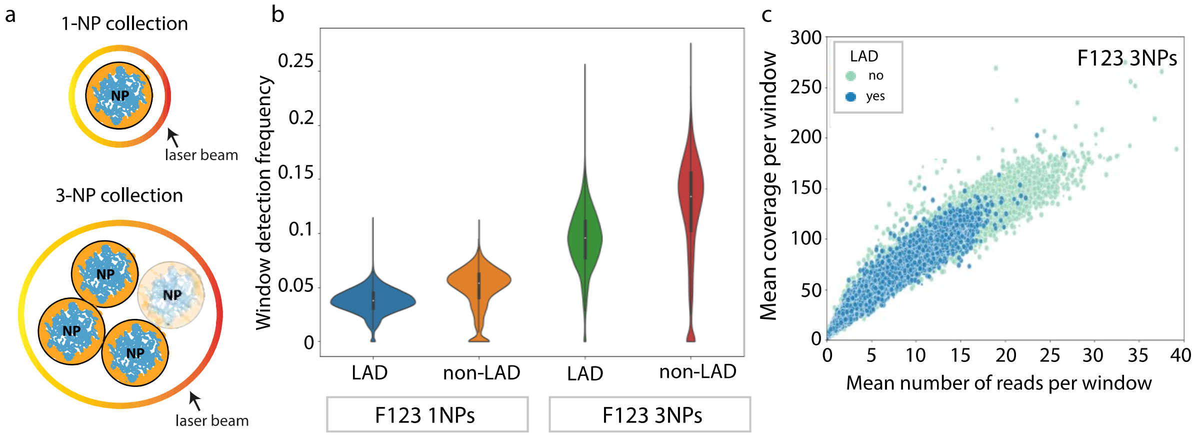


Figure 3. 25: LAD detection bias is present in the 3-NP F123 dataset as well.
(a) Illustration of laser microdissection of single NPs and multiplexed NPs. (b) Detection frequencies of genomic windows (50 kb), grouped according to lamina association in the 1NP and the 3NP F123 dataset. (c) Mean coverage and mean number of reads per window in 3NPs coloured according to lamina-association. LADs (blue) are plotted on top of non-LADs (green).

These observations lead to the conclusion that DNA is less well extracted from LADs because of the heterochromatic nature of LADs. The most crucial part for DNA extraction in the GAM protocol is the WGA, specifically the first reaction, where crosslinked chromatin is treated with a cell lysis buffer and protease for protein removal, so that DNA is accessible for primers and polymerase in the following DNA amplification. One important experimental change in the GAM protocol, is the change from the WGA-4 kit from Sigma to the in-house WGA protocol. The major difference between the Sigma WGA-4 kit and the in-house WGA is how DNA is extracted from nuclear profiles. The Sigma WGA uses a so-called lysis and fragmentation buffer with an unknown component that fragments the DNA in a cell lysis buffer, presumably with a high concentration of detergents that clear protein of the DNA. This step is essential for the subsequent DNA amplification, as DNA needs to be accessible for random oligos to anneal and for the polymerase to read and extend the DNA. In the in-house protocol, cell lysis is achieved by a cell lysis buffer with high concentrations of detergents, but without DNA fragmentation, as we are not aware of a publicly available DNA fragmentation enzyme that can work in the buffer conditions that we found ideal for cell lysis. Looking at the differences in detection frequencies between the 46C and F123 dataset, it seems probable that protein removal from heterochromatin might have been less efficient without simultaneous fragmentation, whilst open chromatin regions suffered less due to lower protein occupancy in euchromatin than in heterochromatin. The additional fragmentation step in the original Sigma WGA4 kit might be able to compensate for these differences in protein occupancy, explaining why the 46C dataset has a more uniform window detection frequency than the F123 dataset.

3.5.6 Impact of LAD under-sampling on locus co-segregation in GAM data

To investigate how the differences in detection frequency between genomic regions impacts on the co-segregation of windows in pairwise matrices, I plotted the co-segregation matrices and window detection frequencies of several genomic regions in the F123 and 46C datasets, and compared them with the genomic positions of LADs, and gene expression counts from total RNA-seq (Figure 3.26). This exploration aimed at better understanding the location of the over- and under-detected loci, but did not intended to study chromatin contacts, which is done in matrices that are appropriately normalised (Beagrie et al. 2017). In 46C, co-segregation frequencies are fairly uniform across entire chromosomes (Figure 3.26a), and large genomic stretches of tens of megabases (Figure 3.26b). Along the chromosome, regions of higher or lower co-segregation frequencies can be observed outside and inside of LADs. In contrast, co-segregation frequencies in F123 are very dependent on window detection frequency and clearly reduced inside LADs (Figure 3.26a-b). To further explore the effect of variable efficiency of window detection, I investigated two shorter regions around the *Sox2* and *Myc* loci. The *Sox2* region contains one small LAD within the surrounding 8 Mb genomic region, resulting in high co-segregation frequencies along the entire locus in both 46C and F123 (Figure 3.26c). At the *Myc* locus, a 4 Mb region with the highly expressed *Myc* gene in the middle of two large LADs, co-segregation frequencies are uniform in 46C, but centred entirely at the few transcribed regions at the locus in F123 (Figure 3.26d). These observations confirmed the variable efficiency of window detection in the F123 data. In the last sections of this chapter, I present our brief efforts to improve our experimental pipeline further to obtain excellent extraction efficiency, and how the variable efficiency was taken into account through normalisation of the pair-wise matrices. Importantly, in Beagrie et al. (2017) a mathematical model, SLICE, was used to detect the most specific contacts at a given distance of interaction, from co-segregation matrices. Although the original SLICE method already considered imperfect efficiency of detection, its current implementation considered only a constant efficiency of detection, and for that reason we have not so far determined the most prominent GAM contacts using SLICE. The Nicodemi lab is currently working on a revised version of SLICE that will consider window-based efficiency, which is especially appropriate for larger datasets, where variability in locus detection is less likely to be due to insufficient sampling and more likely to result from true efficiency of detection.

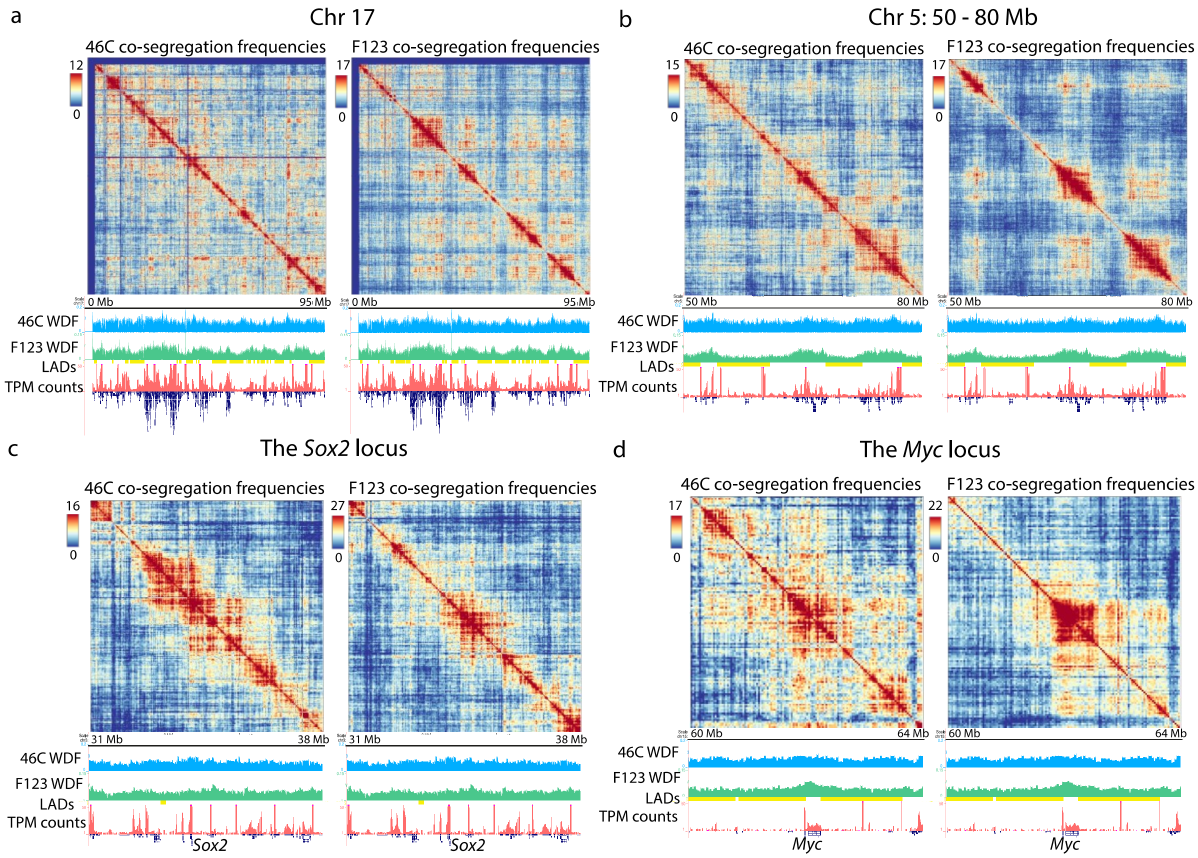


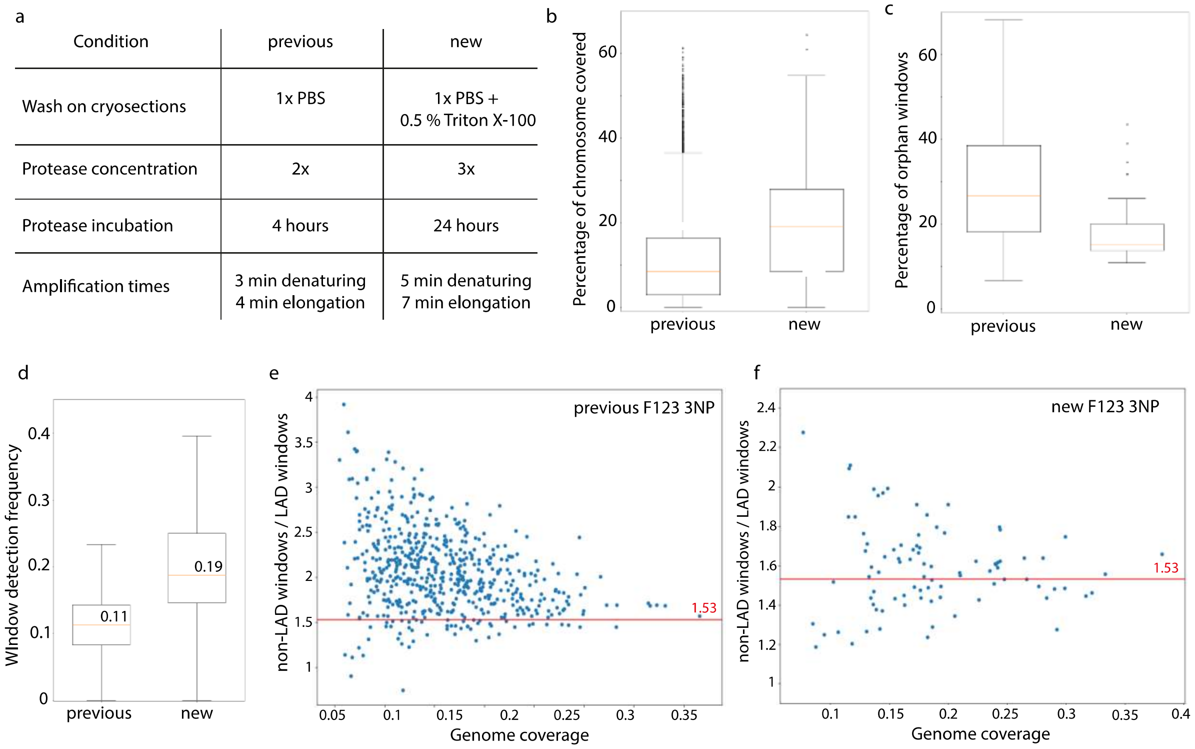
Figure 3. 26: Low detection frequencies in LADs are noticeable as depleted regions in the raw co-segregation frequencies in the F123 dataset but not in the 46C dataset.

(a – d) From top to bottom: Co-segregation frequencies (30 kb) in 46C (left) and F123 (right), window detection frequencies in 46C (30 kb), window detection frequencies in F123 (30 kb), Lamina associated domains (LADs) in mESCs, total RNA-seq TPM values in F123 (biological replicate 1), and below UCSC gene annotation track. All data in mm10. (a) Chromosome 17. (b) Chromosome 5: 50 – 80 Mb. (c) Chromosome 3: 30 – 38 Mb. (d) Chromosome 15: 60 – 64 Mb.

3.5.7 Correcting for biases in DNA extraction efficiencies

(a) Experimental correction

With the problem of LAD under-sampling in mind we decided that our in-house WGA, which can indeed extract DNA from euchromatin with good efficiencies, needed further optimisation to extract condensed heterochromatin more effectively. Thus, we decided to improve the performance of the WGA specifically at the first steps of the WGA reaction, when DNA extraction occurs. I revisited our previous optimisations and found a few parameters in the WGA protocol that could be tested, including (a) increasing the concentration of protease, (b) prolonging the incubation times for cell lysis, (c) increasing the denaturation and elongation times in the DNA amplification, and (d) treating cryosections with a detergent prior to the incubation with protease (Figure 3.27a). We found that the combination of all these protocol changes yields GAM samples with improved quality criteria, leading to increased genome coverage per NP (Figure 3.27b), fewer orphan windows (Figure 3.27c), and higher window detection frequencies (Figure 3.27d). Next, we compared the ratio of non-LAD to LAD windows, and found it to be balanced in the optimised samples (Figure 3.27f; equal to the LAD ratio of 46C) compared with the F123-1123 dataset (Figure 3.27e). This result shows that indeed the DNA extraction and amplification during the WGA are crucial elements in the GAM protocol that determines DNA extraction efficiency, especially from heterochromatic regions, and that optimisations now resulted in a highly promising in-house GAM protocol suitable for collecting future GAM datasets without problems in extracting DNA from heterochromatin.



Condition	previous	new
Wash on cryosections	1x PBS	1x PBS + 0.5 % Triton X-100
Protease concentration	2x	3x
Protease incubation	4 hours	24 hours
Amplification times	3 min denaturing 4 min elongation	5 min denaturing 7 min elongation

Figure 3. 27:Targeted optimisation improve the GAM protocol further and correct for LAD under-sampling.

(a) Overview of new experimental conditions. (b) Percentage of chromosome coverage per NP in the previous dataset and the new experimental conditions. Previous = 565 x 3NPs F123, new = 48 x 3NPs F123. (c) Percentage of orphan windows per NP in the previous dataset and the new experimental conditions. Previous = 565 x 3NPs F123, new = 48 x 3NPs F123. (d) Window detection frequency in F123 3NPs with the previous and the new experimental conditions. Previous = 565 x 3NPs F123, new = 48 x 3NPs F123. (e) Ratio of non-LADs to LADs and genome coverage in multiplexed NPs in the previous dataset (565 x 3NPs). (f) Ratio of non-LADs to LADs and genome coverage in multiplexed NPs with the new experimental conditions (89 x 3NPs).

(b) Computational correction

Although we were able to improve the GAM protocol further so that future datasets will be collected without the bias in detection frequencies between LADs and non-LADs, we aimed to improve the computational analyses of the F123-1123 dataset as much as possible to improve the detection of heterochromatic windows, which might have lower read coverage than expected and might be lost in our window calling approach. In Beagrie et al. (2017), GAM data was normalised using D' (normalised linkage disequilibrium; linkage is the co-segregation of two loci minus the product of their detection frequencies). Previous explorations by Thomas Sparks had shown that normalised point mutual information (NPMI) gives improved quality of normalisation.

In order to explore the NPMI normalisation, I compared the raw co-segregation frequencies and the normalised data (NPMI) side-by-side, and found that it effectively removes biased detection of co-segregation between active genomic windows (Figure 3.28). Applying NPMI generates contact maps with smooth contact decay with increasing genomic distance for LADs and non-LADs.

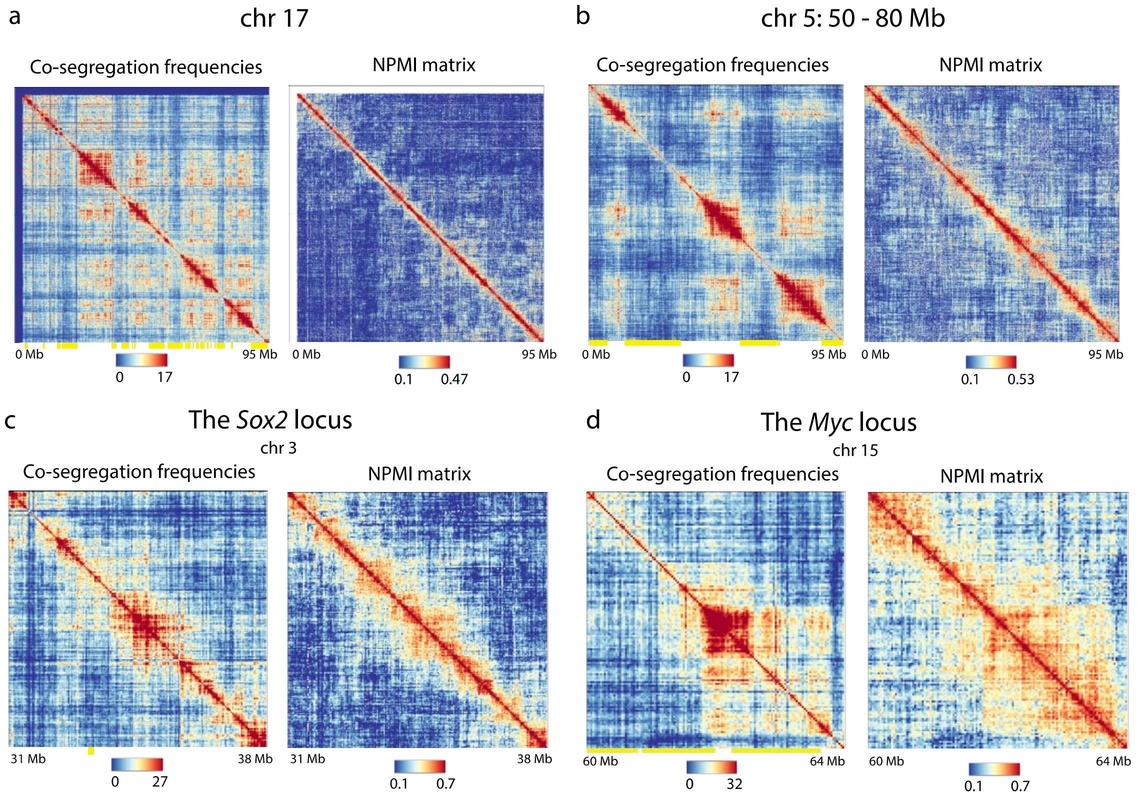


Figure 3. 28: LAD under-sampling in F123 can be normalised effectively at many genomic loci, even within regions containing large proportions of LADs.
(a – d) Raw (left) and normalised (right) co-segregation frequencies (F123). Genomic positions of mESC LADs (yellow) are shown underneath the raw co-segregation matrix. (a) Chromosome 17. (b) Chromosome 5: 50 – 80 Mb. (c) Chromosome 3: 30 – 38 Mb. (d) Chromosome 15: 60 – 64 Mb

To quantify the effect of NPMI normalisation on potential biases of the final matrices, we investigated three potential biases as previously shown in Beagrie at al. (2017), namely window detection frequency, GC content, and read mappability. We found that D' normalisation of the published ESC-46C and NPMI normalisation of the ESC-F123 data both remove these important biases efficiently (Figure 3.29a, b).

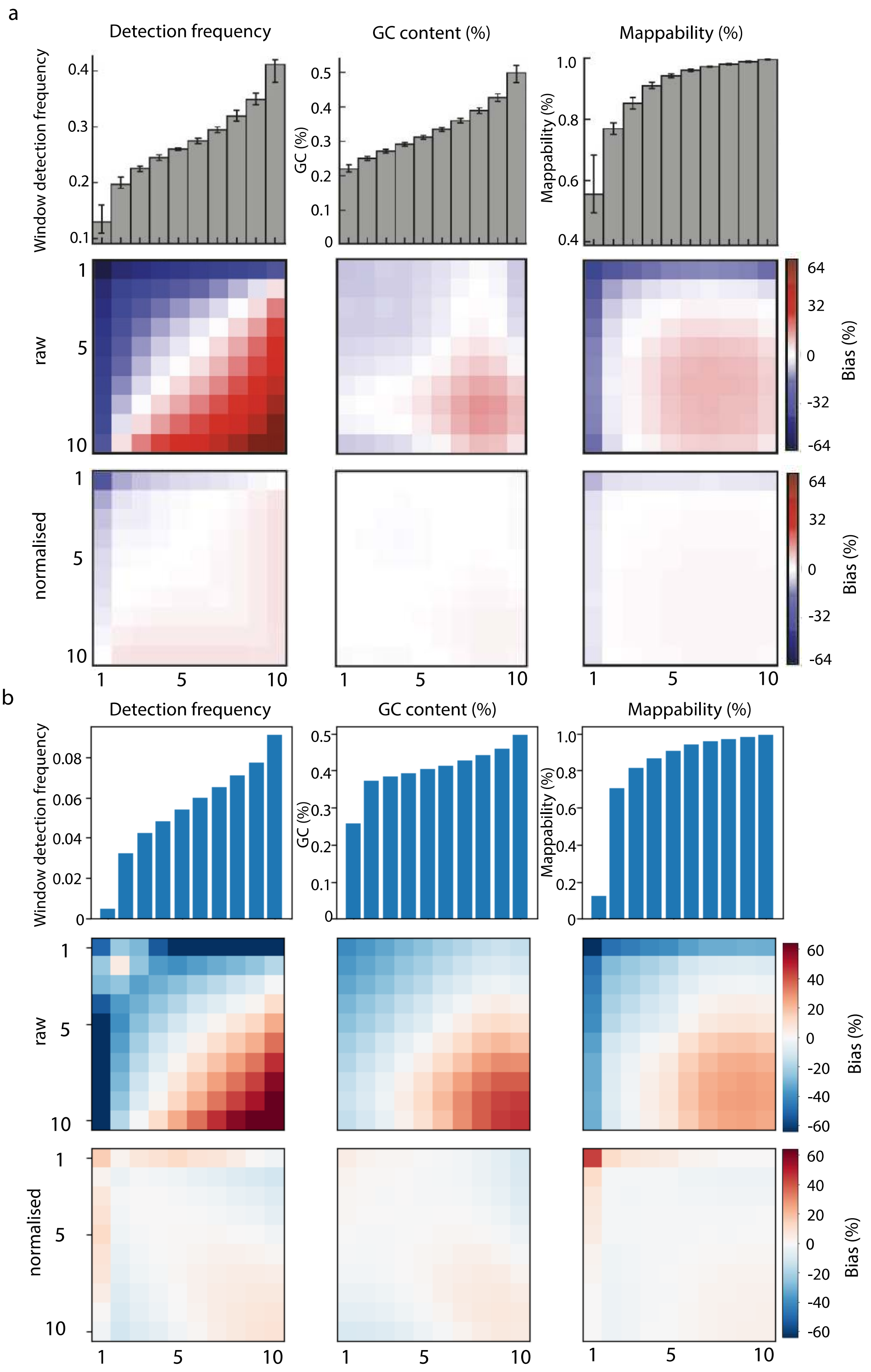


Figure 3. 29: Normalised NPMI effectively reduces bias in GAM datasets.
50 kb windows were divided into equal groups according to their detection frequency, GC content or mappability (left, middle and right respectively), bar plots give the mean. Mean observed over expected values (% bias) between windows in each group are shown for raw GAM co-segregation (middle row) and normalised data (bottom row). (a) Normalisation of 46C GAM data (D' – normalised linkage disequilibrium), shown from (Beagrie et al., 2017). (b) Normalisation of F123 GAM data (NPMI – normalised point mutual information).

Finally, to disentangle biological differences between the ESC lines 46C and F123, which have been grown in different culture conditions, from technical differences between the GAM datasets, I compared the normalised GAM matrices from F123 and 46C, with Hi-C data produced in 46C (described in chapter 5) and F123 (Kubo et al., 2017). A remarkable resemblance of chromatin structures is found at randomly chosen genomic regions both at the level of TADs, and at smaller contact domains, between all four datasets (Figure 3.30). However, in the 46C GAM dataset some additional contacts are present between domains or TADs, that cannot be observed in the corresponding region in the Hi-C datasets, and only to some extend in the F123 data (e.g. Figure 3.30, top row). These contacts might reflect biological differences in gene expression in mESC 46C and F123, which are different ESC lines grown in different conditions (feeder-free and on feeders respectively; different media compositions in Table 2.3).

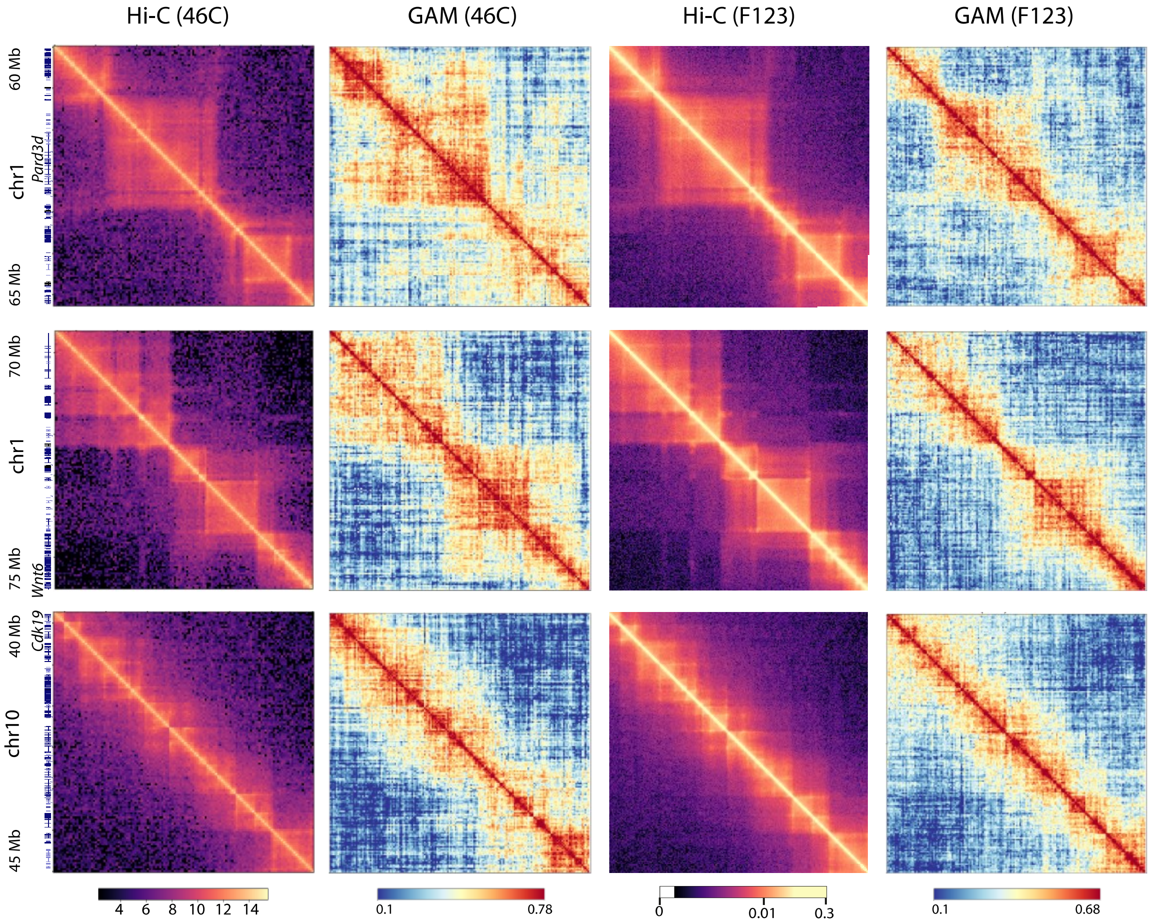


Figure 3. 30: Chromatin contact maps show high resemblance between GAM and Hi-C data from both 46C and F123.

From left to right: USCS gene annotation tracks (mm10), Hi-C interaction frequencies in 46C (30 kb), GAM normalised co-segregation frequencies in 46C (30 kb), Hi-C interaction frequencies in F123 (25 kb), GAM normalised co-segregation frequencies in F123 (30 kb). From top to bottom the following genomic regions are displayed: Chromosome 1: 60 – 65 Mb, Chromosome 1: 70 – 75 Mb, Chromosome 10: 40 – 45 Mb.

3.6 Conclusions

In this chapter, I showed extensive work towards the development of an in-house, affordable GAM protocol that significantly increases the throughput of the GAM experimental pipeline, thereby decreasing the costs and time required to generate GAM datasets. While working on improvements of the GAM protocol, the commercial WGA method from Sigma (WGA4) that had been used until then to extract DNA from NPs, suddenly stopped extracting good-quality DNA from GAM nuclear slices. This unplanned challenge led us to explore other commercially available WGA kits, and ultimate led to the development of an independent WGA protocol, inspired by protocols for single cell genome and transcriptome amplification (Chapman et al., 2015; Zong et al., 2012). I showed here that the newly developed WGA method can robustly extract DNA from nuclear profiles, and thereby allows collection of GAM datasets with an immense cost reduction and without having to rely on the reproducibility of commercial WGA products. We optimised the in-house WGA protocol further, and ultimately generated 1NP and 3NP datasets in mESC (F123) that resulted in a large dataset of a total of 3664 NPs, from two biological replicates.

However, the version of the WGA protocol that was applied to collect these large F123 datasets, turned out to have some drawbacks, which were discovered only after completion of data collection, while analysing the complete F123 dataset. A detailed comparison of various quality metrics between the published 46C dataset and the new F123 dataset, allowed me to reassess all previous optimisations, and find four tested conditions that on their own had only marginally improved the performance of the GAM protocol, but when applied together led to a vast improvement of all quality metrics. These developments complete the efforts to devise an independent, and affordable GAM pipeline which has also been tested in other types of samples (e.g. neurons) and shown to work reliably (Izabela Harabula, unpublished results).

The investigations of the lower detection efficiency found in the F123 dataset compared to 46C, also helped us devise new criteria for quality control of GAM data, and to learn which are most crucial and can be used to test GAM data during collection. This new knowledge will help the consistent production of high-quality GAM datasets in the future. During the entire F123 data collection, we had monitored the percentage of mapped reads and the genome coverage. However, we were unaware of the importance of the quality metric ‘orphan windows’, which had also been tested in Beagrie et al. 2017, but not shown to distinguish lower quality data. Orphan windows represent the frequency of positive windows without a neighbouring positive window, which we found later to be of utmost importance for assessing the quality of single GAM samples. While GAM samples with low percentages of mapped reads or a low genome coverage can still be of good quality, high percentage of orphan windows can only be found in low quality NPs. Thus, the percentage of orphan windows should be monitored throughout the data collection, and compared between batches, as sudden increases in orphan windows are a sign of experimental difficulties, such as reduced DNA extraction efficiencies from NPs during WGA.

Further, we also gained insight into the nature of GAM datasets and how best to assess their quality and ameliorate possible biases, through the comparison of two 1NP datasets from mESCs (46C and F123). First, we compared the quality of NPs produced with the published GAM protocol and NPs generated with the new GAM pipeline. We discovered that reduced DNA extraction occurs predominately in genomic regions associated with the nuclear lamina, suggesting that the heterochromatic nature of these regions was likely to cause this reduced extraction efficiency, rather than the more unlikely scenario of biased collection of equatorial sections that have a lower proportion of heterochromatic DNA. This knowledge allowed us to identify the cause for the reduced efficiency in the WGA reaction, and led us to re-develop new experimental conditions that rescued the DNA extraction efficiency in heterochromatic regions to excellent levels. The latest optimised conditions give a GAM protocol that can be used in future experiments, and will allow the experimentalist to produce GAM datasets with homogeneous DNA extraction efficiencies and in a high throughput, and cost reduced manner.

Ultimately, chromatin contact maps from the new mESC GAM dataset, produced here from the hybrid ESC line F123, largely resembles the published GAM dataset from 46C, and also shows remarkable similarity to published Hi-C data from the same mESC line. Taken together, these results show that our revised GAM pipeline and quality control metrics generate meaningful chromatin contact information.

4. Chromosome topology of the parental alleles

Note on results and author contributions

The experimental work presented in this chapter was done by myself. Dr. Alexander Kukalev and Gesa Loof helped collect nuclear profiles. The analysis presented here was done in collaboration with several members of our laboratory. A. Kukalev generated segregation tables from GAM sequencing data, developed and applied haplotype phasing to nuclear profiles. Dr. Ibai Irastorza Azcarate analysed A/B compartments, and produced the materials for Figure 4.11, and 4.3d. Dr. Christoph Thieme identified differential chromatin contacts between datasets. Julia Markowski (Roland Schwarz laboratory, Max Delbrück Centrum, Berlin) analysed RNA-seq data and identified genes with allele-specific expression. Yingnan Zhang (Lonnie Welch laboratory, Ohio University, US) performed transcription factor enrichment analysis on differential chromatin contacts.

4.1 Objective and research motivation

Monoallelic gene expression is a common feature of genes throughout the genome. Despite its broad occurrence, its regulation and function are largely unknown. Some studies have suggested chromatin folding to be involved in the regulation of alleles at some loci (Dixon et al., 2015; Holwerda et al., 2013; Rao et al., 2014; Tan et al., 2018a). Most genome-wide studies of chromatin folding have not been sensitive enough to detect differences between parental alleles, often due to highly similar genome copies in inbreed mouse lines, or unknown parental genome information in humans. To assign data to the haplotypes, hybrid mouse lines that have genomes with a high density of single nucleotide polymorphisms (SNPs) can be used to distinguish maternal from paternal alleles (Giorgetti et al., 2016; Rivera-Mulia et al., 2018). Importantly, the efficiency of phasing sequencing data depends strongly on the SNP density of the organism’s genome, but also on the length of the sequencing read, and how allele-specific reads are distributed in the genome, which can result in rather low percentages of reads that can be assigned to a haplotype. Thus, there are few studies that have mapped genome-wide allelic chromatin folding, and the results of these studies found only few or mild contact differences between the maternal and the paternal genomes in mammalian cells (Dixon et al., 2015; Greenwald et al., 2019; Holwerda et al., 2013; Rao et al., 2014; Tan et al., 2018a), with exception of the X chromosome in female X inactivation (Giorgetti et al., 2016). In contrast with ligation-based methods, GAM detects positive windows by counting reads that are present in one relatively large window several tens of kilobases (30 kb in Beagrie et al. 2017). Cryosections, as they are used for GAM, mostly contain a single allelic copy, with two being found in only ~8 % of slices containing an allele (Lavitas, 2011). Seeing that genome-wide assays to map chromatin contacts appear to lack sensitivity to allelic differences, I set out to explore allelic chromatin folding with GAM, an orthogonal approach to imaging and 3C-based techniques.

Here, I present allele-specific analyses of RNA-seq and GAM data collected from two biological replicates of mESCs, clone F123. I produced a large GAM dataset with a total of 3664 nuclear profiles. I explore the reproducibility between GAM data replicates and modes of collection (1NP or 3NP) and investigate how effectively allelic-specific reads can be phased in GAM nuclear profiles. The haplotype-assigned sequence information from the nuclear profiles was then used to generate allele-specific chromatin contact maps, which were analysed for differences between the alleles. Finally, allele-specific chromatin contacts were compared with gene expression differences between alleles, and used to explore enrichments for transcription factor binding sites that have the potential to be involved in the establishment of bi- and monoallelic chromatin contacts.

4.2 Experimental design

To map haplotype-specific chromatin contacts, I applied GAM on a hybrid mouse ESC line, F123, which was previously used in genome-wide studies of allelic differences, such as replication timing and chromatin contact frequencies (Giorgetti et al., 2016; Rivera-Mulia et al., 2018). The F123 mESC cell line was derived from an F1 cross between *Mus musculus castaneus* × S129/SvJae (Gribnau et al., 2003), and was kindly provided by the laboratory of Prof. Bing Ren (Ludwig Institute, UCSD, CA, USA). The sequence differences between the inbreed S129/SvJae and the wild *M. castaneus* line result in a diploid genome with a high SNP density, as shown by the median distance between SNPs of ~50 bp across all chromosomes (Figure 4.1).

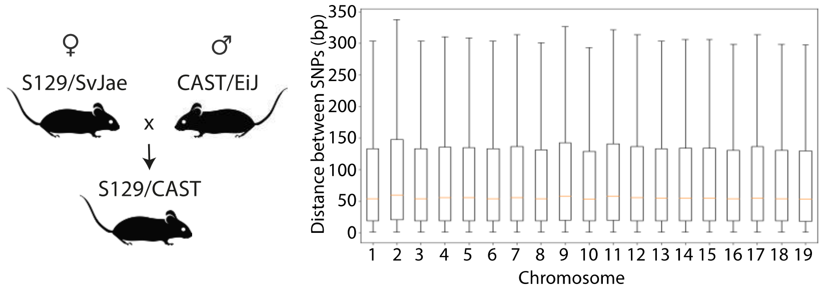


Figure 4. 1: Single nucleotide polymorphism (SNP) density of the F123 genome.
The F123 line was derived from a cross between S129/SvJae and *Mus musculus castaneus* (CAST/EiJ). The hybrid line has a high SNP density, which can be exploited to distinguish between the haplotypes in the sequencing data. The box plot shows the distribution of nearest neighbour distances between SNPs of the CAST and the S129 allele to the reference genome (mm10) for each chromosome.

With the aim to explore allelic differences in chromatin folding and how they relate with differential gene expression, I produced GAM in two biological replicates from F123 mESCs, and simultaneously collected RNA for genome-wide expression analysis (Figure 4.2). I also explored published ChIP-seq and ATAC-seq data produced for F123 mESCs grown in identical conditions (Juric et al., 2019).

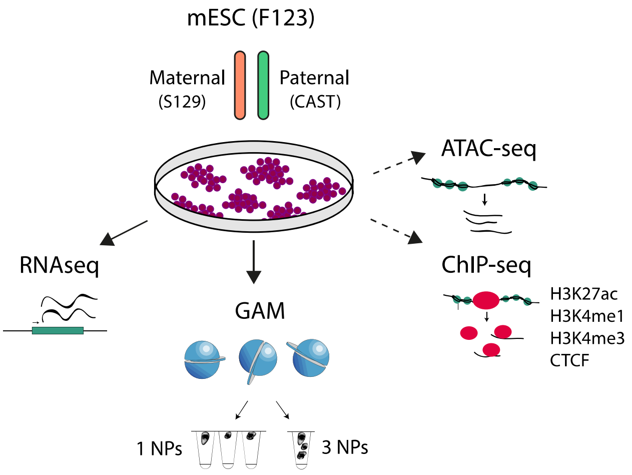


Figure 4. 2: Data collection in F123.
To study chromatin folding with allele specificity, GAM was applied in F123 mESCs, both standard GAM (collecting 1 NP per tube) and multiplexed GAM (collecting 3 NPs per tube) in two biological replicates. For genome-wide information about gene expression, total RNA-seq was collected on the same biological replicates. Published ATAC-seq and ChIP-seq datasets from F123 mESCs grown in the same cell culture conditions were used to study open chromatin regions, histone modifications, and architectural proteins.

4.3 F123 mESCs cell culture preserved expression of pluripotency markers

F123 mESCs are grown in serum/LIF on a layer of mitotically inactivated mouse embryonic fibroblasts (MEFs) which act as feeder cells. Visual inspection of the cultures on a brightfield microscope showed the expected growth of ESCs in clusters on top of the feeder cells (Figure 4.3a). To harvest ESCs for genome-wide mapping of gene expression or chromatin contacts, they are first separated from feeder cells and kept for a few days in culture. To assure that the F123 mESCs were separated successfully from the majority of feeder cells and have not started to differentiate, I monitored the RNA and protein levels of markers of pluripotency and markers of early differentiation in the two biological replicates used to collect GAM data. I confirmed the absence of mycoplasma contamination in the culture (Figure 4.3b) and generated cryoblocks of F123 cells that I used to perform immunofluorescence and later used for GAM. The expression of two pluripotency markers, Nanog and Oct4, was confirmed by immunofluorescence microscopy on cryosections in both replicates (Figure 4.3c). Oct4 expression was detected in 91-94 % of cells, while Nanog expression was more heterogeneous and detected in 61-67 % of cells, as expected (Hatano et al., 2005). ESC cultures grown in serum/LIF often contain a small fraction of cells which express early differentiation markers. To investigate the proportion of cells showing signs of early differentiation, I quantified the number of cells expressing Gata6, which is not expressed in pluripotent cells, and found that the F123 cultures contained only 2-4 % of cells expressing Gata6, within the range expected for ESC growth in serum/LIF on feeder cells (Figure 4.3c). To further quantify the pluripotency levels of both replicates, I also explored the expression levels of the mentioned transcription factors, and additional factors marking pluripotency or differentiation, using the matched total RNA-seq datasets. The expression (in TPM) of several maker genes was quantified using RSEM (Li and Dewey, 2011) and shows comparable levels of all examined transcription factors (Figure 4.3d), confirming the overall pluripotent state of the mESC F123 samples collected.

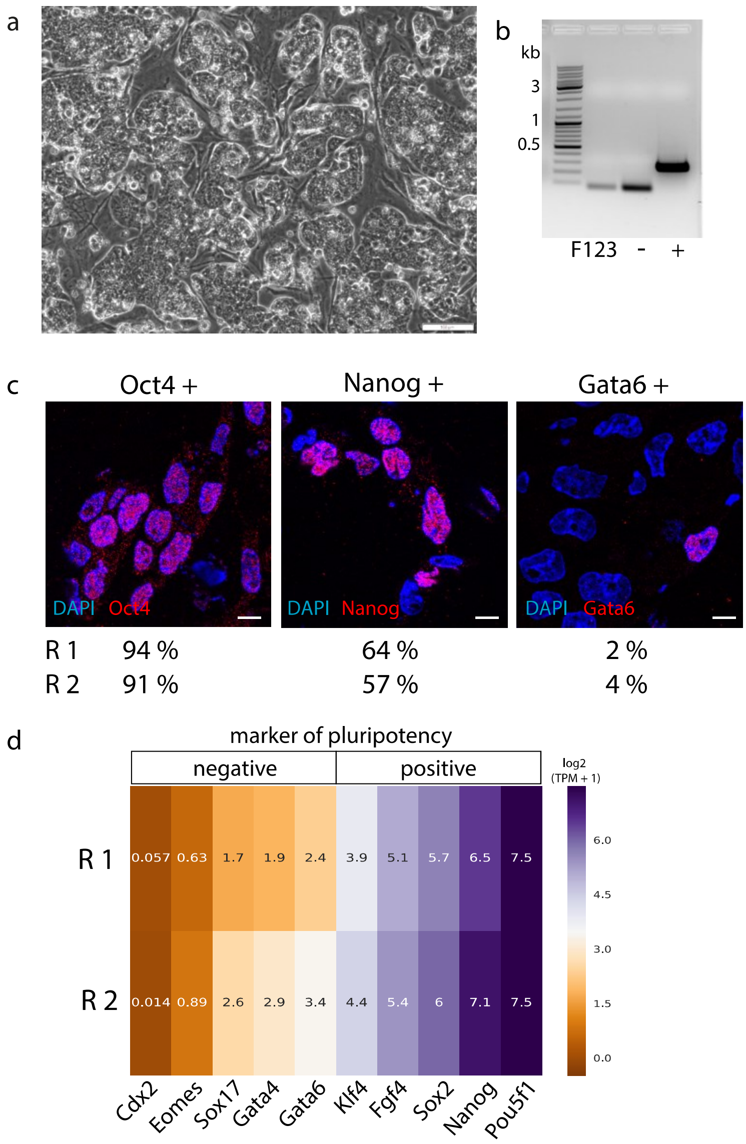


Figure 4. 3: F123 mESCs show important features of pluripotency.

(a) F123 mESCs were grown on feeder cells (irradiated MEFs). Scale bar = 100 µm. (b) Mycoplasma test by PCR on 1.2 % agarose gel, with Gene Ruler 1kb plus ladder, shows that F123 cell culture was free from mycoplasma DNA. (c) Quantification of cells that express pluripotency marker proteins, Oct4 and Nanog, and differentiation marker, Gata6, after indirect immunofluorescence in thin sections from the same cryoblocks produced from 2 biological replicates (R1, R2) for GAM data collection. Example images show Oct4, Nanog, and Gata6 in DAPI-stained cryosections. Scale bars = 8 µm. (d) Gene expression ($\log_2$(TPM +1)) of negative and positive markers of pluripotency measured by total RNA-seq in R1 and R2.

4.4 Allele-specific expression occurs at genes with housekeeping functions

To be able to identify genes with differential expression of the parental alleles, I collected RNA from F123 mESCs and produced RNA-seq libraries from two biological replicates. Gene expression levels were calculated as transcripts per million (TPM). The levels of gene expression (in TPM) correlate well across all genes (Figure 4.4a), and show reproducible read patterns on the genome browser, including good detection of first introns as expected when RNA integrity was retained during extraction and library production (Figure 4.4b).

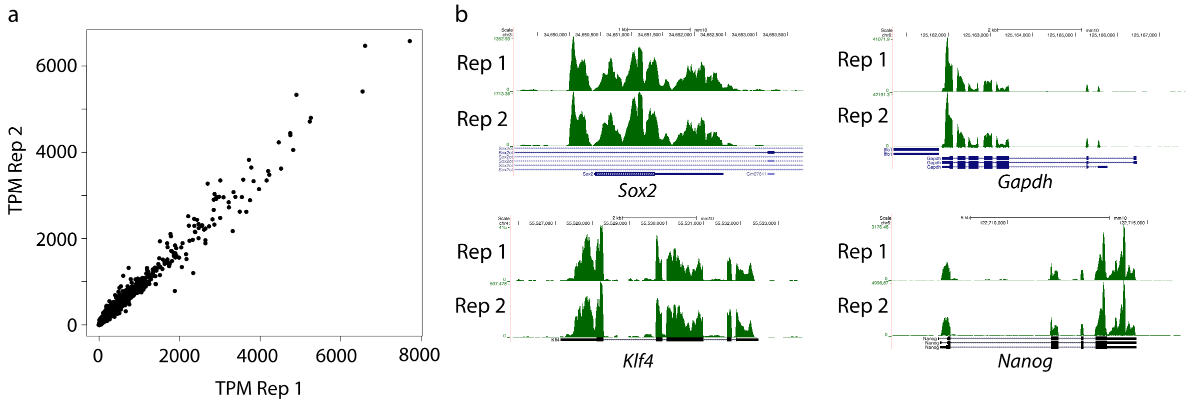
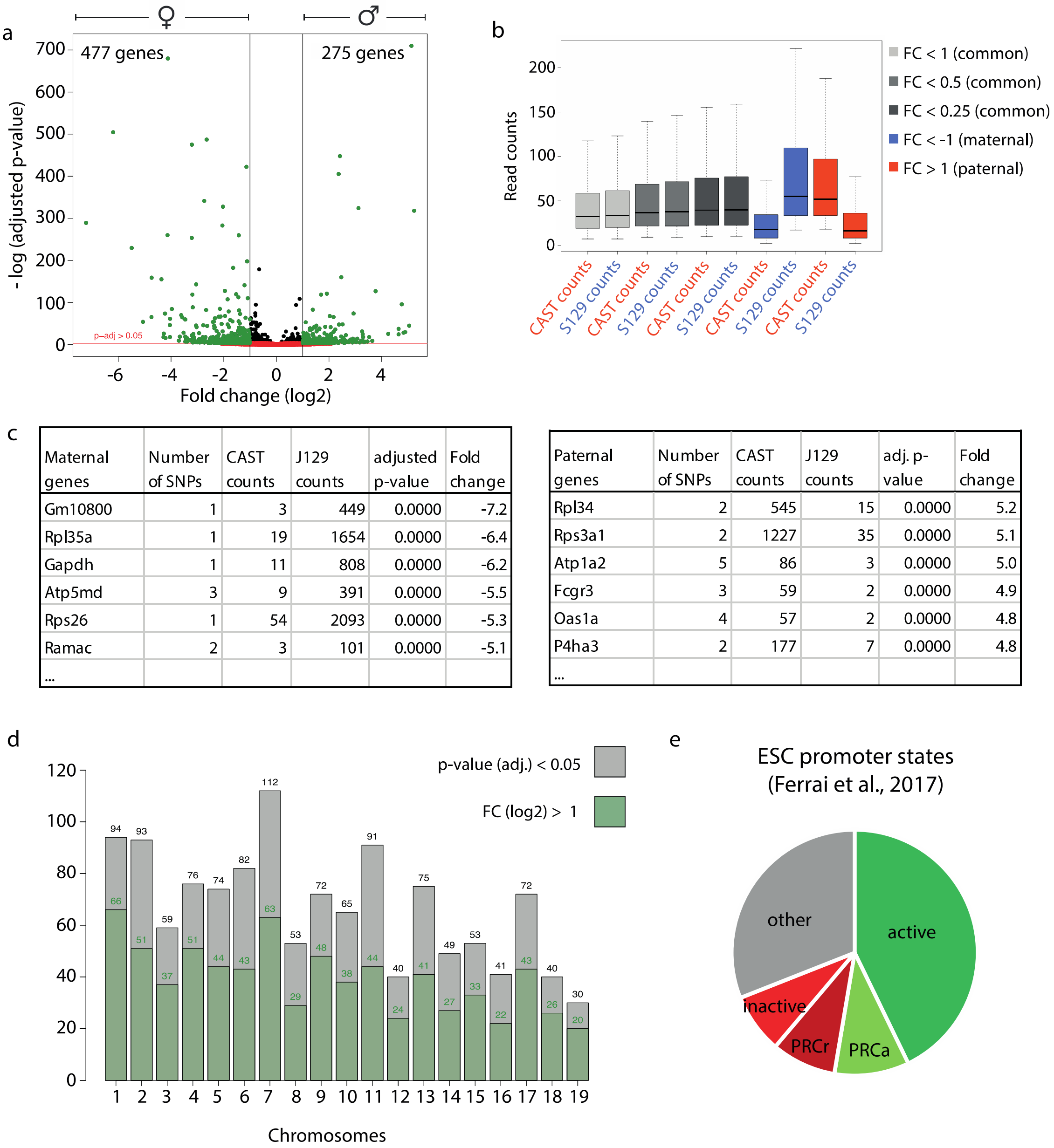


Figure 4. 4: RNA-seq data is reproducible between biological replicates.
(a) Transcripts per million (TMP) values for two biological replicates of F123 mESCs. (b) RNA read counts of examples of expressed genes in both biological replicates.

To phase reads to the S129 and CAST parental alleles, we first mined genome sequencing data for the parental strains of F123 to identify SNP positions in the F123 genome (details in 2.20.8). To differentiate the expression of genes located on the two parental alleles, reads overlapping heterozygous genomic variants were counted in an allele-specific manner. Phased reads were aggregated over all exons of a gene and aggregated counts were tested for significant allele-specific expression (ASE) differences. Genes were defined as differentially expressed by an adjusted p-value < 0.05, and a fold change (log2) > 1 between reads mapping to CAST and S129 (Figure 4.5a). Further, genes with allelically balanced expression were identified based on the fold change between reads mapping to CAST and S129 and for subsequent analysis genes with a fold change (log2) < 0.25 were selected as allelically balanced genes, which show undetectable differences in read counts between the alleles (Figure 4.5b). Interestingly, the top five genes with ASE (Figure 4.5c) on both alleles include four genes encoding for proteins of ribosomal subunits (*Rpl34, Rpl35a, Rps3a1, Rps26*), two genes important for ATP production (*Atp5md*, encoding for ATP synthase and *Atp1a2*, an ATPase). Additional genes included *Gapdh*, an important metabolic gene, *Ramac*, encoding the regulatory subunit of the mRNA-capping methyltransferase RNMT:RAMAC complex, required for efficient gene expression, *Fcgr3*, encoding for an immunoglobin receptor, *Oas1a*, encoding for an enzyme involved in the innate antiviral response of the cell, and *P4ha3*, encoding for a key enzyme in collagen synthesis. These genes are key to general cellular processes, and most of them are expressed in many cell types or tissues. In total, we identified 477 maternally expressed genes, and 275 paternally expressed genes, distributed throughout the genome (Figure 4.5d). Gene ontology analysis of genes with ASE using the GO-Elite package, versus all expressed genes with SNPs (i.e. all phased genes) as background group, confirmed their role in housekeeping functions of the cell, such as various metabolic processes and translation (Table 4.1).

For a first exploration of the possible regulation of allele-specific genes, I investigated whether Polycomb regulation was potentially involved, by matching the list of ASE genes with published data of promoter states from mESCs collected in 46C cells (Ferrai et al., 2017). The classification of gene promoters into active, inactive, and different states of Polycomb regulation in mESCs is based on the occupancy of RNA polymerase II (RNApol II) and H3K27me3, a mark set by Polycomb repressive complex 2 (PRC2). Different states of RNApol II can be identified at active and Polycomb repressed genes, which are also poised for activation. Active genes that produce mature mRNAs are associated with RNApol II phosphorylated on serine-5 and serine-7 residues of the C-terminal domain of the largest subunit of RNApol II. In contrast, Polycomb-repressed genes are found associated with a poised form of RNApol II marked only by serine-5. The most interesting class in the context of allele-specific expression is the class of PRC active (PRCa) genes. These genes are marked by H3K27me3 and RNAPII-S5p on one allele and by actively transcribing RNApol II on the other allele (Brookes et al., 2012; Ferrai et al., 2017). Finding this promoter state at genes with allele-specific expression would provide a possible mechanism for their allelic regulation. In this case, H3K27me3 would be specific to only one allele, and Polycomb-mediated silencing could specifically repress the transcription of one allele, while the other allele is transcribed. Of all 752 genes with allele-specific expression, 127 were not classified in Ferrai et al. because their promoters were too close to other active or Polycomb-repressed promoters, but 74 have been described as PRCa (Figure 4.5e). This first exploration of promoter regulation of the ASE genes suggests that Polycomb regulation may not be the primary reason for allele-specific gene expression, however differences between the two mESC lines and their culture conditions should be considered. Further studies could take advantage of simply using the H3K27me3 classification itself to avoid the high number of NA genes, and of exploring other promoter classifications, for example obtained by our lab in the ESC-OS25 line (Brookes et al. 2012). Further work will also explore the 74 genes with the PRCa classification, for instance regarding the fold change of expression of those genes, or whether there are differences between PRCa and other genes with regard to the expression level of the silent allele.



Maternal genes	Number of SNPs	CAST counts	J129 counts	adjusted p-value	Fold change
Gm10800	1	3	449	0.0000	-7.2
Rpl35a	1	19	1654	0.0000	-6.4
Gapdh	1	11	808	0.0000	-6.2
Atp5md	3	9	391	0.0000	-5.5
Rps26	1	54	2093	0.0000	-5.3
Ramac	2	3	101	0.0000	-5.1
...					

Paternal genes	Number of SNPs	CAST counts	J129 counts	adj. p-value	Fold change
Rpl34	2	545	15	0.0000	5.2
Rps3a1	2	1227	35	0.0000	5.1
Atp1a2	5	86	3	0.0000	5.0
Fcgr3	3	59	2	0.0000	4.9
Oas1a	4	57	2	0.0000	4.8
P4ha3	2	177	7	0.0000	4.8
...					

Figure 4. 5: Allele-specific gene expression in F123 mESCs.

(a) Allele-specific analysis of gene expression. Plot shows the fold change of phased reads between the maternal (S129) and the paternal (CAST) allele and the negative logarithm of the adjusted p-value (including only expressed genes with a TPM value > 1) for each gene. Allele-specific expression (ASE) is defined by an adjusted p-value < 0.05, and a log2 fold change > 1. Green dots represent genes with ASE, black and red dots represent genes with allelically balanced expression. (b) Counts of reads mapping to the CAST or the S129 allele in genes without ASE (common, sorted by fold change) and genes with ASE (maternal and paternal). (c) Examples of genes with ASE. Number of SNPs, read counts, adjusted p-value, and fold change (log2) of the top five genes with ASE are shown for the maternal and the paternal allele. (d) Number of genes with ASE per chromosome. Grey bars show all genes with significant ASE (p< 0.05), green bars represent all genes with significant ASE and a fold change (log2) > 1. (e) Percentage of allele-specific genes with published promoter states in mESCs (Ferrai et al., 2017). Promoters were classified based on the presence of different states of RNA polymerase II (RNAPII) phosphorylation (S5p and S7p, which mark both active and polycomb-repressed genes, or only active genes, respectively) and H3K27me3, a mark for Polycomb repressive complex 2 (PRC2). Promoter states include active genes (with RNAPII-S5p and -S7p), inactive genes (no RNApol II), PCRa genes (with RNAPII-S5p and -S7p, and H3K27me3), and PCRr genes (with RNAPII-S5p and H3K27me3).

Table 4. 1: Gene ontology (GO) analysis of genes with allele-specific expression (ASE). Gene class enrichment of genes with ASE (752) over a group of background genes (Ref) containing all expressed genes (9725) in F123 mESCs (with TPM> 1) that contain SNP information (Number of SNPs > 1). Total number of ASE genes with GO term: 712, total number of background genes with GO term: 9161. GO terms for biological processes, cellular components, and molecular functions are ranked by z-score. For relevance purposes, only GO terms with more than 10 genes in the reference list are shown. GO enrichment analysis was performed with GO-Elite.

Ontology-ID	Ontology name	Genes with ASE	Genes in ref	Genes in ontology	Percent with ASE	Z-score	Permuted P-value	Examples
Biological process								
GO:0006412	translation	31	134	223	23.1	6.73	0	Mrpl52, Rpl13, Rpl14, Rpl17, Rpl21, Rpl27a, Rpl28
GO:0019748	secondary metabolic process	6	13	31	46.2	5.19	0	Akr1b3, Bdh2, Cyp1b1, Ephx2, Pam, Star
GO:0042273	ribosomal large subunit biogenesis	5	12	12	41.7	4.41	0	Bop1, Npm1, Rpl14, Rpl35a, Rpl7
GO:0034308	primary alcohol metabolic process	5	12	34	41.7	4.41	0.001	Aldh1b1, Cyp1b1, Pecr, Rbp1, Retsat
GO:0007129	synapsis	5	12	19	41.7	4.41	0.001	Mael, Mlh1, Stra8, Sycp3, Tex15
GO:0016101	diterpenoid metabolic process	6	17	51	35.3	4.26	0.002	Crabp2, Cyp1b1, Pecr, Rbp1, Retsat, Star
GO:0043436	oxoacid metabolic process	53	414	726	12.8	3.96	0	Aldh9a1, Bdh2, Cd74, Crabp2, Cyp1b1, Gapdh, Ldhb, Pam
GO:0031214	biomineral tissue development	7	25	64	28.0	3.80	0.0025	Axin2, Enpp1, Fgfr2, Msx2, Pkdcc, Srgn, Tuft1
GO:0015908	fatty acid transport	5	16	37	31.3	3.53	0.004	Anxa1, Got2, Slc25a17, Slc27a2, Slco2a1
GO:1901568	fatty acid derivative metabolic process	5	17	59	29.4	3.35	0.0065	Cd74, Cyp1b1, Pdpn, Ptgr2, Rnpep
GO:0042274	ribosomal small subunit biogenesis	4	12	16	33.3	3.32	0.01	Npm1, Rps15, Rps19, Rps6
GO:0051258	protein polymerization	7	29	57	24.1	3.32	0.004	Arl6, Fbxo5, Gas7, Tuba1c, Tuba3b, Tuba4a, Tubb6
GO:0006383	transcription from RNA polymerase III promoter	5	18	20	27.8	3.19	0.0085	Crcp, Polr1d, Polr2h, Polr2k, Polr3h
Cellular component								
GO:0022627	cytosolic small ribosomal subunit	11	26	38	42.3	6.62	0	Rps12, Rps15, Rps15aRps19, Rps2, Rps29, Rps3a1Rps8
GO:0005840	ribosome	27	129	217	20.9	5.66	0	Mrpl35, Rpl13, Rpl29, Rps15, Rps15a, Rps3a1, Rps8
GO:0009897	external side of plasma membrane	14	68	240	20.6	3.99	0.0005	Abcg1, Cd55, Cd59a, Clptm1, Fcgr2b, H2-K1, Ly75, Spn
GO:0005576	extracellular region	34	244	1219	13.9	3.69	0	Adam23, , Bola1, Dpp7, Fcgr2b, Fgfr2, Hsd17b11
GO:0000795	synaptonemal complex	5	17	24	29.4	3.35	0.0065	Ccnb1ip1, Mlh1, Stag3, Syce1, Sycp3
GO:0032993	protein-DNA complex	11	58	144	19.0	3.22	0.002	H3f3b, Hist1h1t, Hist1h2ab, Hist1h2ak, Hist1h2bb
GO:1990104	DNA bending complex	8	37	116	21.6	3.17	0.004	H3f3b, Hist1h1t, Hist1h2abHist4h4, Kat6b
GO:0015934	large ribosomal subunit	9	48	64	18.8	2.87	0.013	Mrpl12, Rpl10l, Rpl14, Rpl17, Rpl35a, Rpl7
Molecular function								
GO:0003735	structural constituent of ribosome	28	119	208	23.5	6.51	0	Mrpl12, Rpl34, Rpl35a, Rpl41, Rpl7, Rps12, Rps15, Rps15a, Rps18
GO:0004364	glutathione transferase activity	5	11	32	45.5	4.69	0.001	Gsta3, Gstm2, Gstm5, Gsto1, Gstz1
GO:0017134	fibroblast growth factor binding	5	14	20	35.7	3.93	0.002	Api5, Fgf2, Fgfr2, Klb, Rps19

GO:0016829	lyase activity	16	98	174	16.3	3.21	0.002	Amd1, Aplf, Car2, Car8, Cd38, Echdc2, GlulShmt1, Uros
GO:0051287	NAD binding	8	37	59	21.6	3.17	0.0055	Ahcyl2, Aldh9a1, Bdh2, Gapdh, Ldhb, Phgdh, Sirt5
GO:0005544	calcium-dependent phospholipid binding	4	13	25	30.8	3.11	0.0145	Anxa1, Anxa4, Anxa7, Sytl3

4.5 Generation of GAM datasets with single and multiplexed nuclear profiles

With the aim to generate contact maps from F123 mESCs, I collected GAM samples from two biological replicates (R1 and R2), in two collection modes (1NP and 3NP). The 1NP collection led to the production of a GAM dataset with 1123 1NP samples from R1, which was discussed in the previous chapter (section 3.5) in comparison with the previously produced GAM dataset in 46C mESCs. To further increase the resolution of the F123 GAM dataset, I collected a larger number of NPs using multiplex GAM with 3NP/sample (see Figure 3.3 for details) in both R1 and R2. Although 3NP datasets lose the single cell information that is inherent to 1NP GAM data, 3NP collection allows three-fold faster production of GAM data, suitable for the aim of increasing the genomic resolution to analyse chromatin contacts. The 1NP data has the advantage that it can be used to study single-cell dependencies and radial positioning. Additionally, I explored the possibility of combining 1NP and 3NP data into one large GAM dataset with improved resolution, and use this data to study pairwise chromatin contacts that are consistent between collection modes (1NP and 3NP) and biological replicates (R1 and R2).

To study the quality of each dataset and identify criteria for exclusion of low-quality samples between datasets, I first compared the quality of NPs collected with the single and the multiplexed collection mode (Figure 4.6). As expected from combining genomic information from several NPs, the 3NP samples have an overall higher number of chromosomes and a higher genome coverage than the 1NP samples (Figure 4.6a), with similar range of percentage of orphan windows, consistent with the extractability of genomic windows being independent of whether the WGA reaction occurs on 1 or 3 NPs. Further, the quality of the samples in terms of the number of uniquely mapped reads is also consistent between 1NP and 3NP samples, and between 3NP samples from different biological replicates (R1 and R2) (Figure 4.6b). Thus, we applied the same quality control criteria chosen previously for the F123 data (Chapter 3) to pass the 3NP samples (> 20,000 uniquely mapped reads and < 60 % orphan windows) (Figure 4.6a, b, Table 4.2). Next, I compared NPMI-normalised co-segregation matrices of all three datasets, and found that the pairwise chromatin contact maps are visually similar between all the datasets (1NP and two 3NP datasets from different biological replicates; Figure 4.6c). Visual inspection of the matrices suggests that the 3NP data from the second biological replicate (R2) has a slight increase in noise, visible by non-continuous contact patches at larger genomic distances, which suggests that the dataset is not fully saturated, and might be difficult to analyse on its own with the current number of NPs.

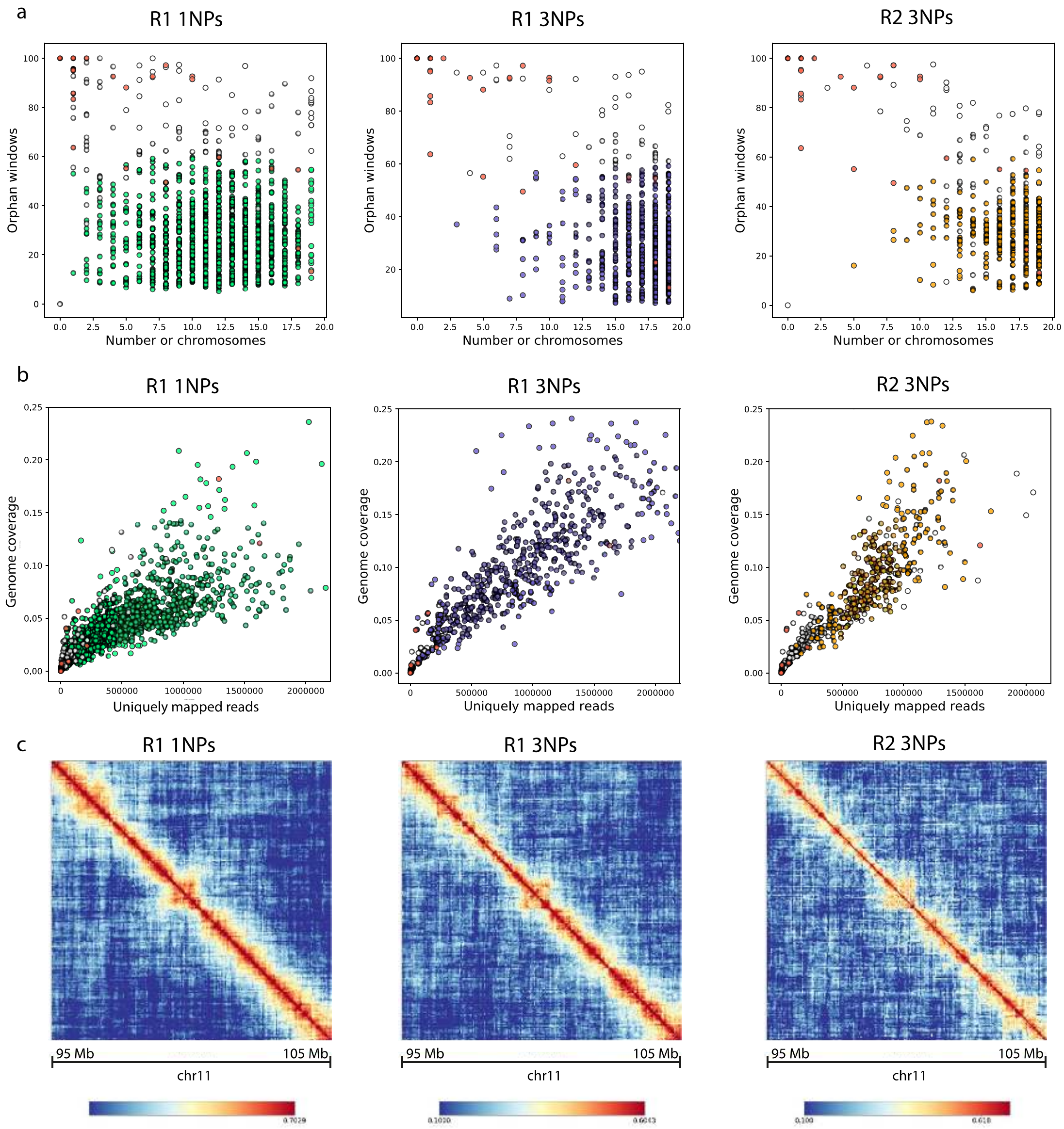


Figure 4. 6: Comparison of data quality in F123 datasets.
Quality metrics of replicate 1 (R1) – 1NPs (green), R1 – 3NPs (purple), and R2 – 3NPs (orange). (a) Percentage of orphan windows vs number of chromosomes. (b) Percentage of genome coverage vs. number of uniquely mapped reads. For both, negative controls are shown in red, samples that failed the quality metrics are shown in grey. (c) Normalised co-segregation frequencies (NPMI) of a 10 Mb region on chromosome 11.

Table 4. 2: Number of nuclear profiles (NPs) per F123 dataset.

Dataset	**1NP - R1**	**3NP - R1**	**3NP - R2**
Samples collected	1283	586	322
Cross-contaminated	22	1	0
Samples passed QC	1123	534	313
Number of NPs in final dataset	1123	1602	939
Total number of NPs	3664		

4.6 Comparing GAM datasets from biological replicates and collection modes

For the purpose of combining NPs from different biological replicates and collection modes into one large GAM dataset, I first explored the possibility of merging the two biological replicates into one dataset and subsequently study primarily contacts that are frequent in both replicates, while at the same time reduce the noise in the data due to the increased number of NPs. From exploration of the RNA-seq data, we found that gene expression highly correlates between the two biological replicates (Figure 4.4a, b). To test whether this correlation is also present in the GAM datasets, I compared the window detection frequencies of the two replicates of 3NP datasets (Figure 4.7a). There is a strong correlation ($r^2 = 0.8$) between the average window detection frequencies of R1 and R2, which shows that genome sampling is highly comparable between the two replicates. Further, principal component analysis (PCA) can be used to identify A and B compartments in chromatin contact maps. We compared the localisation of A and B compartments between GAM datasets as an approach to assess the reproducibility of 3D genome topology between the datasets. A and B compartments overlap largely between the replicates (Figure 4.7b), with a correlation of $r = 0.6$. Future analyses will investigate the overlap of compartments A and B across the linear genome as additional comparison metrics between biological replicates. Based on these observations, we combined the two replicates into one merged 3NP dataset. Inspection of NPMI-normalised matrices show less apparent noise compared to the replicates (Figure 4.7c).

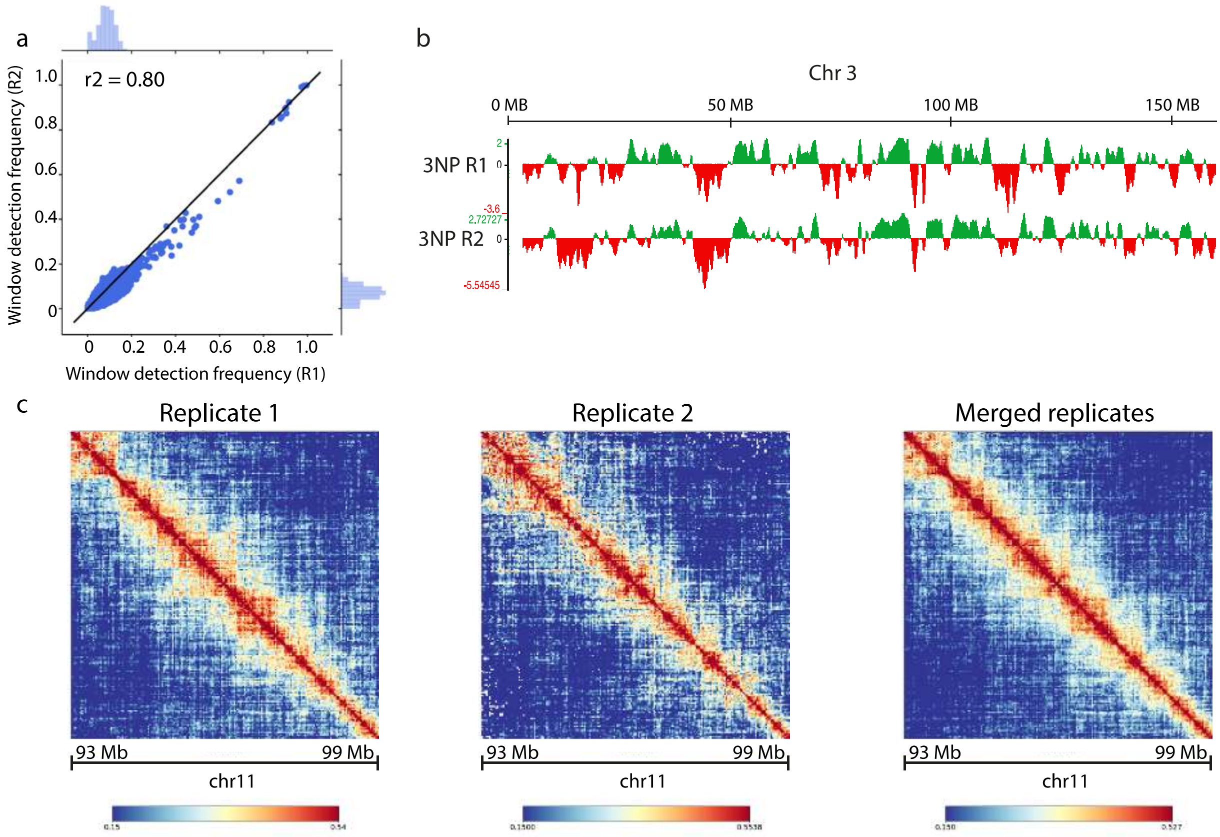
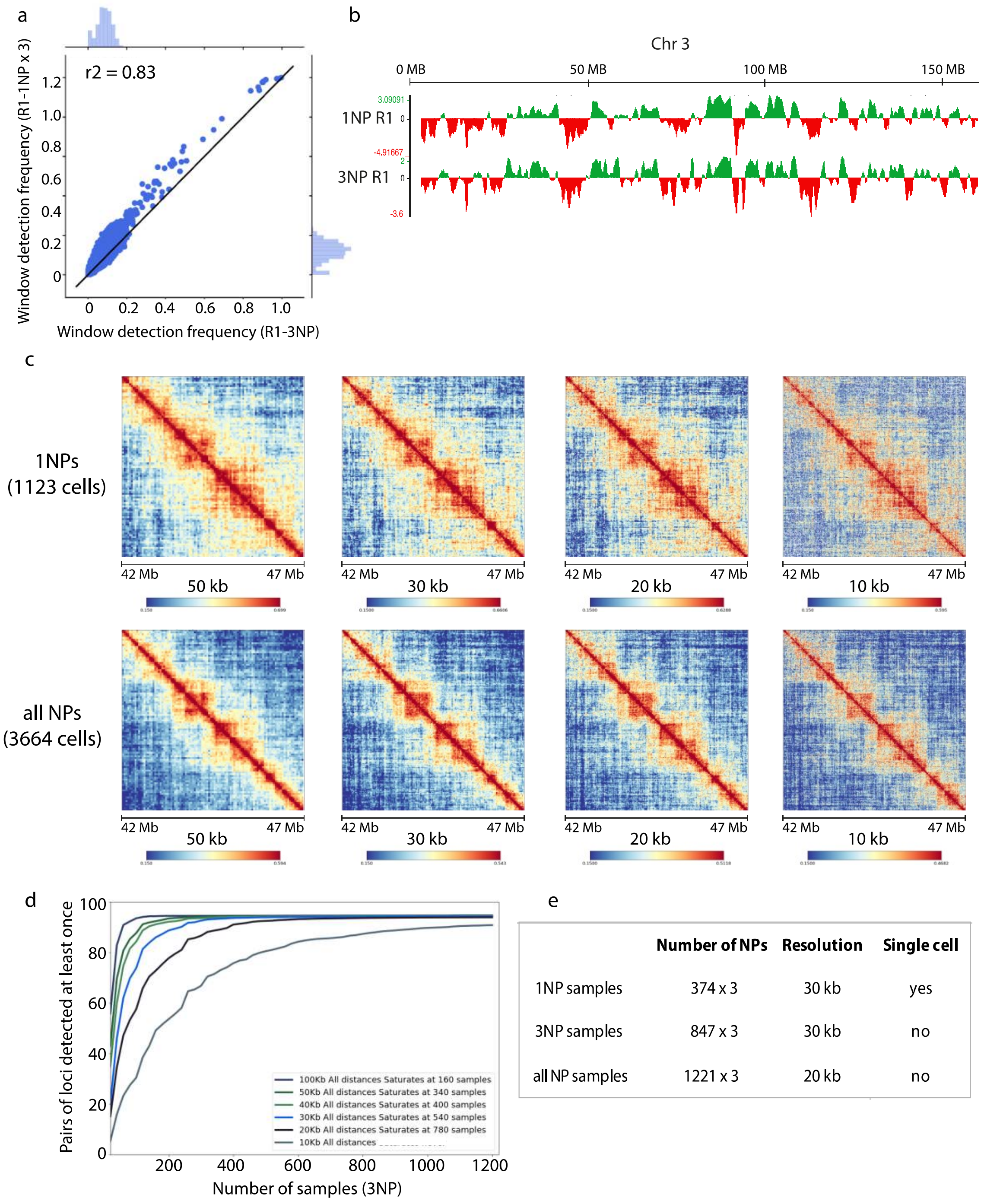


Figure 4. 7: Comparison of biological replicates from 3-NP GAM datasets.

(a) Correlation of the window detection frequencies in biological replicate 1 and 2 (R1, R2). (b) Eigenvalues of the principal component analysis (PCA) of contacts on chromosome 3 in R1 and R2, showing the A (green) and the B (red) compartment. (c) Normalised co-segregation frequencies (50 kb) for a 6 Mb genomic region on chromosome 11 of R1, R2, and the merged replicates.

Next, I explored the differences between the 1NP and 3NP datasets. The genomic content in 1NP samples is inherently less than in a 3NP sample, which results in proportionally different window detection frequencies. As GAM data is normalised for window detection frequencies, the normalisation of a mixed population of 1-NP and 3-NP samples can result in biases. To avoid potential biases, the single NP data was combined into an in-silico 3-NP dataset (1NP x 3) by combining the positive windows from 3 random single NPs into 1 multiplexed NP. The in-silico 3NP dataset should have the same properties as real 3NP samples.
To test the effects of combining NPs for an in-silico multiplex GAM dataset, I compared window detection frequencies of the in-silico data with the real 3NP dataset (Figure 4.8a). The correlation between the datasets is strong ($r^2 = 0.83$), showing that combining single NPs after sequencing results in similar detection frequencies as collecting them together. Further, compartments A and B identified in the 1NP dataset overlap to large extent with the compartments identified in the 3NP data (Figure 4.8b), with a correlation of r = 0.59, which is comparable to the correlation between 3NP biological replicates (r = 0.60). Thus, I decided to merge all nuclear profiles together as a 3NP dataset that corresponds to a total of 3664 single NPs; this corresponds to a 9-fold increase relative to the published 408-1NP GAM dataset collected in mESC-46C.
A main advantage of combining different datasets is the improvement in resolution that accompanies the increased number of NPs in the dataset. The saturation of this large dataset can be tested by performing an erosion of samples, which provides the information about how often each pair of windows in the genome is detected at least once with increasing number of samples. As the number of samples increases, more pairs of genomic windows are detected and at higher frequency, until the dataset reaches saturation. There is a proportion of never detected windows, which overlap with unmappable regions of the genome (see Beagrie et al., 2017). While the separate 1NP-1123 and 3NP-2541 datasets only saturate at 30 kb (data not shown), the combined 3NP-3664 dataset saturates at 20 kb resolution after 2340 NPs (730 samples) and almost reaches saturation at 10 kb resolution (Figure 4.8c). In comparison to the 1NP dataset, contact maps of the all-NP dataset tend to show reduced noise at all resolutions (Figure 4.8d, e).



	Number of NPs	Resolution	Single cell
1NP samples	374 x 3	30 kb	yes
3NP samples	847 x 3	30 kb	no
all NP samples	1221 x 3	20 kb	no

Figure 4. 8: Combining all NPs for higher resolution analysis of chromatin contacts.
(a) Correlation of window detection frequencies of 1NPs (1NP x 3, in silico) and 3NPs from biological replicate 1 (R1). (b) Eigenvalues of the principal component analysis of contacts on chromosome 3, showing A (green) and B (red) compartments in 1NPs and 3NPs. (c) Normalised co-segregation frequencies in the 1NP and the all-NP dataset at different resolutions. (d) Pairs of loci detected at least once in the all-NP dataset with increasing number of 3NP samples. Saturation test of the all-NP dataset shows that all pairs of genomic loci are detected at least once up to a resolution of 20 kb. (e) GAM datasets and currently used resolutions.

4.7 Chromatin contacts of the parental alleles

Next, we aimed to develop allele-specific mapping of GAM data, taking advantage that the large dataset produced for this PhD thesis was obtained in the hybrid F123 mESC line (Figure 4.9a), with a high SNP density. The first challenge for analysing chromatin contacts with allele specificity was to phase reads to maternal and paternal genomes separately in each GAM sample. For each GAM sample, reads were assigned to their haplotype using SNPsplit, a tool for phasing reads that identifies the haplotype origin of a read by examining each SNP position for the presence of a SNP, and then assigning it to a haplotype accordingly. When for instance a paternal SNP is detected, the read is phased to the paternal genome; conversely, when the SNP corresponding to the genomic region of a given read is absent, the read is assigned to the maternal genome (Figure 4.9b). After phasing, an average of 36.5 % of the mapped reads of each sample were assigned to one allele (17.8 % to the paternal, 18.7 % to the maternal genome) (Figure 4.9c). Examples of parental read distribution in the linear genome show that reads often go the same haplotype for large consecutive regions, up to entire chromosomes, as expected to the linear collinearity of parental genomes (Figure 4.9d, NP 1 and 2). As expected from the slicing of nuclei in random orientations, also both alleles are found in the same nuclear profile (Figure 4.9d, NP 3).

Next, we used the phased reads to call allele-specific positive windows at different resolutions. Reproducibly for different resolutions, up to 85 % of all positive windows could be assigned to the alleles, with equal proportions for the maternal and the paternal allele (Figure 4.9d). We observed a smaller percentage of windows that were assigned to both alleles (5.8 – 10.8 % of positive windows, depending on the resolution), which is expected and confirmed by cryo-FISH of specific genomic loci in ultrathin cryosections (Lavitas, 2011). After phasing the windows, the detection frequencies for each allele decrease compared to the non-phased data by the expected 2-fold reduction (Figure 4.9e).

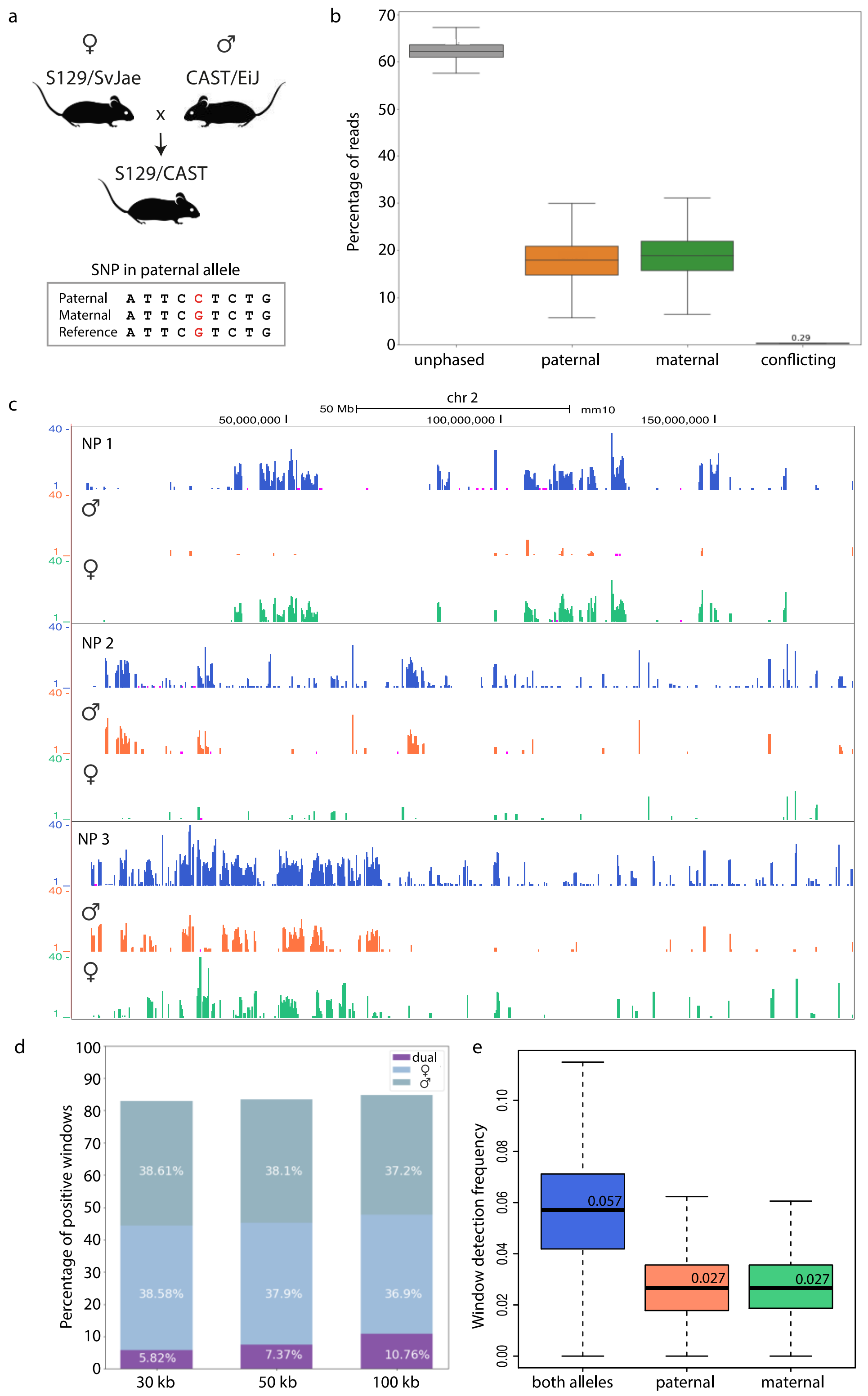


Figure 4. 9: Phasing of F123 nuclear profiles.

(a) SNPs between the alleles S129 (maternal) and CAST (paternal) are compared to the reference genome (mm10) for phasing of sequencing reads. (b) Percentage of reads per NP that have no SNP information (unphased), that are mapped to the maternal or the paternal allele, and conflicting reads that have SNP information from both alleles and are discarded from the analysis. (c) Sequencing tracks of three example NPs for chromosome 2. Examples are shown for the non-phased, the paternal, and the maternal reads. (d) Percentage of positive windows in the F123 1NP dataset that can be assigned to one or both alleles. (e) Average window detection frequencies for the non-phased 1NP F123 dataset (both alleles), the paternal, and the maternal allele.

Using the phased positive windows, pairwise contact maps were generated and NPMI normalised separately for each parental genome (Figure 4.10). At the time of completing this PhD thesis, all analyses were performed in the F123 1NP dataset, with the future aim to reproduce the same analyses in the combined 3NP dataset. Remarkably, the parental-specific matrices display seemingly more clear contact maps than the merged data, suggesting that the latter is confounded by the allelic differences, which seem more prominent than anticipated from previous Hi-C analyses (Rao et al., 2014: Dixon et al., 2015, Rivera-Mulia et al., 2018).

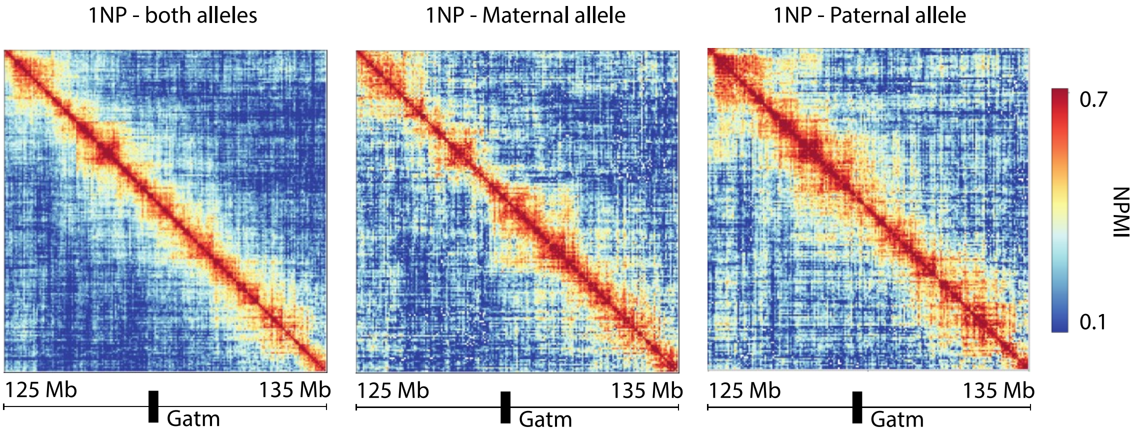


Figure 4. 10: Chromatin contacts of the paternal and the maternal allele.
Chromatin contacts of the non-phased 1NP dataset and of the parental alleles for a 10 Mb region on chromosome 2 (50 kb), centred at the paternally imprinted gene *Gatm*.

4.8 Allele specificity of compartments A and B

To explore genome-wide chromatin topologies in the allelic contact maps, A and B compartments were identified in the non-phased and the phased F123 datasets, at 100 kb resolution (Figure 4.11). Visual comparisons show that the allele-specific compartments often overlap, resembling the pattern of the non-phased data (Figure 4.11a). However, there are many regions of the genome, where the allele-specific compartments deviate from each other. To understand how different the compartments of the alleles are from each other, in comparison to the compartments of different non-phased datasets, we compared them to the two 3NP datasets, R1 and R2 (Figure 4.11a). Interestingly, we observed that differences between the alleles occur more often than differences between non-phased datasets, even when they come from different biological replicates. This observation could be confirmed when calculating the overlap of the compartments between all F123 datasets, which is lowest between the two allelic datasets (Figure 4.11b). In total, 25.8 % of the F123 genome shows differences in compartments between the maternal and the paternal allele, with equal contributions of A and B compartment (Figure 4.11c).

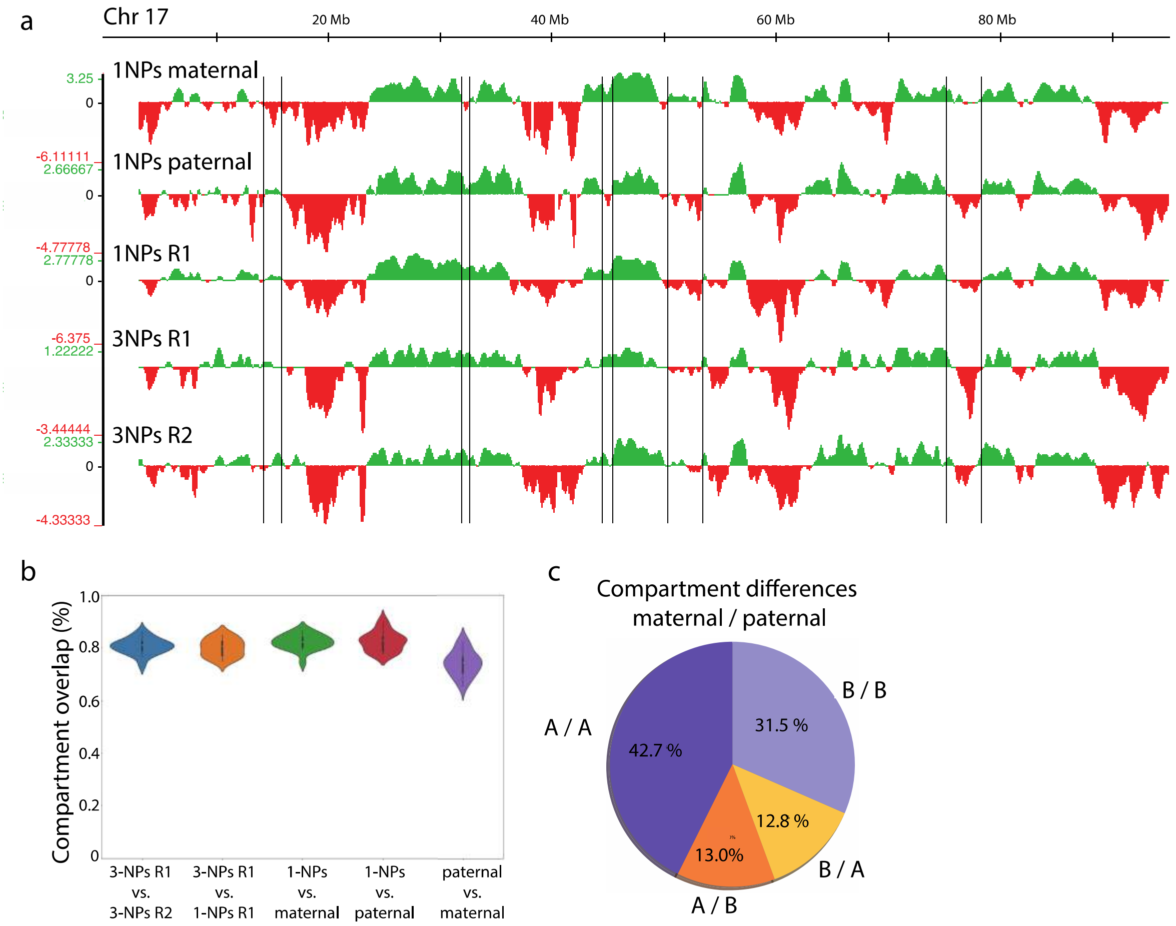


Figure 4. 11: Large-scale differences between A/B compartments of the maternal and the paternal allele. (a) Principal component analysis (PCA) at 100 kb resolution showing the A and B compartment of chromosome 17 in different F123 datasets. Black lines mark example regions with compartment changes between the maternal and the paternal allele. (b) Overlap of A/B compartments of all chromosomes between F123 datasets. (c) Compartment differences between the maternal and the paternal allele in percentage of genomic windows at 100 kb.

To further explore the differences between allele-specific compartments, we asked whether they contained genes with different expression patterns. Genes were classified according to whether they are active (expressed, TPM-value >1) or inactive (not expressed, TPM-value <1). Although compartments A and B are enriched for active and inactive chromatin, respectively (Lieberman-Aiden et al., 2009, Dixon et al., 2012), active and inactive genes do not have strict memberships to A or B compartments (Figure 4.12a). Compartments A that are biallelic in F123, contains 60 % of all genes, with the majority of them (69 %) being active. In contrast, biallelic B compartments, which comprise 17 % of all genes, contains 31 % active genes. The monoallelic compartments (A paternal and B maternal, or vice versa) harbour a total of 23 % of genes, and equal percentages of active and inactive genes (51 % of active genes). These results suggest differences in expression between bi- and monoallelic compartments. Noteworthy, when comparing TPM values of genes within the different compartments, a clear difference in expression can be observed between the biallelic A and B compartment, while the monoallelic compartments show intermediate gene expression between the more active A and the more inactive B compartment, with no difference between the maternal and the paternal allele (Figure 4.12b).

The gene expression differences between bi- and monoallelic compartments connect gene activity with differences between alleles in the presence of eu- and heterochromatin, and thus provide an orthogonal validation of the detected compartment differences. Further, it suggests two possible mechanisms for the establishment of monoallelic compartments. First, slightly lower and heterogeneous gene expression might lead to variable association with the A or B compartment, and in some cases change the compartment on one of the alleles. Second, the association of one allele with the A and the other allele with the B compartment could be a result of allele-specific gene expression. To test this hypothesis, the fold change of expression between the alleles was compared for genes in mono-and biallelic compartments (Figure 4.12c). Generally, the majority of genes in all compartments have balanced expression between the alleles. Genes with allele-specific expression (ASE) are located predominantly in the biallelic A compartment (70 % of maternally and 72 % of paternally expressed genes), and to lesser extend in equal proportions in the biallelic B, or the monoallelic compartments, with no noticeable differences in fold change of expression with regard to their location. This first exploration of compartments and allele-specific expression revealed no genome-wide connection between the expression of a gene on one allele and a corresponding change of the compartment. However, further analysis that takes into account the distribution of genes with balanced and allele-specific expression in the genome, and their linear distance to each other, might help reveal cases where gene expression differences connect to compartment changes between the alleles.

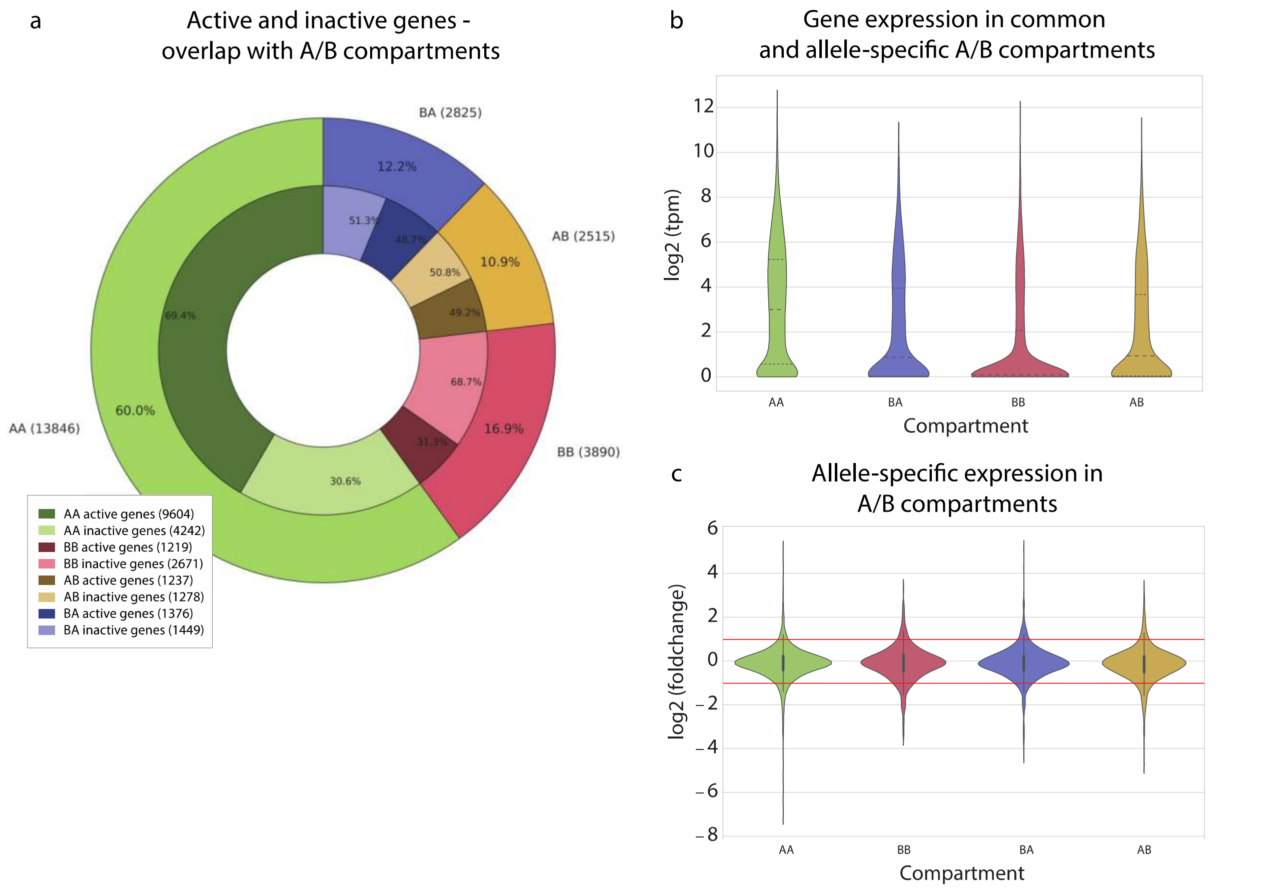


Figure 4. 12: Allele-specific A or B compartments have equal proportions of active and inactive genes. (a) Overlap of active and inactive genes with the A/B compartment in F123 mESCs. Outer ring: Percentage of genes per compartment. Inner ring: Percentage of active (TPM >1) and inactive genes (TPM <1) per compartment. (b) Expression (log2(TPM)) of genes inside biallelic (A/A, B/B) and monoallelic (A/B, B/A) compartments. (c) Fold change (log2) of read counts per allele (paternal / maternal) for genes in bi- and monoallelic compartments. Compartment labels (A/B, B/A) in a - c: maternal allele / paternal allele.

4.9 Differential contacts between maternal and paternal alleles

To explore allele-specific chromatin contacts, maternal and paternal contact maps were subtracted from each other to extract both differential and common contacts (Figure 4.13a). The data from each allele was first normalised by calculating z-scores for each genomic distance, before subtracting the z-score normalised contacts from each other, resulting in a delta z-score contact map with all differential contacts (Figure 4.13b). Next, the allele-specific contacts were extracted by setting a threshold for the significant 5 % allele-specific contacts, which most frequently contact on one allele and not the other. To explore the contacts that have equally strong intensities on both alleles, we selected the top 10 % most common and strong contacts of the alleles. This selection leaves us with three datasets, maternal-specific, paternal-specific, and common contacts, which we can use for the exploration of mono- and biallelic chromatin conformations (Figure 4.13b). Visual inspection of the allele-specific and common contact maps shows regions of clustered differential contacts which cover specific genomic windows within a neighbourhood.

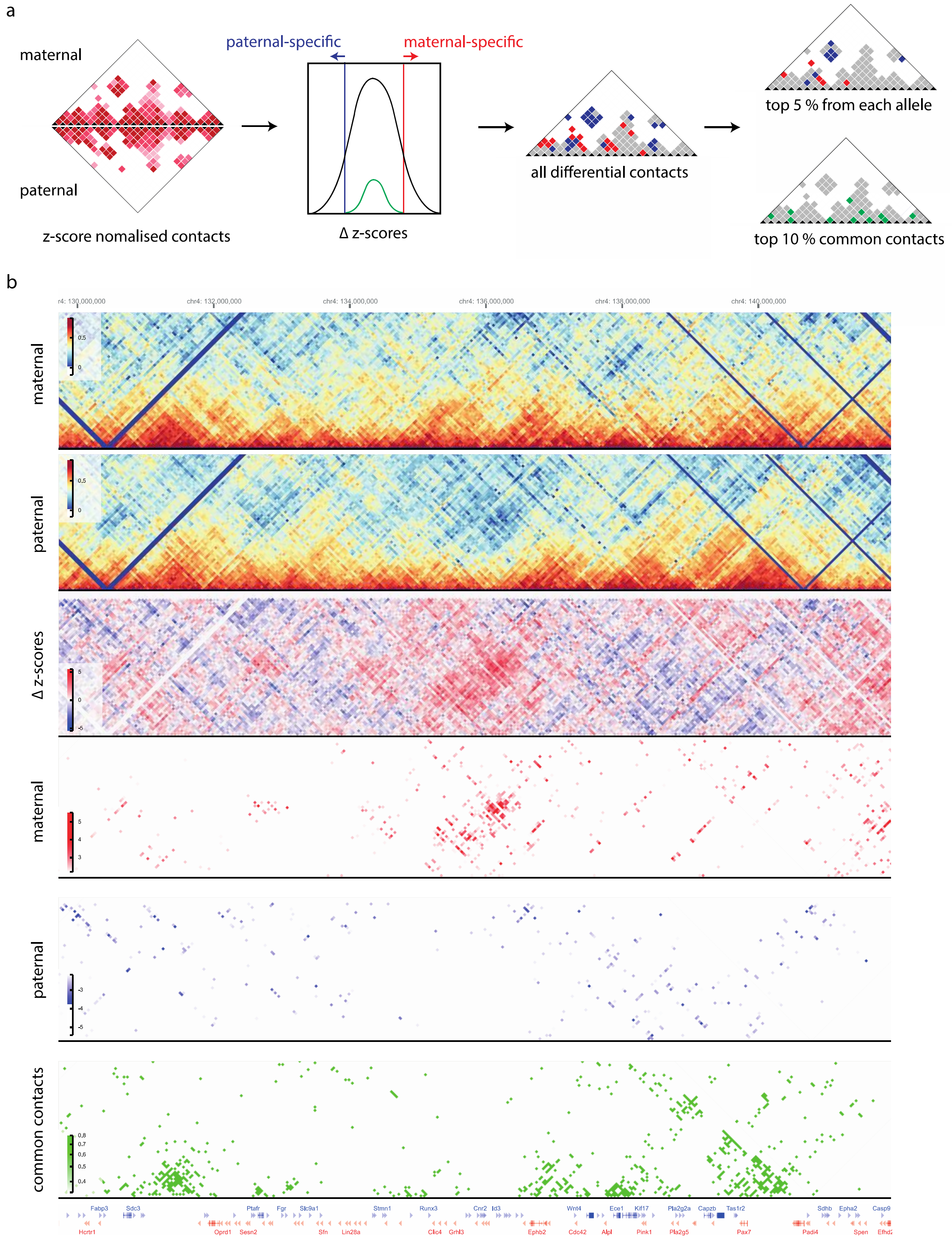


Figure 4. 13: Differential contacts between the alleles.

(a) Pipeline to identify significant differential contacts between the alleles. (b) Chromatin contacts of a 12 Mb region on chromosome 4 in the F123 1NP dataset. From top to bottom: Normalised contact frequencies (NPMI) of the maternal allele, normalised contact frequencies (NPMI) of the paternal allele, delta z-score of the maternal minus the paternal allele, significant maternal-specific contacts, significant paternal-specific contacts, top interactions that are shared by the alleles, mm10 gene annotation. Differential contacts were identified with a threshold for genomic distance. Only contacts within the range of 4 Mb are considered and shown here.

4.9.1 Most imprinted genes have allele-specific contacts

Genome-wide studies of 3D genome topology at genomic regions that have allelic differences in gene expression have not so far found a clear relationship between chromatin contacts and allelic expression differences. However, a few examples with allele-specific chromatin contacts have been reported for imprinted genes (Dixon et al., 2015; Greenwald et al., 2019; Holwerda et al., 2013; Rao et al., 2014; Tan et al., 2018a). Thus, I started the exploration of allele-specific contacts detected by GAM and their relationship with genes with allele-specific expression at imprinted gene loci.

Imprinted genes often cluster in specific regions of the genome, and differences between the folding of the parental alleles have been observed for several of them. At the *H19*/*Igf2* imprinted gene cluster on chromosome 7, which contains the first described imprinted genes (Barlow et al., 1991; Bartolomei et al., 1991; DeChiara et al., 1991; Ferguson-Smith et al., 1991), many allele-specific contact are detecable, especially on the maternal allele (Figure 4.14a). GAM matrices show that the *H19/Igf2* locus has many maternal-specific contacts with regions further downstream, and selected paternal-specific contacts with upstream regions. The difference in the directionality of these contacts is accompanied by a local drop of interactions over the paternal *H19/Igf2* locus, corresponding to the location of a paternal-specific TAD border at the *H19/Igf2* locus. These differences are not only visible in the allele-specific contact maps, but also amongst the top 5 % of allele-specific contacts on both alleles, showing maternal contacts between the H19 locus and other maternally active genes (Figure 4.14b). The majority of genes in the cluster, including *H19*, are active on the maternal allele (paternally imprinted). Interestingly, on the maternal allele, the *H19/Igf2 locus* interacts with other imprinted genes (*Tnfrsf22*, *Tnfrsf 23*, *Tnfrsf26*). On the paternal allele, where those genes are silenced, this interaction is lost, and the *H19/Igf2* locus interacts with regions further upstream. Generally, almost all imprinted genes have allele-specific contacts in the top 5% monoallelic contacts (Figure 4.14c), however, the specific contacts are not always on the active allele. There are examples where an imprinted gene has more contacts on the silent allele, and others where it has more contacts on the expressed allele. In the future, it will be interesting to perform a thorough exploration of these contacts considering the expression state of the imprinted gene in F123, which might help distinguish these two scenarios, as imprinted genes may engage in different contacts depending on whether they are expressed in mESCs or not.

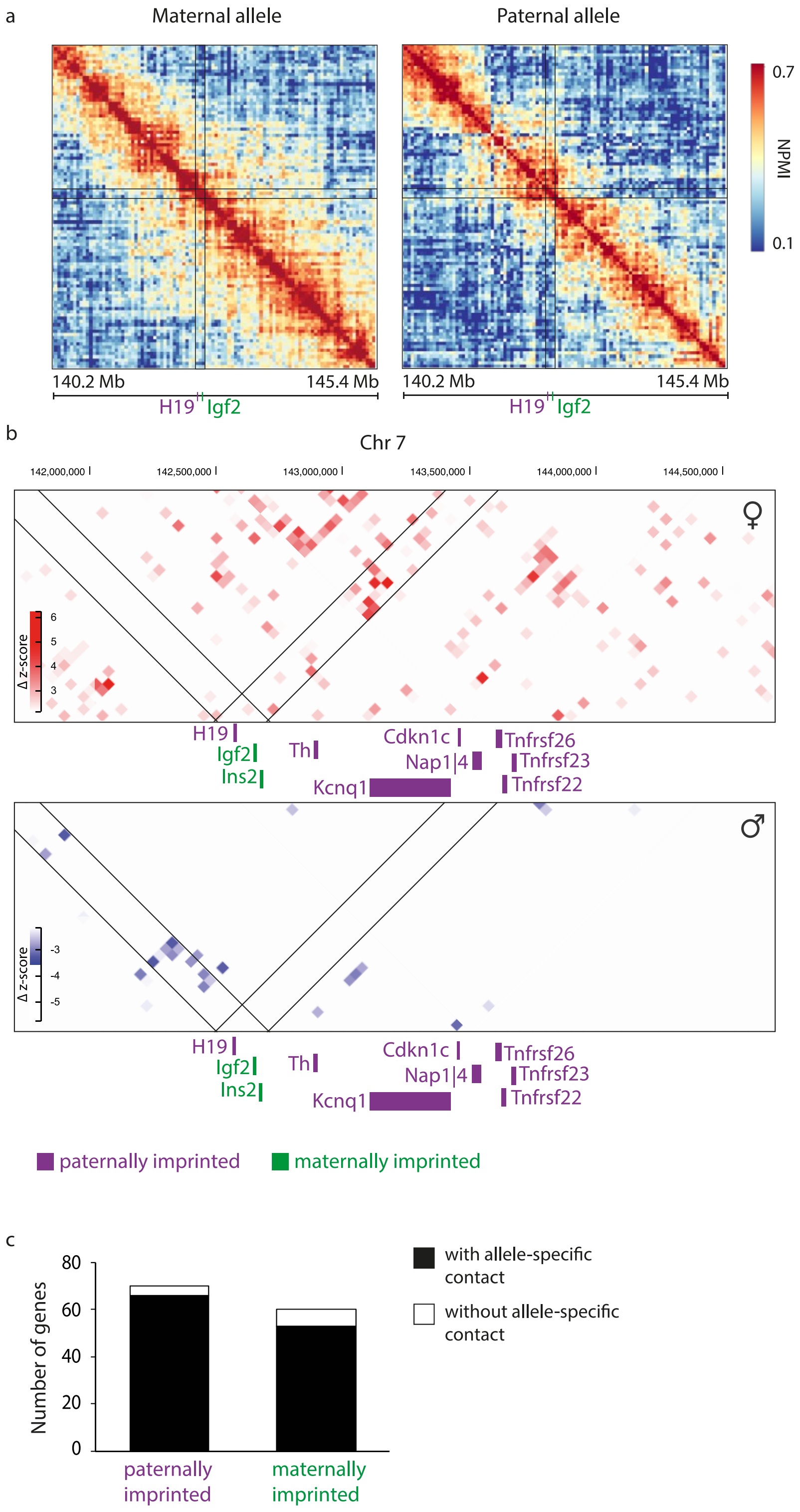


Figure 4. 14: The majority of imprinted genes has specific contacts on one parental allele.
(a) Paternal and maternal contact maps of the imprinted gene cluster on chromosome 7, including *H19* and *Igf2* (b) Maternal- (top) and paternal- (below) specific contacts at the imprinted gene cluster. Cluster contains paternally (purple) and maternally (green) imprinted genes. (c) Number of imprinted genes with and without allele-specific contacts (top 5%).

4.9.2 Allele-specific expression and differential chromatin contacts

With the aim to study a possible connection between allele-specific expression (ASE) and chromatin contacts, I explored different genomic loci containing genes with ASE and found that allele-specific chromatin contacts can be found at many of these loci (examples in Figure 4.15a). To explore the biology underlying these differential contacts, I inspected the linear genome occupancy of the enhancer mark H3K27ac and identified candidate regulatory regions using H3K27ac ChIP-seq data for F123 mESCs (Juric et al., 2019). Allele-specific interactions between genes with ASE and enhancers can be found at different loci. The maternally expressed genes *Pus7* and *Gufl* for example engage in long-range contacts specifically on the maternal allele (Figure 4.15a, top). Conversely, the paternally expressed genes *Snhg14* and *Snrp* (imprinted), or *Col4a1* and *Cars2* interact with enhancer-rich regions specifically on the paternal allele (Figure 4.15a,bottom).

To understand the genome-wide prevalence of enhancer contacts for genes with ASE, I quantified how many enhancer contacts occur at paternally and maternally expressed genes in both allele-specific contact maps (Figure 4.15b). For this comparison, I chose genes with maternal-specific (n = 477), and paternal specific (n = 275) expression, but also genes with allelically balanced expression (n = 3790). Genes were chosen based on the fold change between the read counts mapping to the parental alleles and the p-value (details in chapter 4.4). An enhancer contact was defined by the presence of H3K27ac in the genomic window interacting with a gene with ASE. Furthermore, I investigated contacts for the presence of H3K4me1 (Figure 4.15c), a mark for active and poised enhancers, H3K4me3 (Figure 4.15d), which can be found predominantly at promoters of active genes, and CTCF (Figure 4.15e), an architectural protein enriched at DNA loops. In all three groups of genes, balanced, maternally, and paternally expressed genes, very similar percentages of contacts are found overlapping with the investigated ChIP-seq peaks. Generally, common contacts have a larger overlap with all investigated ChIP-seq peaks compared to the allele-specific contacts. This first exploration of chromatin contacts is limited to active histone marks and CTCF, which are the publicly available ChIP-seq datasets for F123 mESCs. Future work will be extended to the analyses of other histone marks, such as H3K9me3 and H3K27me3, to explore histone marks connected with gene silencing and heterochromatin.

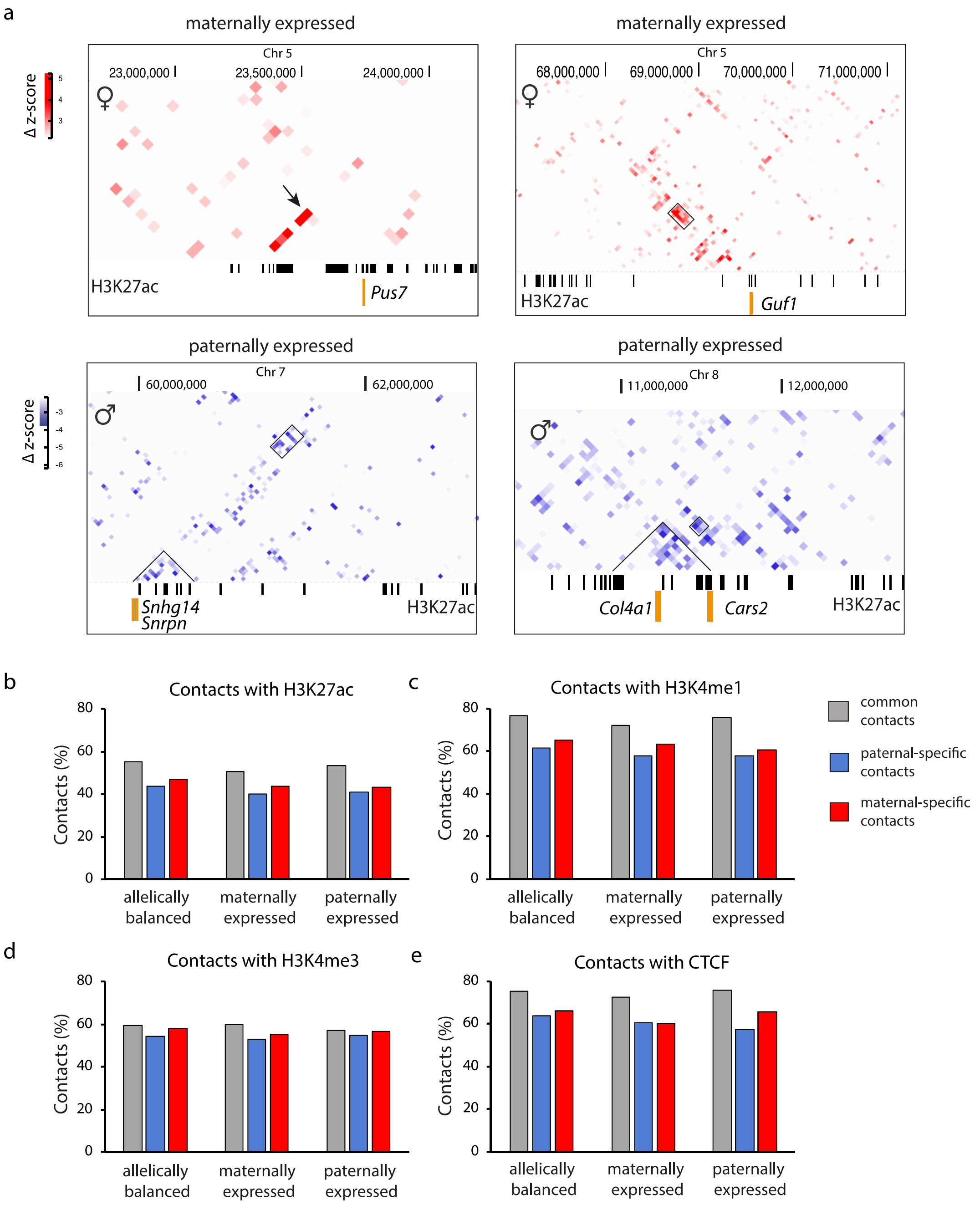


Figure 4. 15: Connecting maternal and paternal chromatin contacts with allele-specific expression. (a) Examples of allele-specific chromatin contacts (5 and 95 % of differential z-scores) of maternally or paternally expressed genes (orange bars) and F123 mESC enhancers defined by the presence of H3K27ac (ChIP-seq peaks, black bars). (b - e) Percentage of contacts (common, paternal-specific, maternal-specific) of allelically balanced, maternally, and paternally expressed genes with different chromatin features. (b) Percentage of contacts with H3K27ac. (c) Percentage of contacts with H3K4me1. (d) Percentage of contacts with H3K4me3. (e) Percentage of contacts with CTCF. Contacts were scored positive for a feature when the underlying genomic bin contained one or more ChIP-seq peaks.

Many studies have described contacts between co-transcribed, active genes, suggesting their co-localisation in active chromatin hubs (Ferrai et al., 2010; Osborne et al., 2004; Osborne and Eskiw, 2008; Schoenfelder et al., 2010b). Further, silenced genes have also been found in repressed chromatin hubs, for instance in Polycomb bodies (Mifsud et al., 2015). Previous work in mESCs showed that active genes preferentially contact other active genes, while the combination of active and inactive genes is less frequent (Beagrie et al., 2017). To test whether this preference can also be observed for allele-specific expression, I investigated how often maternally or paternally expressed genes contact each other (Figure 4.16). In fact, of all possible gene-gene combinations investigated (Figure 4.16a), genes with expression on the maternal allele preferentially contact each other, followed by genes with expression on the paternal allele (Figure 4.16b). Contacts between maternally and paternally expressed genes occur less frequent. This difference was strongest when analysing the common contacts, suggesting that gene-gene contacts of maternal or paternal genes occur predominantly in biallelic configurations.

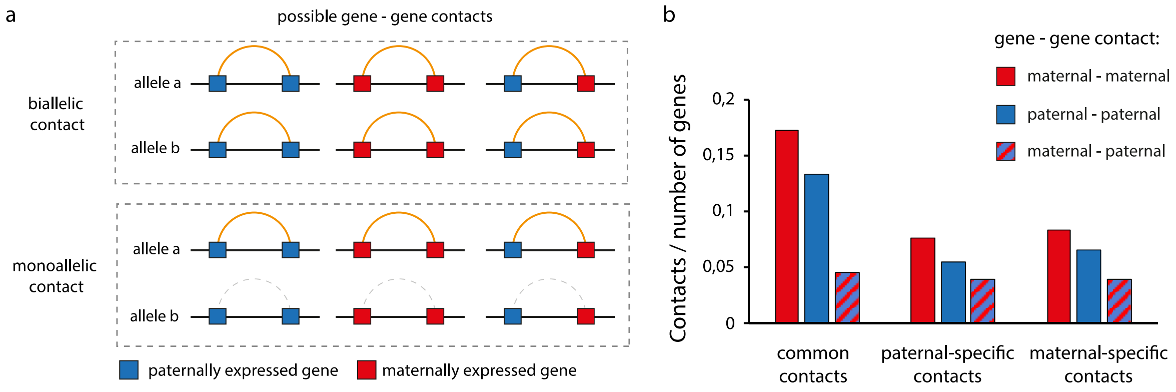


Figure 4. 16: Contacts between genes with allele-specific expression.

(a) Possible gene-gene contacts between maternally or paternally expressed genes. Biallelic contacts are present in the common contacts, monoallelic contacts are either maternal-specific or paternal-specific. (b) Gene-gene contacts present in the common, paternal-specific, and maternal-specific contacts. Contacts are normalised to the number of genes present in each group of genes; maternal – maternal (477 + 477 genes); paternal – paternal (275 + 275 genes); maternal – paternal (477 + 275 genes).

4.10 Transcription factor analysis of common and allele-specific chromatin contacts

Transcription factors (TFs) are key to cellular identity and their binding to regulatory DNA regions determines not only the expression of genes in different cell types, but also corresponds to clustering of occupied TF binding sites in 3D space (Liu et al., 2014; Ma et al., 2018b). Clustering can be observed for regions occupied by the same TF, but also for regions bound by different TFs, of which many have been reported to engage in physical protein-protein interactions (Ma et al., 2018b). Thus, we aimed to investigate which TF binding sites are present at frequent chromatin contacts in F123 mESCs, that are shared by the alleles, or that are specific to the maternal or paternal allele. For the discovery of potential TF candidates, genomic windows engaging in common or allele-specific contacts were filtered for ATAC-seq peaks (ATAC-seq data from Juric et al., 2019) to narrow down the size of the genomic region that is scanned for the presence of TF binding sites to open chromatin regions (Figure 4.17). Then, all TF motifs that were found inside open chromatin regions were filtered for expression of the corresponding TF, as a proxy for potential TF binding. TF binding sites that are enriched in open chromatin regions are analysed for co-localisation in 3D space based on the percentage of chromatin contacts that harbour the two TF motifs. This percentage is expressed as the coverage of a motif pair. Here, we report the top 10 TF motif pairs that are enriched in common or allele-specific chromatin contacts, ranked according to their coverage (Table 4.3). Interestingly, the top TF motif pairs in all three lists of chromatin contacts are combinations of CTCF, SALL1, UBIP1, and MAZ, followed by SMAD3, CTCFL, and KLF15. Seeing that common and allele-specific contacts are enriched for the same TF binding site motifs, encourages the analysis of the binding of these TFs in F123 with respect to allelic differences. Consequently, further analysis of these TFs could include haplotype phasing of ChIP-seq data, to identify regions with differential TF occupancy between the alleles.

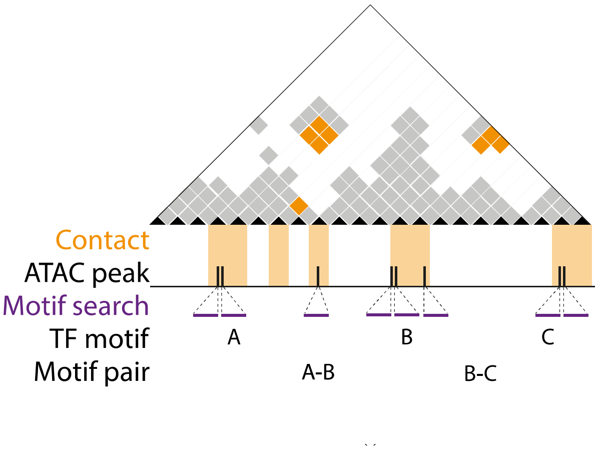


Figure 4. 17: Pipeline to identify transcription factor (TF) motif pairs.
The search for TF binding sites in a set of chromatin contacts (here: common, maternal-specific, paternal-specific) is restricted to ATAC-seq peaks inside genomic bins engaged in specific chromatin contacts. The top TF motifs (A, B, C, ...) are analysed for motif pairs that preferentially contact each other. Enriched TF motif pairs are chosen based on their coverage (percentage of contacts connecting two TF motifs).

Table 4.3: Top 10 transcription factor motif pairs ranked according to their coverage in common, and allele-specific chromatin contacts.

Common contacts		
Rank	Motif pair	Coverage
1	CTCF - SALL1	0.388573
2	UBIP1 - SALL1	0.387422
3	UBIP1 - MAZ	0.382355
4	CTCFL - SALL1	0.38231
5	CTCF - MAZ	0.381977
6	SMAD3 - SALL1	0.381546
7	CTCFL - MAZ	0.376608
8	SMAD3 - MAZ	0.376451
9	UBIP1 - KLF15	0.374472
10	CTCF - KLF15	0.373856

Paternal specific contacts		
Rank	Motif pair	Coverage
1	CTCF - SALL1	0.438536
2	UBIP1 - SALL1	0.437067
3	UBIP1 - MAZ	0.432189
4	CTCF - MAZ	0.431321
5	CTCFL - SALL1	0.430094
6	SMAD3 - SALL1	0.430053
7	SMAD3 - MAZ	0.426974
8	CTCFL - MAZ	0.424006
9	UBIP1 - KLF15	0.424
10	CTCF - KLF15	0.421752

Maternal specific contacts		
Rank	Motif pair	Coverage
1	UBIP1 - SALL1	0.466328
2	CTCF - SALL1	0.46381
3	UBIP1 - MAZ	0.461196
4	SMAD3 - SALL1	0.458978
5	CTCFL - SALL1	0.457502
6	CTCF - MAZ	0.455532
7	SMAD3 - MAZ	0.453117
8	UBIP1 - KLF15	0.452604
9	CTCFL - MAZ	0.449556
10	KLF5 - UBIP1	0.449514

4.11 Conclusions

This chapter addresses allele-specific differences between chromosome topologies and chromatin contacts in F123 mESCs using GAM. Here, I collected three GAM datasets in the hybrid mESC line F123, which comprise a total of 3664 nuclear profiles. The data show high reproducibility between biological replicates and GAM collection modes, which is shown by comparing window detection frequencies and A/B compartments between datasets. By merging all three datasets into one, complete saturation of the data is achieved at a genomic resolution of 20 kb and contacts maps visually appear well detected even at 10 kb resolution, which shows the potential of the F123 data to discover specific chromatin contacts at higher resolution than previously achieved with GAM (30 kb; Beagrie et al., 2017).

Further, this work for the first time distinguishes chromatin contacts between the alleles in GAM, and shows that 80 % of all positive windows in GAM data can be assigned to a haplotype. The efficient phasing of GAM data allows genome-wide explorations of chromatin contacts at 50 kb resolution in a 1123 NP large dataset, with the potential to achieve even higher resolution in the full F123 dataset of 3664 NPs. Comparisons of the allele-specific chromatin contacts revealed genome-wide differences between the alleles for association with the A or B compartment. About 25 % of the F123 genome is present in monoallelic compartments, with equal contributions of the A and B compartment on both alleles. Interestingly, while biallelic compartments have clear and expected preferences for the presence of active and inactive genes in the A and the B compartment, the monoallelic compartments contain equal amounts of active and inactive genes.

Furthermore, allele-specific analysis of RNA-seq data revealed hundreds of genes with allele-specific expression, many of them with housekeeping functions. While those genes were not enriched in monoallelic compartments, but more often found in biallelic A compartments, many of the allele-specific and almost all imprinted genes are engaged in allele-specific chromatin contacts. Finally, to address mechanism for the establishment of common and allele-specific contacts, transcription factors were identified that have the potential to bind to open chromatin engaged in mono- and biallelic contacts. Further work is required to reveal the importance of these transcription factors for common and allele-specific chromosome topologies in mESCs.

5. The effects of the 16p11.2 deletion on chromatin contacts and gene regulation

Note on results and author contributions

The following chapter presents results obtained in collaboration with Dr. Ehsan Irani, Dominik Szabo, Dr. Christoph Thieme, Dr. Marta Slimak-Mastrobuoni, and Julieta Ramirez Cuellar. E. Irani provided an approach for z-score normalisation of Hi-C data, mapped the positions of A/B compartments, and produced differential Hi-C scores. C. Thieme and D. Szabo analysed all RNA-seq data in this chapter, and D. Szabo provided the materials for the figures 5.2b, 5.3 to 5.6 and 5.11. C. Thieme called TAD borders in the Hi-C data, and calculated correlations between Hi-C datasets. M. Slimak-Mastrobuoni and J. Ramirez performed the neuronal differentiation, and M. Slimak-Mastrobuoni generated RNA-seq libraries for two out of three biological replicates. All other experiments and analysis were performed by myself. The three ESC lines used in this chapter were previously produced in the laboratory of Dr. Alea Mills, Cold Spring Harbour Laboratory, Cold Spring Harbour, NY, USA (Horev et al., 2011) and kindly shared by Yon Chang and Dr. Alea Mills.

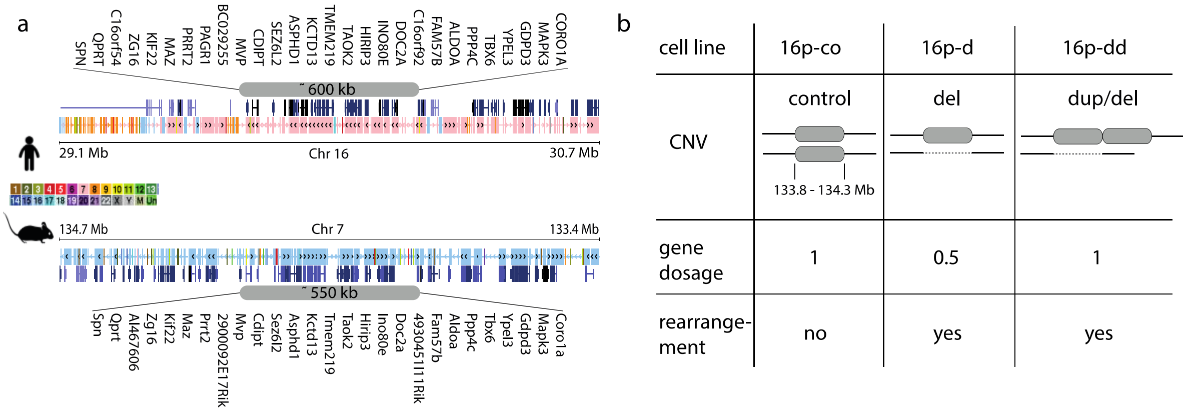
5.1 Introduction and research motivation

The aim of this thesis is to expand the knowledge about chromatin structures and their functionality. Here, I aimed to investigate the impact of chromosomal rearrangements on 3D genome folding and gene regulation, using a model system of congenital disease. The relationship between changes in chromatin topology and disease has so far been investigated mostly at the level of TADs or promoter-enhancer contacts within a few megabases of the disease-associated genomic locus, with few exceptions (Barutcu et al., 2015; Loviglio et al., 2016; Maass et al., 2018a; Zhang et al., 2018a). An intriguing study on chromatin contacts in neurological disorders shows the effects of a ~600 kb chromosomal deletion at the 16p11.2 locus in humans on gene expression and long-range chromatin contacts (Loviglio et al., 2016). CNVs at this locus have been associated with disease, namely a high prevalence of autism spectrum disorder (ASD), increased risk of intellectual disability and psychiatric disorders in carriers of the deletion and the duplication, as well as other phenotypes, such micro- or macrocephaly in the duplication and deletion, respectively (Kumar et al., 2008; Niarchou et al., 2019; Steinman et al., 2016). Changes in long-range contacts involving the 16p11.2 locus were observed both within the same chromosome and between chromosomes

(Loviglio et al., 2016). Although these changes were only mildly connected to gene expression differences, this study raised the possibility that a genomic re-arrangement in one chromosome may affect chromatin folding genome-wide. To investigate the effects of these long-range topology changes on gene expression and better understand the mechanisms of ASD, we decided to develop an in vitro system to study the effects of the 16p11.2 deletion on chromatin contact changes during neuronal development. Further, our in vitro system mimics a developmental stage when genes relevant for ASD are regulated. The cell types affected in ASD are currently not clear, although the complexity of the disease suggest multiple brain regions may be affected; however, several studies suggest a critical role for dopamine signalling in ASD (Anderson et al., 2008; Bariselli et al., 2018; Gadow et al., 2008; Jo et al., 2016; Nakamura et al., 2010; Nguyen et al., 2014; Staal et al., 2012) and dopamine antagonists are approved to treat some behavioural problems associated with ASD (Hellings et al., 2017; LeClerc and Easley, 2015). Thus, we chose to work with a neuronal differentiation system that drives pluripotent cells towards dopaminergic neuron fate (Ferrai et al., 2017). Applied in mouse ESCs, the differentiation undergoes early stages of neuronal fate decisions to derive neuronal precursor cells, and results in premature dopaminergic neurons after 16 days. We obtained mouse ESCs carrying the 16p11.2 locus CNV that is most common in humans and that phenocopied the human disease in mice (Horev et al., 2011). By studying gene expression and chromatin topology in the context of the 16p11.2 deletion, we aim to understand which genes are misregulated in neuronal development in ASD, and how this misregulation is connected to topology changes in the genome. We next used a genome-wide approach to investigate chromatin contacts connected to changes in gene expression in ASD. This approach seeks to gain insight into the functionality of long-range and inter-chromosomal contacts, a topic which is still highly debated in the field.

5.2 Experimental design

In order to study chromatin folding in the 16p11.2 deletion, we obtained mouse ESC lines which contain rearrangements in 7qF3, the mouse region syntenic to human 16p11.2, which were previously generated in the laboratory of Dr. Alea Mills (CSHL, Cold Spring Harbour, NY, USA) starting from ESCs generated from wildtype C57BL/J6 mice, and used to generate mouse models of the 16p11.2 deletion (Horev et al. 2011). The human 16p11.2 deletion is located in a genomic region which is conserved between human and mouse genomes. The equivalent murine region spans 550 kb and contains 27 protein-coding genes, which comprise a consecutive synteny block shared between humans and mice (Figure 5.1a). The high gene density in the deletion region raises some additional challenges in studying broader structural effects of the 16p11.2 deletion in gene regulation, because of the effect of gene dosage changes in a large number of protein-coding genes in the deletion region. Thus, we considered three different murine ESC lines described in Horev et al. (2011). First, the *df/+* mESC line that contains the heterozygous 16p11.2 deletion; this line is referred to as 16p-deletion (or '16p-d') in this thesis. Second, we chose a mESC line (*df/dp*) containing both the 16p11.2 deletion and the corresponding duplication (here referred to as '16p-dd' ESC line), with the aim of balancing out gene dosage effects (Figure 5.1b). The 16p-dd ESC line carries the 16p11.2 deletion on one allele, and the duplication of the 16p11.2 region on the other allele. If gene dosage would be effectively recovered in the 16p-dd line, it would make it possible to discover alterations in gene expression due to the structural rearrangement. Last, we used the wildtype ESC line as a control line; here referred to as '16p-co'. This complex experimental design explores gene expression differences between the wildtype cell and the heterozygous deletion (16p-d), and searches for matching gene expression differences between the control and duplication-deletion (16p-dd) cells. This approach allows us to specifically investigate changes in gene expression that are likely to be due to the large 16p11.2 genomic deletion, rather than from halving gene dosage. Overall, this should help identify genes that change solely because of downstream effects of the reduced expression of genes within the deletion region. It should be noted, that our experimental design does not aim to ignore gene dosage effects of the 16p11.2 deletion, which are also likely to contribute to the autism phenotypes of the 16p11.2-deletion carriers. While the effects of gene dosage include an interesting set of target genes for autism research, however they are not the focus of this study. At this time, only three of the genes within 16p11.2, *Taok2*, *Mapk3*, and *Kctd13*, have been assigned to the SFARI list of ASD genes (https://www.sfari.org). In this thesis, I retain the human nomenclature of the 16p11.2 rearrangement, even though it is modelled in the 7qF3 mouse region, as is standard in this field of research (Horev et al., 2011).



cell line	16p-co	16p-d	16p-dd
CNV	control	del	dup/del
gene dosage	1	0.5	1
rearrange-ment	no	yes	yes

Figure 5. 1: Copy number variations (CNVs) at the 16p11.2 locus.
(a) Genes and synteny in the human 16p11.2 deletion region are conserved between human and mice. Similarity between human and mouse genomes and corresponding gene tracks are shown, colour coded by chromosome. The 16p11.2 locus is syntenic to a genomic region in the mouse 7qF3 locus. (b) The 16p11.2 cell lines used in this project carry different CNVs at the 16p11.2-syntenic murine locus on chromosome 7.

5.3 Neuronal differentiation of the 16p11.2 cell lines

To investigate the effects of the 16p11.2 deletion in ASD pathology, we chose to investigate changes in gene expression and 3D chromatin structure, not only in pluripotency but also during neuronal differentiation. We took advantage of a reliable neuronal differentiation system that was previously established in our lab, which generates a homogeneous population of neuronal precursor cells (NPCs) and dopaminergic neurons (Figure 5.2a), by manipulating FGF and Shh signalling (Ferrai et al., 2017). First, we collected RNA and Hi-C-ready chromatin in the three ESC lines (co, d and dd). Second, the three ESC lines were differentiated into stable EpiStem cell (EpiSC) lines, which can be amplified and stored. The EpiSCs were then induced to differentiate by blockage of FGF/ERK signalling and differentiated towards the dopaminergic fate by stimulation with FGF and Shh (Ferrai et al., 2017). Five days after induction, the cells differentiate into NPCs; at this time, we collected RNA and Hi-C-ready chromatin. Finally, at day 16, we collected only RNA. The differentiation of all three 16p11.2 ESC lines was performed in several replicates, and showed no consistent differences in pluripotency levels, growth rate, or differentiation, based on the morphology of the cells (Figure 5.2a), or protein levels of stage markers validated by RT-qPCR and immunofluorescence, respectively (J. Ramirez and M. Slimak-Mastrobuoni, personal information).

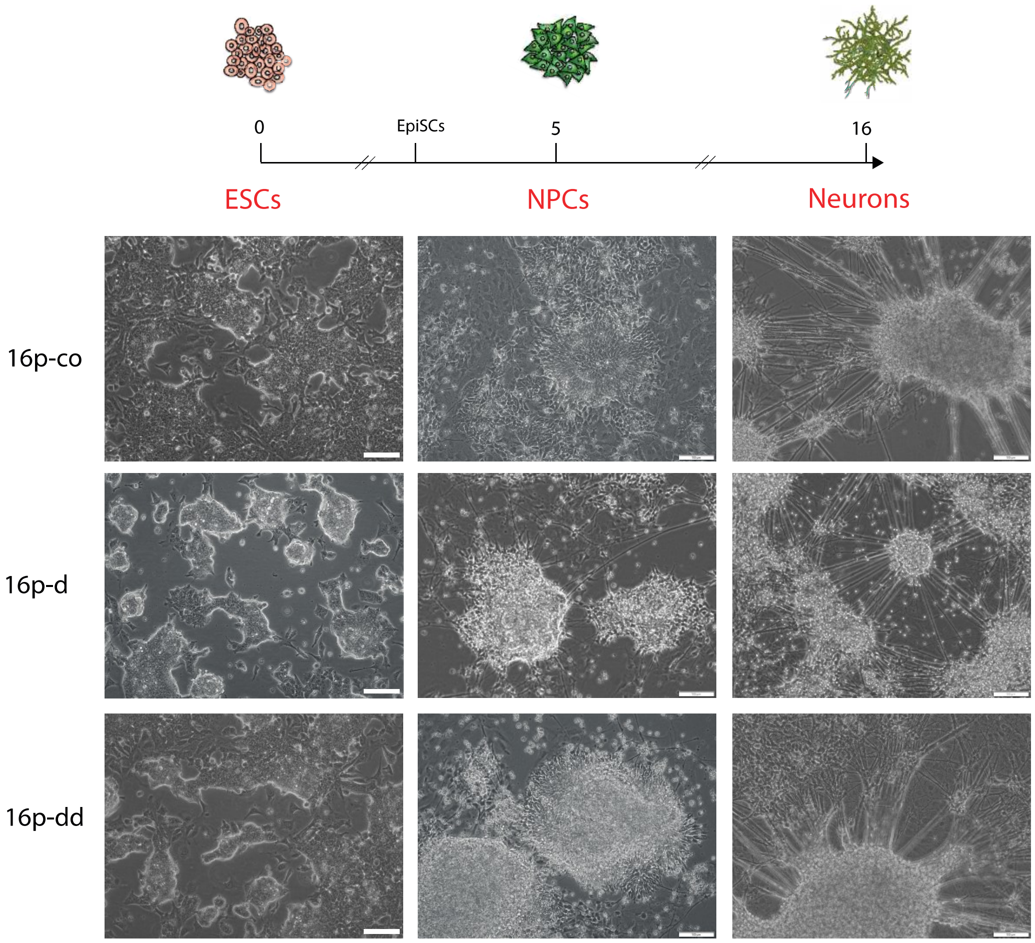


Figure 5. 2: Successful differentiation of 16p11.2 cell lines.
Morphology of 16p-co, 16p-d, and 16p-dd at different stages of the neuronal differentiation; embryonic stem cells (ESCs), neuronal precursor cells (NPCs), and premature dopaminergic neurons (Neurons). Scale bar: 100 µm.

5.4 RNA-seq libraries are highly similar between replicates and cell lines and cluster into cell types

To investigate whether the 16p11.2 deletion or deletion-duplication had major effects on the cell state and to identify changes in gene expression, we produced total RNAseq libraries for the 16p11.2 lines (16p-co, 16p-d, and 16p-dd) in several replicates, for ESCs, NPCs (day 5), and neurons (day 16) (Figure 5.3a). Comparison of expression levels show that biological replicates correlate well (e.g. NPC/16p-co, replicate 1 and 2, r = 0.957; Figure 5.3b), which can be confirmed for all cell lines (data not shown), indicating little variability between the neuronal differentiations. Further, principal component analysis (PCA) of the first two principal components (explaining 80 % of the variance between samples) shows that all experiments cluster according to their differentiation stage, regardless of the CNV of the cell line. We also analysed the expression of differentiation stage markers (Figure 5.3d). As expected, the expression of pluripotency markers *Pou5f1* (Oct4) and *Nanog* is high in ESCs, showing large similarity between biological replicates and 16p11.2 lines. In NPCs, *Nanog* and *Pou5f1* expression could no longer be detected, while the expression of early neuronal markers, such as *Hes5*, *Nes*, and *Tubb3* is detected. *Nes* expression is no longer present in neurons, while *Hes5* expression peaks in NPCs, and is still present in neurons. *Tubb3* starts being expressed in day 5, and its expression is highest in day 16. Lastly, neurons express the dopaminergic marker *Th*, indicating the commitment of the neurons towards the dopaminergic fate. The expression of stage markers observed in the three 16p11.2 lines are equivalent to the expression observed for other mESC lines (e.g. ESC-46C; Fraser et al., 2017), and confirms successful differentiation of all three cell lines to premature dopaminergic neurons. These results show that the 16p11.2 deletion and deletion-duplication do not alter the broad signature of the RNA expression in cells undergoing neuronal differentiation, neither the time nor efficiency of differentiation from pluripotency to NPCs and early dopaminergic neurons.

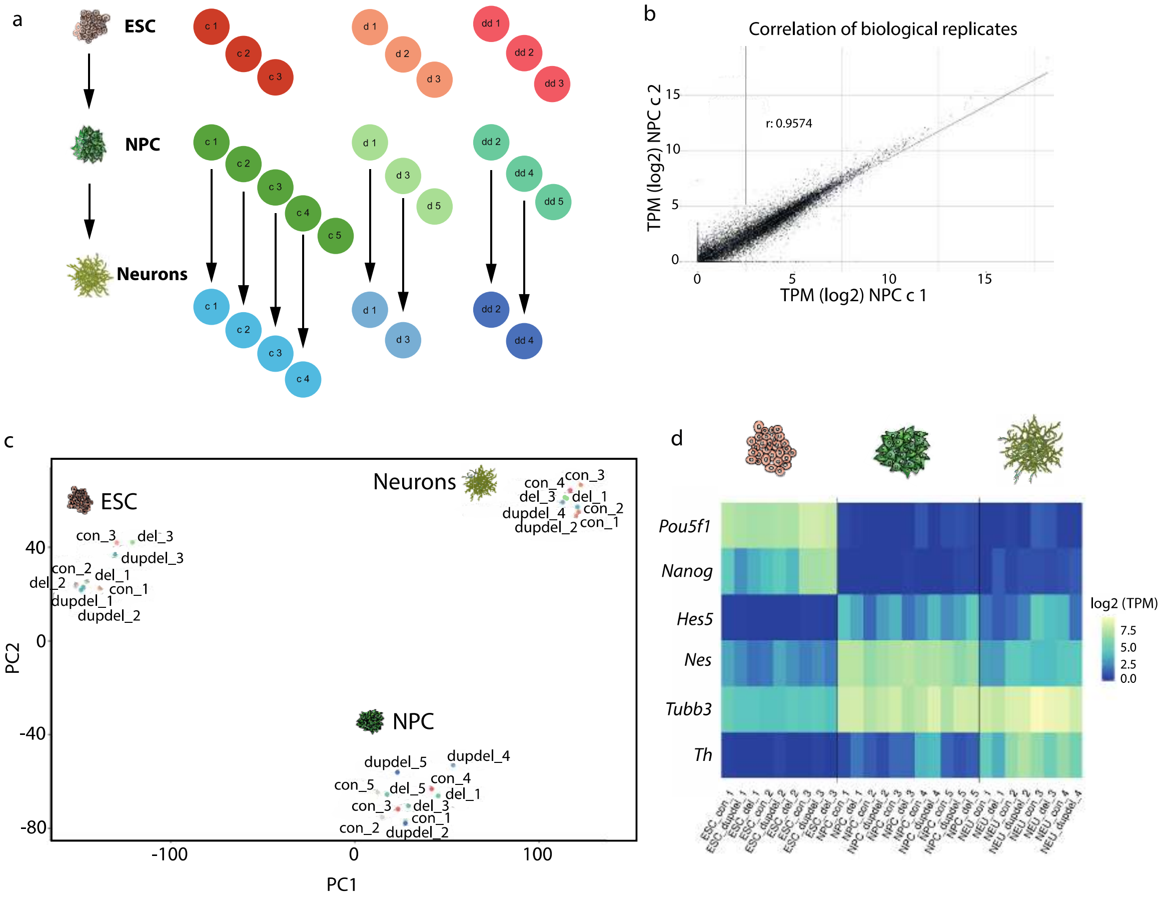


Figure 5. 3: RNA-seq libraries show large similarity between replicates and cell lines.
(a) Libraries are collected in biological replicates at three stages in the neuronal differentiation (ESC, NPC, neuron). Each differentiation of cells carrying the deletion (d1, d3) or duplication-deletion (dd2, dd4) was accompanied by the wildtype control (c1, c3 and c2, c4), respectively. (b) Correlation of TPM values (log2) between biological replicates of NPC 16p-co (c1, c2). (c) Principal components (PCs) 1 and 2, explaining 62%

(PC1) and 18% (PC2) of variance between samples, shows clustering of samples according to their differentiation stage, but not according to their cell lines. (d) Expression (RNA-seq, log2(TPM)) of pluripotency marker genes (*Pou5f1, Nanog*), NPC marker genes (*Hes5, Nes, Tubb3*), and marker genes of dopaminergic neurons (*Tubb3, Th*) in 16p-co, 16p-d, and 16p-dd at different stages of neuronal differentiation.

5.5 Differential RNA-seq analysis at the 16p11.2 locus confirm dosage effects in the deletion, and dosage compensation in the duplication-deletion

To investigate gene expression changes induced by the 16p11.2 deletion that are independent of the 0.5x gene dosage, we searched for changes in gene expression that were also found in the 16p-dd line. To this end, we analysed the gene expression differences between the 16p11.2 control cell line (16p-co), as well as the line carrying the deletion (16p-d) and the one carrying the duplication-deletion (16p-dd), and searched for common genes with altered expression. To determine gene expression changes above the internal fluctuations within each type of sample, we computed significant gene expression changes within the same timepoint in the different cell lines by considering the variability in transcript levels between all biological replicates within each cell line.

First, we started by investigating the levels of gene expression at the 16p11.2 locus itself, and specifically whether the 16p-dd line, which has both copies of the 16p11.2 locus in one allele, effectively compensated dosage. First, we compared the transcript levels of genes within the 16p11.2 deletion and their immediate neighbours in the genome (Figure 5.4). As expected, the genes within the 16p11.2 deletion showed significantly reduced gene expression in the 16p-d cell line compared to wildtype. This reduction could be observed in all three timepoints, ESCs, NPCs, and neurons. In the 16p-dd cells, the reduced gene expression due to the deletion on one allele could indeed be compensated by the duplication of the whole 16p11.2 deletion on the other allele for two timepoints, ESCs and NPCs. In day 16 neurons, the dosage compensation did not fully restore the wildtype transcript levels of the genes at the 16p11.2 locus; in fact, nine genes inside the 16p11.2 region showed reduced expression compared to the NPC control in the 16p-dd line. These results show that the 16p-dd cells can be considered as a control for dosage compensation in ESCs and NPCs, but cannot fully compensate for gene dosage in day 16 neurons. Thus, we decided to focus our analysis of differential gene expression and their connection to genome topology mostly on the earlier timepoints in the differentiation, ESCs and NPCs.

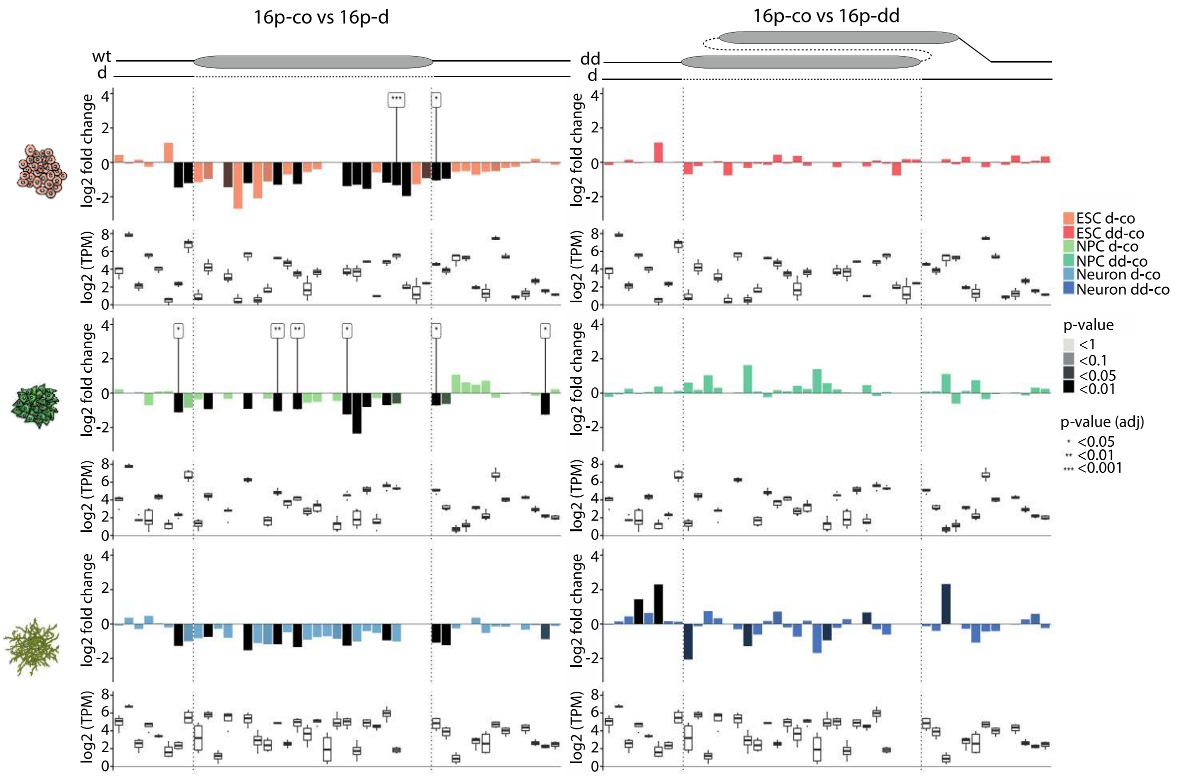
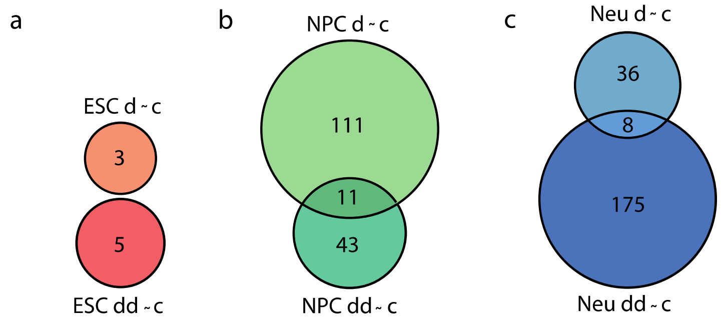


Figure 5. 4: Gene expression differences between control and mutant cell lines at the 16p11.2 locus. Gene expression differences (log2 fold change) of the 16p-d and the 16p-dd TPM values compared to 16p-co TPMs. Under each graph, gene expression (log2 TPM) of 16p-co is shown. Inside the 16p11.2 deletion 24 genes are expressed (genes between dotted lines), which were downregulated in 16p-d in ESCs, NPCs, and neurons. In 16p-dd, ESCs and NPCs genes inside the deletion show only minor differences in expression compared to the wildtype, but are slightly downregulated in neurons.

5.6 A small number of differential expressed genes are common to NPCs from the 16p-d and 16p-dd

To find genes that are differentially regulated due to the 16p11.2 deletion and not to gene dosage, we intersected the lists of differentially expressed genes in the 16p-d and the 16p-dd line in ESCs, NPCs and neurons (Figure 5.5). We found only 3 and 5 differentially expressed genes in the ESC 16p-d and 16p-dd lines, respectively, with none of them being shared, suggesting that the 16p11.2 deletion has little influence on gene regulation in pluripotency, even though most of the genes it contains are expressed (Figure 5.4). However, in NPCs, the number of differentially expressed genes increases to 123 and 54 in 16p-d and 16p-dd, respectively, with 11 common genes. Remarkably, none of the 11 overlapping genes were located on chromosome 7, which carries the 16p11.2 deletion. In neurons, the number of differentially expressed genes was highest in the 16p-dd line, suggesting that not only the incompletely compensated gene dosage, but the duplication itself contributes to a large number of differentially expressed genes. The more complex results in neurons led us to focus our next analysis in the NPC timepoint.



d Overlap of differentially expressed genes in 16p-d and 16p-dd

NPCs		**Neurons**	
Chr 1	*Col5a2*	Chr 2	*Rbm43, 4930412013Rik*
Chr 5	*Zfp951*	Chr 4	*Penk*
Chr 14	*Rnf17, Ak157947*	Chr 6	*Parp12*
Chr 15	*Mtmr12*	Chr 14	*Rnf17*
Chr 18	*Pcdhb* : 16,17,19,21,22	Chr 18	*Pcdhgb2*
Chr X	*Gm5124*	Chr X	*Gm5124, Cldn34c1*

Figure 5. 5: Cells carrying the 16p11.2 deletion have only a small number of differentially expressed genes that are shared with their dosage-compensated control.
Number of significant differentially expressed genes (adjusted p-value < 0.05) in 16p-d and 16p-dd compared to 16p-co in (a) ESCs, (b) NPCs, and (c) neurons (NEU). Venn diagrams show the overlap of differentially expressed genes in the two mutant cell lines. (d) List of all differentially expressed genes shared by 16p-d and 16p-dd in NPCs and neurons.

5.7 Most gene expression changes common in 16p-d and 16p-dd in NPCs are located on chromosome 18

Surprised by the identification of differentially expressed genes in other chromosomes that are common to the 16p-d and 16p-dd NPCs, we asked whether these genes had a preferential genomic location (Figure 5.6). First, we confirm that ESCs have little changes in gene expression both in 16p-d or 16p-dd relative to wildtype 16p-co. In NPCs, there are extensive changes in gene expression found across all chromosomes, especially in 16p-d. A small number of changes are conserved, and these are found in chromosomes 1, 5, 14, 15, 18 and X. We were most interested in the changes found in chromosome 18, all of which were protocadherin genes found in a single genomic region. Unexpectedly, none of the significant gene expression changes in chromosome 7 where common in 16p-d and 16p-dd NPCs. Finally, we found that the 16p-dd leads to stronger gene deregulation in neurons, and affecting genes found across all chromosomes, including many in chromosome 7, where the deletion is located.

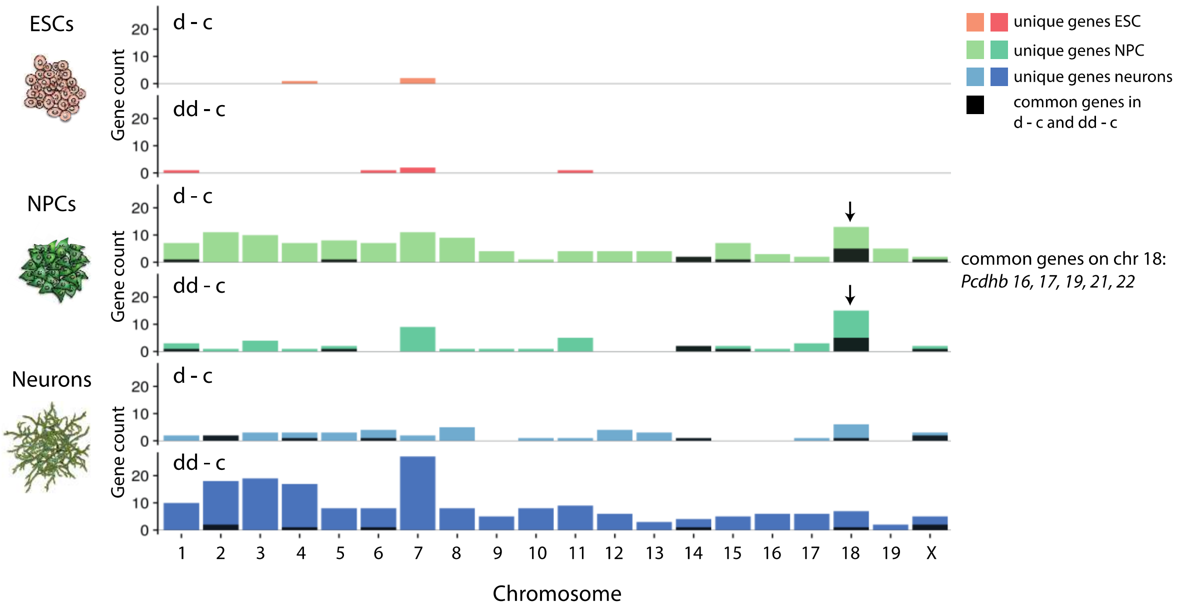


Figure 5. 6: Differentially expressed genes shared by 16p-d and 16p-dd are mostly found on chromosome 18.

Number of differentially expressed genes per chromosome of 16p-d, and 16p-dd in comparison to the control cell line 16p-co in ESCs (top), NPCs (middle), and day 16 neurons (bottom). Common genes between 16p-d and 16p-dd are indicated in black.

5.8 Mapping chromatin contacts in 16p11.2 cells

The aforementioned observations of differentially expressed genes both within chromosome 7 and across the genome, particularly in chromosome 18, raised the possibility that the deletion leads to broader alterations in genome organisation. To explore the 3D genome folding of the 16p11.2 cells, we mapped chromatin contacts genome-wide using Hi-C, which identifies pairwise chromatin contacts via proximity-ligation.

First, I established the Hi-C protocol in the ESC line 46C, which is commonly used in our laboratory, by exploring a range of parameters described in the published protocols (Fraser et al., 2015; Lieberman-Aiden et al., 2009). Important protocol steps that required adjustment were the concentration of the restriction enzyme and incubation time of the restriction digestion, the composition of the ligation buffer, the temperature and timing of the biotin removal step, and the conditions of adapter ligation and amplification of DNA from streptavidin beads (see Methods chapter 2.11 for details). I performed quality controls at all stages of the protocol, as suggested in Belton et al. (2012), and optimised the protocol to match the published recommendations for digestion efficiency (Figure 5.7a) and efficiency of biotin incorporation (Figure 5.7b). Library preparation was adjusted according to these efficiencies, as well as effective DNA extraction, sonication, and size selection of DNA fragments of 300 – 600 bp (Figure 5.7c). All adapter ligated and amplified libraries were run on a high sensitivity Bioanalyzer to determine the average fragment size of the Hi-C library and to check for uniform amplification of all fragment sizes (Figure 5.7d).

After sequencing, I mapped reads to the mm9 reference genome and analysed them with TADbit (Serra et al., 2017) to determine the percentage of reads that contain ligation events between two restriction fragments, or in other words, valid Hi-C pairs. When filtering reads for valid Hi-C pairs, additional classes of reads can be identified, such as dangling ends (suggesting inefficient ligation or biotin-end removal), self-circles (suggesting a low concentration of crosslinked chromatin fragments in the ligation reaction), or PCR duplicates (a result of low amounts of input material in the library preparation and a general hallmark of low quality sequencing libraries). I compared the efficiency of my protocol with published Hi-C data in mESCs, which at the time was considered as the gold standard for Hi-C (Dixon et al., 2012). With my optimised Hi-C protocol, I reached almost identical quality metrics compared to Dixon et al. (Figure 5.7e). The Hi-C data was then ICE-normalised using TADbit and matrices were plotted for visual inspection and comparison with the published matrices from Dixon et al. (2012). Overall, Hi-C contact frequencies in 46C mESCs closely resembled the contact frequencies detected in Dixon et al. (2012) (Figure 5.7f, g).

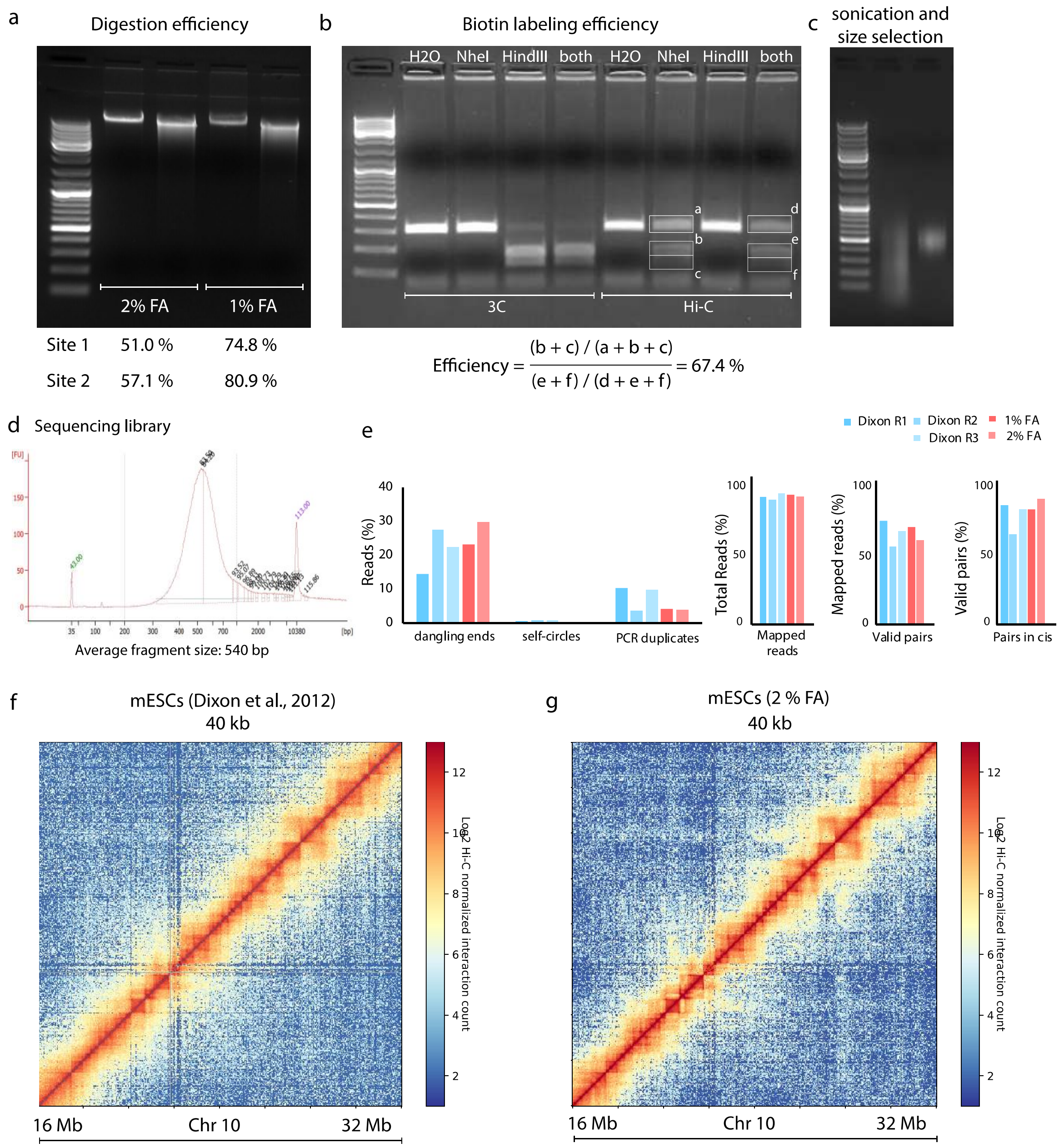


Figure 5. 7: Set-up of the Hi-C protocol and quality controls.

(a) Undigested and digested Hi-C libraries, with the respective digestion efficiencies for two genomic regions below. Digestion efficiencies have been estimated by RT-qPCR. 2% formaldehyde (FA) fixation results in reduced digestion efficiencies compared to 1% FA. (b) Assay to determine biotin labelling efficiency. Amplified ligation product from two adjacent HindIII restriction fragments is digested with HindIII and with NheI. Fragments without biotin incorporation carry the HindIII restriction site, fragments that incorporated biotin carry a NheI restriction site. The efficiency of incorporation is calculated from the proportion of DNA that can be digested with NheI compared to HindIII. (c) Sonication and size selection of a purified Hi-C library results in fragments of 300 – 600 bp in size. (a – c) DNA is loaded on a 1.2% agarose gel, next to the Gene Ruler 1kb plus ladder. (d) Electropherogram showing the fragment size distribution of a Hi-C sequencing library with an average fragment size of 540 bp. FU, fluorescence units. (e) Percentage of reads with dangling ends (unligated fragments), self-circles (fragments ligated to themselves), and PCR duplicates in Hi-C libraries from Dixon et al., 2012 (blue) and libraries produced with 1% or 2% formaldehyde fixation (pink). Below, reads after mapping, filtering, and assigning them as *cis* reads (reads within a chromosome), all as percentage of the previous step, mapped reads as percentage of total reads. (f) ICE normalised Hi-C ligation frequencies in mESCs from Dixon et al., (2012) and (g) 46C mESCs (2% FA).

Next, I applied Hi-C on ESCs and NPCs of the three 16p11.2 cell lines, 16p-co, 16p-d, and 16p-dd, resulting in six Hi-C libraries. Mapping and filtering showed good quality metrics for all libraries (Figure 5.8a, b). HiCRep (Yang et al., 2017) was used to compare Hi-C datasets and check for similarity between the different lines. The pipeline determines the correlation between Hi-C data, while giving weight to different genomic length scales, a process that makes the comparison of datasets more accurate and sensitive to differences between samples, which is often not given when calculating the Pearson correlation between datasets. With HiCRep higher correlations between libraries of the same timepoint can be observed, compared to libraries from different timepoints (Figure 5.8c), similar to what we observed in the matching RNA-seq data.

a

Name	Sample	Genome	Number of cells	Enzyme	Reads (total)	Mapped reads	Valid pairs
16pco ESC	ESC, control	mm9	3.07x10^7	HindIII	248925802	219647200	100785734
16pd ESC	ESC, deletion	mm9	2.8x10^7	HindIII	290687736	248272353	140248606
16pdd ESC	ESC, duplication-deletion	mm9	2.8x10^7	HindIII	259862977	224830175	121986344
16pco NPC	Day 5 NPC, control	mm9	2.6x10^7	HindIII	367511951	333835460	174936591
16pd NPC	Day 5 NPC, deletion	mm9	3.4x10^7	HindIII	436759673	396743558	201902805
16pdd NPC	Day 5 NPC, duplication-deletion	mm9	2 - 3 x10^7	HindIII	250193553	215637289	94388241

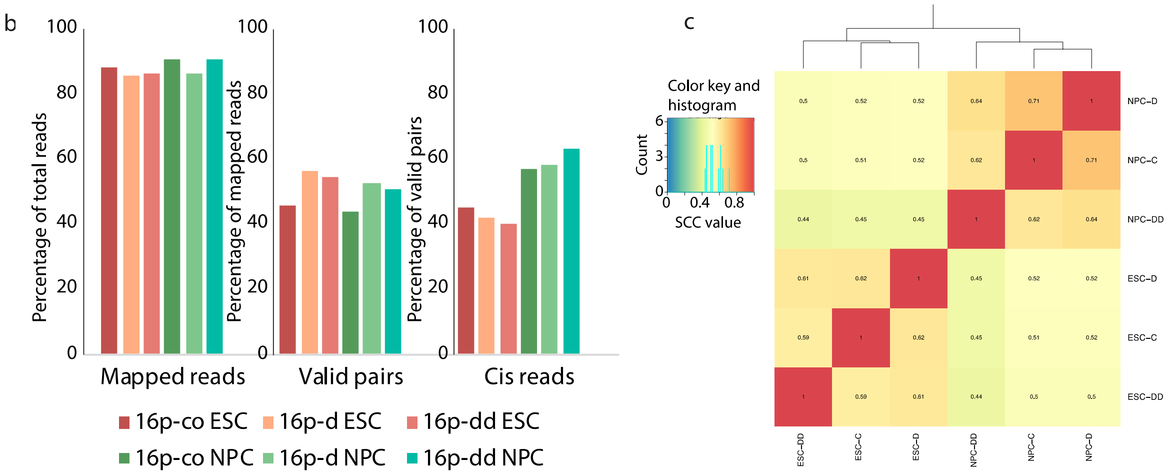


Figure 5. 8: Comparable quality metrics between 16p11.2 Hi-C datasets.
(a) Hi-C data collected in the three 16p11.2 cell lines. (b) Mapped reads in percentage of total reads, valid read pairs after filtering, and *cis* reads in percentage of valid pairs. (c) Spearman's correlation coefficient (SCC) of the 16p11.2 Hi-C datasets, calculated with HiCRep.

5.9 Chromatin contacts around the 16p11.2 locus are largely unchanged between cell lines

Next, I studied the chromatin contacts at the 16p11.2 locus and neighbouring regions. Interestingly, these analyses show that the 16p11.2 region corresponds to a single TAD in

ESC and in NPCs, and together with the neighbouring regions there are no noticeable major changes in TAD organisation between the two differentiations stages (Figure 5.9a, b).
To explore the impact of the deletion on chromatin topology, I compared the Hi-C contact frequencies at the 16p11.2 locus between 16p-co and 16-d or 16p-dd in both ESCs and NPCs (Figure 5.9a, b). In 16p-co, the wildtype cells, ESCs and NPCs show highly similar contact frequencies. Noteworthy, the 16p11.2 deletion is found inside compartment A of chromosome 7 in NPCs (Figure 5.9b; analyses of compartments are ongoing in ESCs). The comparison with 16p-d and 16p-dd shows that overall contact frequencies do not show noticeable changes, except for increased contact frequencies between the flanking regions of the 16p11.2 deletion, an expected result of reducing the genomic distance between these regions. The same effect is visible in the 16p-dd data, but the difference to the 16p-co line appears less pronounced, potentially balanced out by reduced contacts between these regions on the allele carrying the duplication. To quantify these observed differences, we subtracted Hi-C contact frequencies after z-score normalisation of either the 16p-d line or the 16p-dd line from the control 16p-co (Figure 5.9c, d). Indeed, the subtraction revealed only a few contact changes, locally and chromosome wide. The most prominent change occurs at the 16p11.2 locus in NPCs, where chromatin contact frequencies are clearly deleted for contacts with the deletion and increased upstream and downstream of the deletion in 16p-d and 16p-dd. However, these changes do not alter the positions of TAD borders at the locus nor do they affect the association with the A compartment. Further, the genes at the locus that significantly change their expression in 16p-d (*9130019O22Rik* and *Tgfb1i1*) and in 16p-dd (*Zfp764*) do not show specific contact changes in the respective Hi-C data.

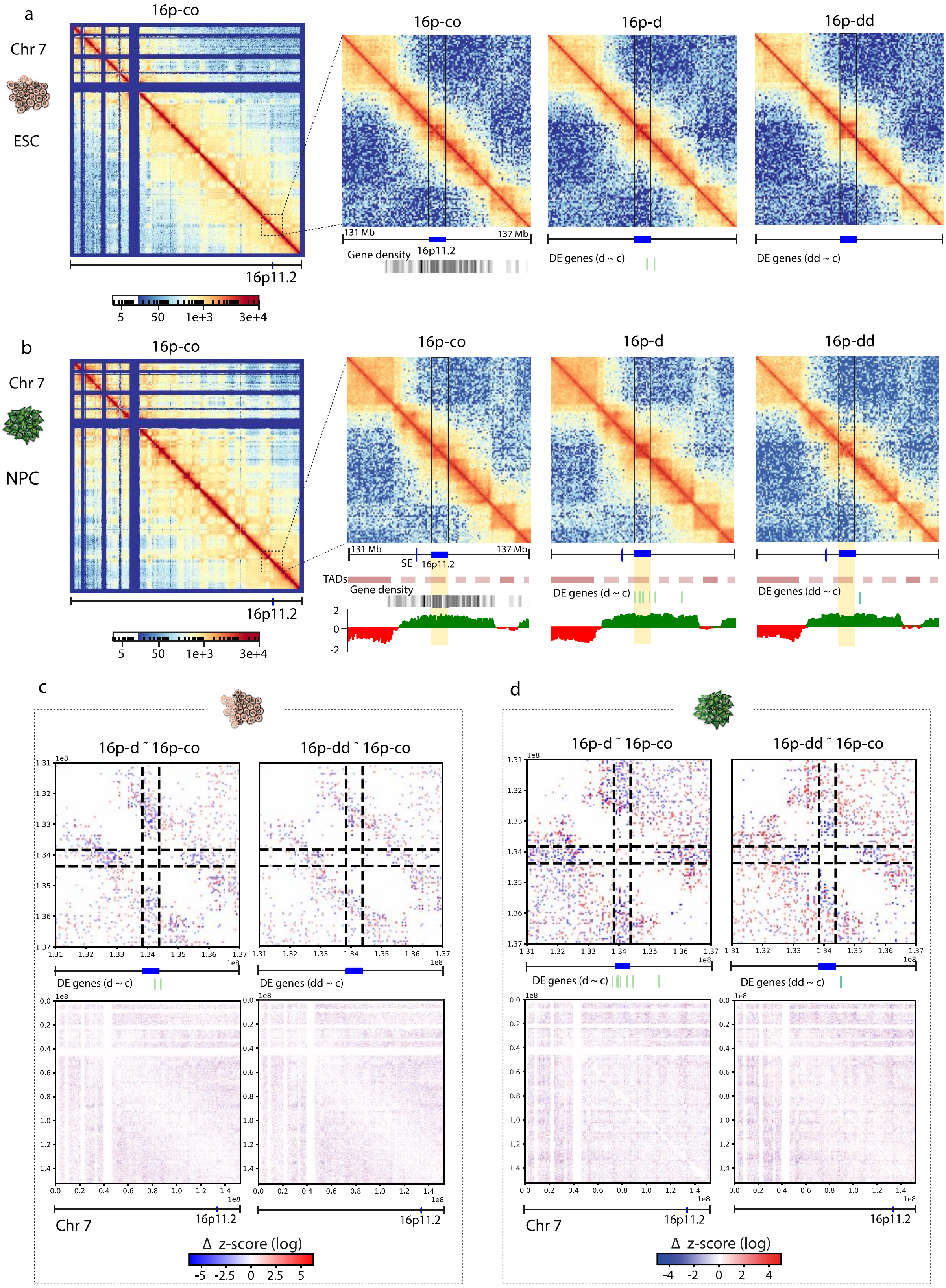


Figure 5. 9: Chromatin contacts around the 16p11.2 locus change only marginally between cell lines. (a) Hi-C contact frequencies (log2) of chromosome 7 in 16p-co and the zoom to the genomic locus carrying the 16p deletion in 16p-co, 16p-d, and 16-dd in mESCs. Below: genomic position of the 16p11.2 deletion, gene density, and positions of differentially expressed genes. (b) Hi-C contact frequencies (log2) of chromosome 7 in 16p-co and the zoom to the genomic locus carrying the 16p11.2 deletion in 16p-co, 16p-d, and 16-dd in NPCs.

Below the zoom to the 16p11.2 locus, from top to bottom: Genomic position of E14.5 brain super-enhancers, genomic position of the 16p11.2 deletion, position of TADs, gene density (shown in 16p-co), differentially expressed genes (shown in 16p-d, and 16p-dd), A (green) and B (red) compartments. (c) Delta z-score of Hi-C contacts in ESCs for 16p-d minus 16p-co (left), and 16p-dd minus 16p-co (right). Top matrices show differential contacts of the 16p11.2 locus, bottom matrices show differential contacts of chromosome 7. (d) Delta z-score of Hi-C contacts in NPCs for 16p-d minus 16p-co (left), and 16p-dd minus 16p-co (right). Top matrices show differential contacts of the 16p11.2 locus, bottom matrices show differential contacts of chromosome 7.

5.10 The 16p11.2 locus contact other regions on chromosome 7 with no direct connection to gene expression changes

Although most contacts on chromosome 7 are largely unchanged in both timepoints, we observed significant differential gene expression in NPCs for a number of genes on the same chromosome. These genes were not consistent between the deletion and the dosage compensated control, however, in order to understand if these changes are connected to the 16p11.2 locus, we explored the chromatin topology of chromosome 7 and the contacts of the 16p11.2 region with the remaining chromosome 7 in NPCs. The 16p11.2 deletion comprises a gene-dense, active genomic region, located in the A compartment of the chromosome (Figure 5.9b). Interestingly, a large proportion of the differentially expressed genes of 16p-d and 16p-dd on chromosome 7 are located in the B compartment (Figure 5.10a). Further, the positions of the A and B compartments on chromosome 7 were largely unchanged between 16p-co, 16p-d, and 16p-dd (Figure 5.10a). To explore the intrachromosomal contacts established by the 16p11.2 locus, we used z-score normalised Hi-C contacts and extracted the contacts of 16p11.2 with chromosome 7. The z-score normalisation takes into account for every genomic distance the average contact frequencies of the chromosome as a background signal. Contact frequencies of a specific region of interest are normalised to this background signal. Consequently, z-score Hi-C tracks (Figure 5.10b) show distance-normalised depletion or enrichment of contacts between a genomic region and the remaining chromosome compared to the background signal of that chromosome. Interestingly, the 16p11.2 deletion and its surrounding genomic regions (+/- 500 kb) establish long-range contacts with other genomic regions from compartment A in chromosome 7, while they are depleted for contacts with their immediate surrounding (Figure 5.10b, viewpoint 16p, viewpoint 16p - 500 kb). These long-range contacts appear as distinct patches along the chromosome, a characteristic that cannot be observed for regions further up- or downstream of the 16p11.2 locus. The contacts of other A compartment regions outside the immediate 16p11.2 locus (the locus comprises 16p11.2 +/- 500 kb) are distributed along the chromosome, without the 16p11.2-characteristic depletion between contact patches (example shown in Figure 5.10b, viewpoint 16p - 1.5 Mb).

The 16p11.2 locus, together with all its contact patches on chromosome 7, is located in the A compartments, however, not every A compartment is connected to 16p11.2, which might indicate that 16p11.2 is involved in a network of active chromatin regions, that has some specificity (Figure 5.10b), especially, as the A compartments directly up- and downstream of 16p11.2 are not a part of this network. However, most of the contacts of 16p11.2 within its own chromosome do not contain differentially expressed genes (Figure 5.10b). In the 16p-d cell line, 1 (Prr19) out of the 6 differentially expressed genes outside the deletion on chromosome 7 can be found inside a contact patch of 16p11.2 (chr7:24,881,000-26,661,000). In the dosage compensated control, 2 (*2310022K01Rik, 2610034B18Rik*) out of 9 differentially expressed genes are located inside a contact patch (chr7:27,115,000-31,993,000; chr7:82,122,000-92,028,000). Further, the contact patches of the 16p11.2 locus are preserved between the three cell lines, as shown here for the contacts of the region upstream of 16p11.2 (Figure 5.10b, viewpoint 16p – 500 kb). At this time, we did not pursue further how the intrachromosomal contacts of the 16p11.2 with other regions of chromosome 7 may relate with changes of gene expression within compartment A, or how they may affect the organisation of the regions that contain differential expressed genes located in compartments B. In the next sections, we instead decided to focus on exploring changes found common to 16p-d and 16p-dd, namely those affecting the protocadherin gene locus in chromosome 18.

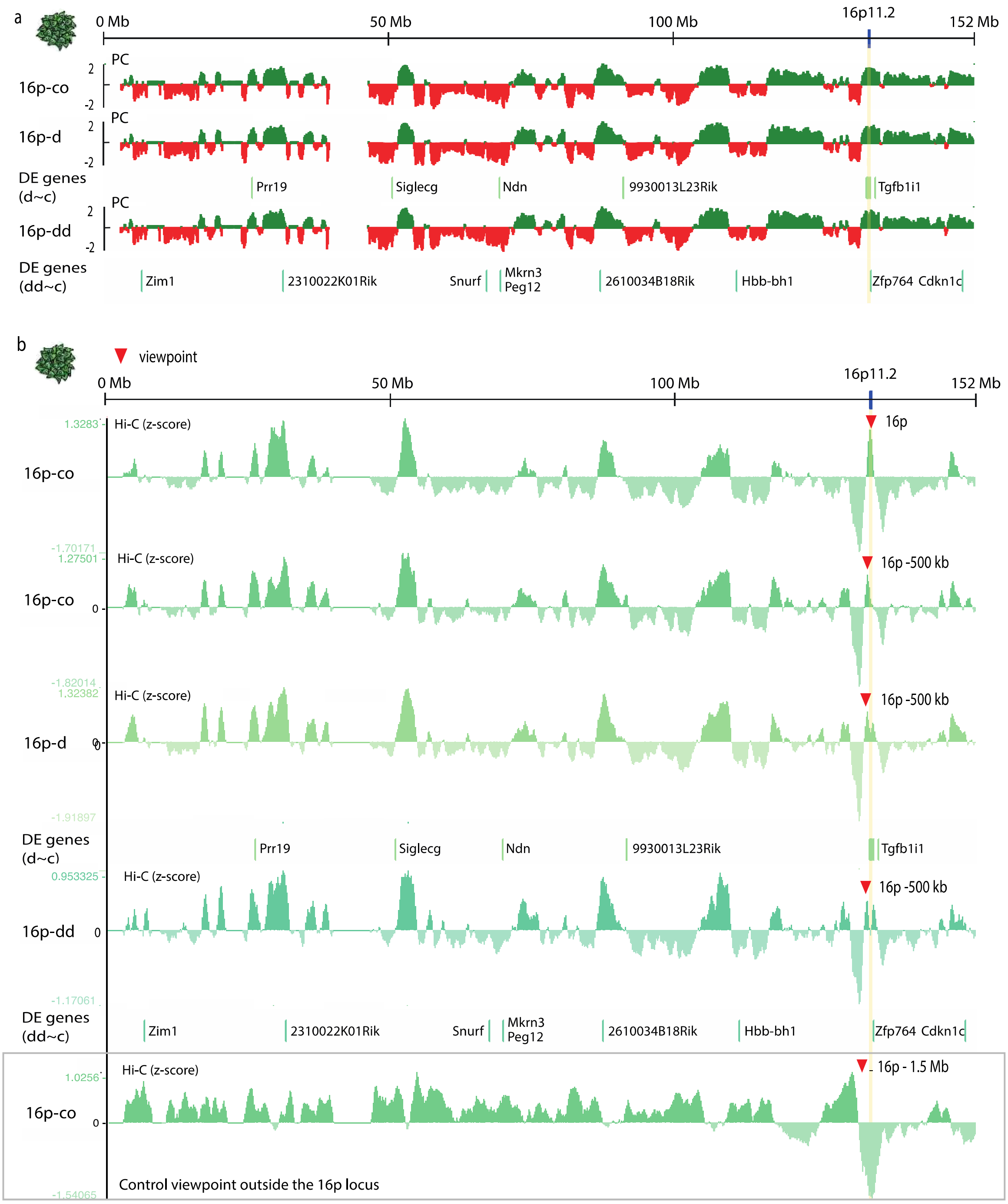


Figure 5. 10: The 16p11.2 locus engages in long-range contacts with distinct parts of the A compartment of chromosome 7

(a) Principal component analysis (PCA) of chromosome 7 in NPCs for 16p-co, 16p-d, and 16p-dd, A compartment is shown in green, B compartment in red. Below, significantly differentially expressed genes in 16p-d and 16p-dd. The position of the 16p11.2 deletion is highlighted in yellow. (b) Z-score normalised Hi-C contacts, showing contacts of one 500 kb viewpoint with chromosome 7. From top to bottom: z-score (log) 16p-co with the viewpoint in 16p11.2, z-score (log) 16p-co with the viewpoint 500 kb upstream of 16p11.2, z-score (log) 16p-d with the viewpoint 500 kb upstream of 16p11.2, significantly differentially expressed genes of 16p-d, z-score (log) 16p-dd with the viewpoint 500 kb upstream of 16p11.2, significantly differentially expressed genes of 16p-dd, z-score (log) 16p-co with the viewpoint 1.5 Mb upstream of 16p11.2.

5.11 *Pcdh* gene clusters on chromosome 18 are deregulated in 16p-d and 16p-dd

When exploring the chromatin topology of chromosome 7, we were able to identify specific long-range chromosomal contacts involving the 16p11.2 locus, but these contacts could not readily explain gene expression differences observed between the control and 16p-d or 16p-dd on chromosome 7, potentially because none of the significantly differentially expressed genes in 16p-d on chromosome 7 are also found in the dosage-compensated control 16p-dd. Thus, indirect effects of the reduced expression of the genes within the 16p11.2 deletion might contribute to these changes. Thus, we decided to explore also gene expression changes on chromosome 18 with respect to chromatin topology, and focus on the *Pcdh* genes that are differentially expressed in both 16p-d and 16p-dd (Figure 5.11). While the *Pcdhb* cluster is upregulated both in 16p-d and 16p-dd, the neighbouring *Pcdha* and *Pcdhg* clusters show different behaviours. The *Pcdha* cluster is upregulated in 16p-dd and unaffected in 16p-d, the *Pcdhg* cluster is downregulated in 16p-d and unaffected in 16p-dd (Figure 5.11). The *Pcdh* clusters *a*, *b*, and *g* comprise the majority of all protocadherin (*Pcdh*) genes, in a total of nearly 60 out of 80 genes, which are expressed in neurons and have been demonstrated to play a central role in self-recognition of neuronal cells in the central nervous system (Chen and Maniatis, 2013). The importance of *Pcdh* genes in neuronal development and their co-regulation in the 16p11.2 deletion and the dosage compensated control cells not only suggest a role of *Pcdh* genes in the manifestation of the 16p11.2 phenotypes, but it further suggests a causal relationship to the deletion, that is not connected to gene dosage effects.

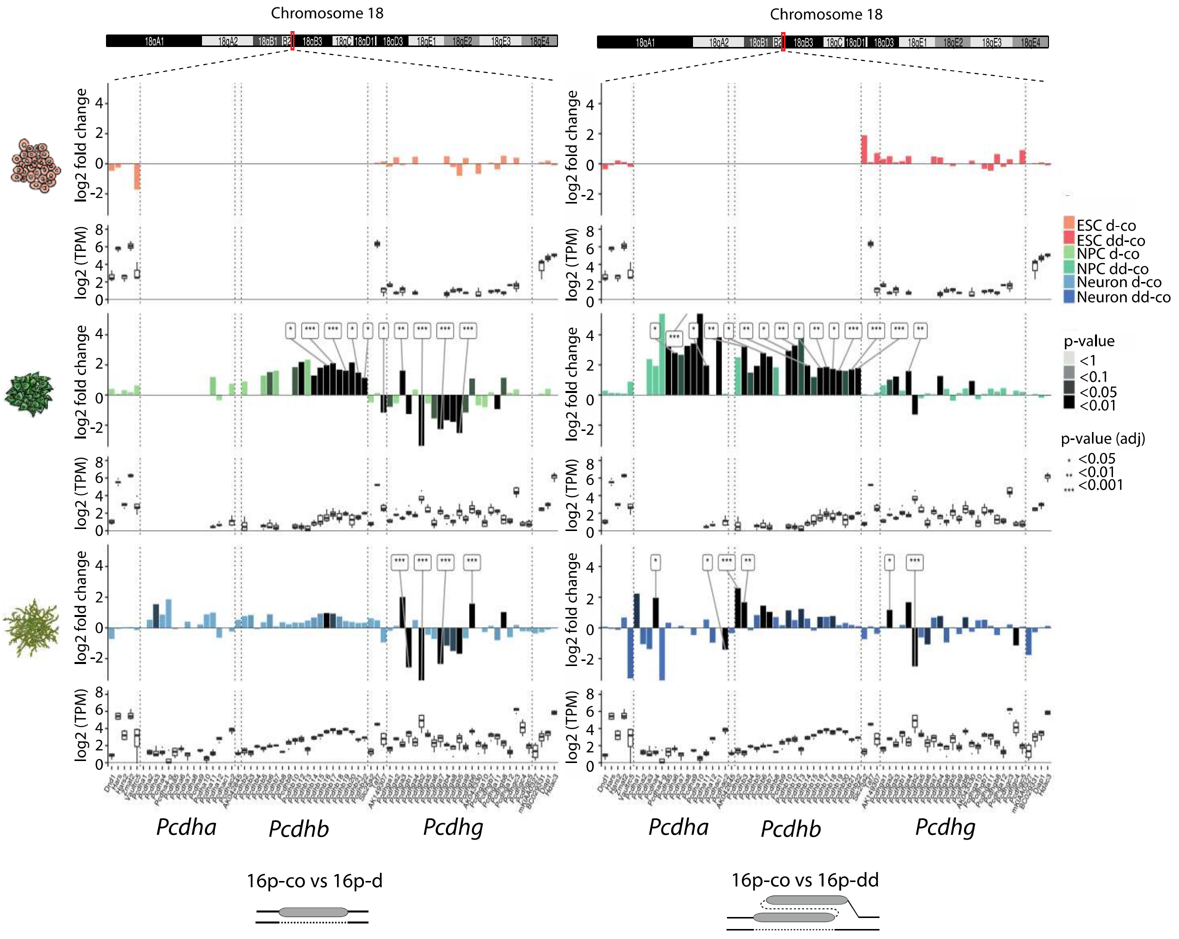


Figure 5. 11: *Pcdh* clusters are differentially expressed in 16p-d and 16p-dd.
Fold change (log2) of genes in the *Pcdh* clusters *a*, *b*, and *g* in 16p-d (left) and 16p-dd (right) compared to the control cell line 16p-co in ESCs (top), NPCs (middle), and day 16 neurons (bottom). Adjusted and unadjusted p-values are indicated by stars and by colour gradient, respectively. Under each graph showing the fold change, the gene expression (log2 TPM) of 16p-co is shown.

5.12 Chromatin topology of the *Pcdh* locus changes in 16p-d and 16p-dd

The *Pcdh* gene clusters *a*, *b*, and *g* are located next to each other, inside a 1 Mb genomic region of the B compartment of chromosome 18 (Figure 5.12a). As expected, no expression is detected in *Pcdha* and *Pcdhb* in ESCs, while *Pcdhg* is weakly expressed. In NPCs, the three clusters remain located in the B compartment, although the *Pcdhb* and *g* clusters become expressed or increase expression, respectively, while the *Pcdha* cluster remains silent (Figure 5.11). The association with the B compartment remains in NPCs for both 16p-d or 16p-dd, although changes in gene expression are detected in all the clusters (Figure 5.12a). Further, all three *Pcdh* clusters are located inside one TAD (boundary positions indicated in Figure 5.12b), with clearly demarcated borders in 16p-co, 16p-d and 16p-dd. Noteworthy, there are sub-TAD structures within the *Pcdh*-TAD (Figure 5.12b), which may contain information about the differential regulation of the *Pcdh* clusters.

The *Pcdhb* cluster is located in the middle of the *Pcdh* TAD and upregulated in 16p-d and 16p-dd NPCs. To explore the chromatin contacts of the *Pcdhb* cluster, we quantified the z-score normalised contact frequencies of the *Pcdhb* cluster with the neighbouring region (Figure 5.12c). Interestingly, in ESCs, where *Pcdhb* genes are not expressed, the cluster has high contact frequencies within itself and with the *Pcdha* and *g* clusters in 16p-co, 16p-d and 16p-dd, with a clear drop of contacts at both ends of the cluster.

In NPCs, the contacts of *Pcdhb* with itself are lower than in ESCs in the control line 16p-co. Upstream of *Pcdhb*, both the drop of contact frequencies at the border of the *Pcdhb* cluster and the contacts with the neighbouring *Pcdha* cluster remain. Downstream of the *Pcdhb* cluster we find a more marked decreased in contacts, which taken together with the increased gene expression of the *Pcdhb* and *Pcdhg* clusters in NPCs, suggests that this structure acquires a slightly more open chromatin conformation in NPCs compared to ESCs. In the 16p-d and 16p-dd NPCs the contacts of *Pcdhb* with the neighbouring regions change drastically. The self-contacts of the *Pcdhb* cluster are highly reduced, and while the contacts with the *Pcdha* clusters are mostly unchanged, the *Pcdhg* cluster interacts more frequently with *Pcdhb*.

The expression of *Pcdhb* genes in neurons is known to be regulated by CTCF-mediated looping between the promoter of a *Pcdhb* gene and a distant enhancer containing six DNaseI hypersensitive sites, the cluster control region (CCR) of *Pcdhb* (Yokota et al., 2011). The CCR is located downstream of the *Pcdhg* cluster, and its deletion leads to downregulation of *Pcdhb* genes, and also some isoforms of the *Pcdhg* genes. Interestingly, the contacts between the *Pcdhb* cluster and the CCR increase in the 16p-d and 16p-dd NPCs in comparison to the control (Figure 5.12c), in concordance with the upregulation of the *Pcdhb* cluster in the two cell lines.

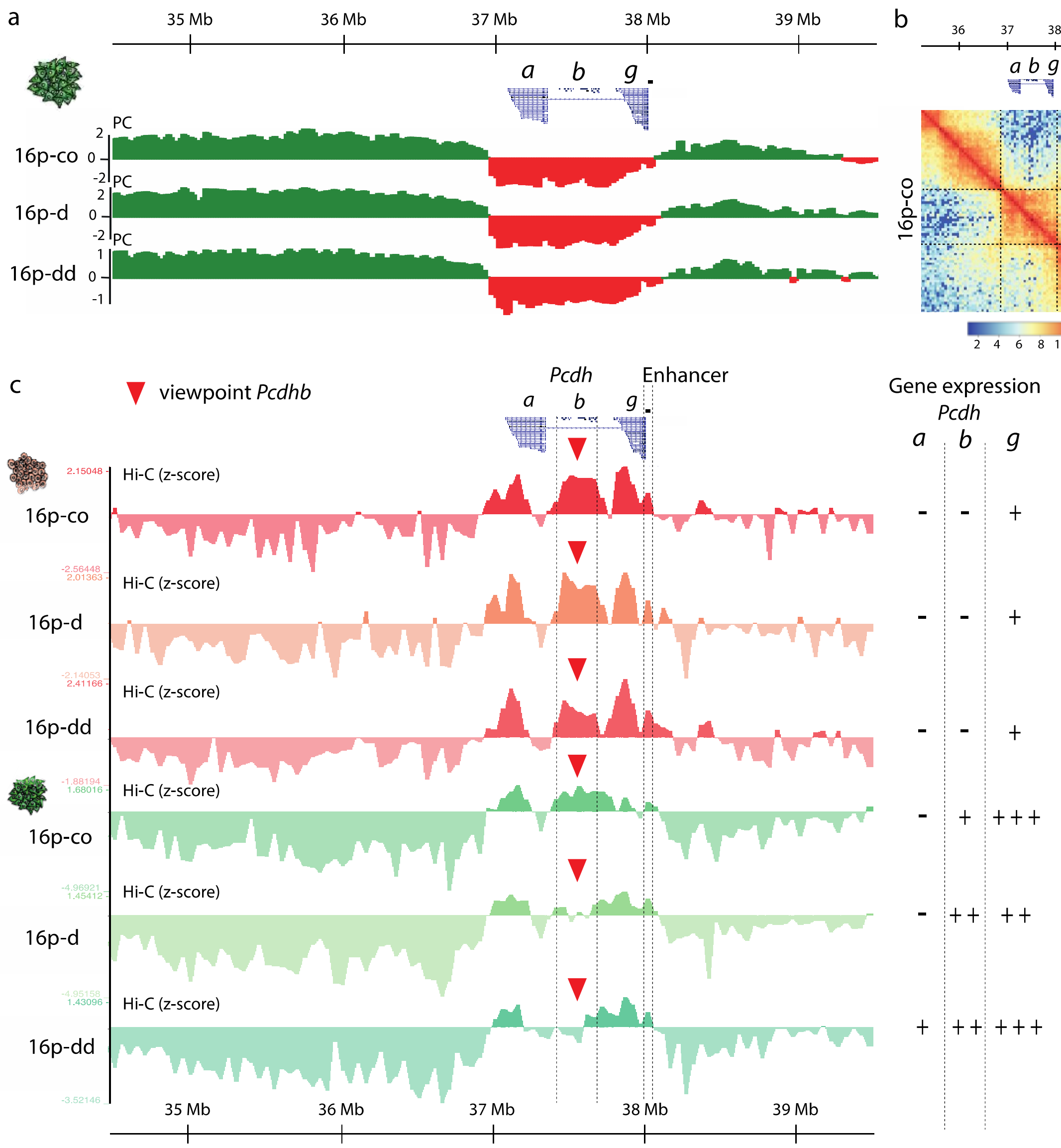


Figure 5. 12: Chromatin contacts of the *Pcdhb* cluster change in the 16p-d and 16p-dd NPCs but not in ESCs.

(a) Principal component analysis (PCA) of chromosome 18 in NPCs; Eigenvalues of the PCA shown for the *Pcdh* locus in 16p-co, 16p-d, and 16p-dd, A compartment in green, B compartment in red. (b) Hi-C contact frequencies (log2) of 16p-co NPCs at the *Pcdh* locus. TAD boundaries are indicated by dotted lines. (c) Z-score normalised Hi-C contacts, showing contacts of the *Pcdhb* cluster with its local surrounding on chromosome 18. From top to bottom: z-scores (log) of 16p-co ESCs, 16p-d ESCs, 16p-dd ESCs, 16p-co NPCs, 16p-d NPCs, 16p-dd NPCs. Table at the right indicates gene expression of the *Pcdh* clusters in ESCs and NPCs in each cell line, in the same order as the z-scores.

5.13 Inter-chromosomal contacts connect the 16p11.2 locus with chromosome 18

To study how the 16p11.2 deletion on chromosome 7 might be connected to the gene expression and chromatin changes at the *Pcdh* clusters on chromosome 18, we examined chromatin contacts between chromosomes. Inter-chromosomal contacts detected with Hi-C are magnitudes lower than contacts within the same chromosome, however, their specificity has been shown in several studies (Apostolou and Thanos, 2008; Horta et al., 2018; Spilianakis et al., 2005). Inter-chromosomal contacts between the 16p11.2 locus and other chromosomes have also been reported in human blood cells (Loviglio et al., 2016). To this end, we explored the z-score normalised inter-chromosomal contacts of the 16p11.2 locus on chromosome 7 with chromosome 18 (Figure 5.13a, viewpoint 16p). As inter-chromosomal contacts cannot be assigned to a certain genomic distance, we used the average contact frequencies between chromosome 7 and 18 as a background signal for the z-score normalisation. Interestingly, in NPCs the 16p11.2 locus interacts with several regions on chromosome 18, most frequently with the region immediately upstream of the *Pcdh* TAD. The *Pcdh* clusters are clearly depleted for contacts with the genomic region upstream of the *Pcdh* TAD, more than with any other region along chromosome 18 (Figure 5.13a, shown for viewpoint *Pcdhb*). Noteworthy, this inter-chromosomal contact of 16p11.2 is present in NPCs, but not in ESCs (Figure 5.13a), which also do not show detectable changes in *Pcdh* gene expression. To further validate the contacts between the 16p11.2 locus and the *Pcdh*-adjacent TAD, I examined *trans* contacts in published Hi-C data in NPCs (Bonev et al., 2017). The Hi-C data was generated from NPCs after 60 hours differentiation into cortical neurons. Despite the biological differences between NPCs from different neuronal differentiation systems, the inter-chromosomal contact can be clearly seen in the contact map of chromosome 7 and 18 (Figure 5.13b).

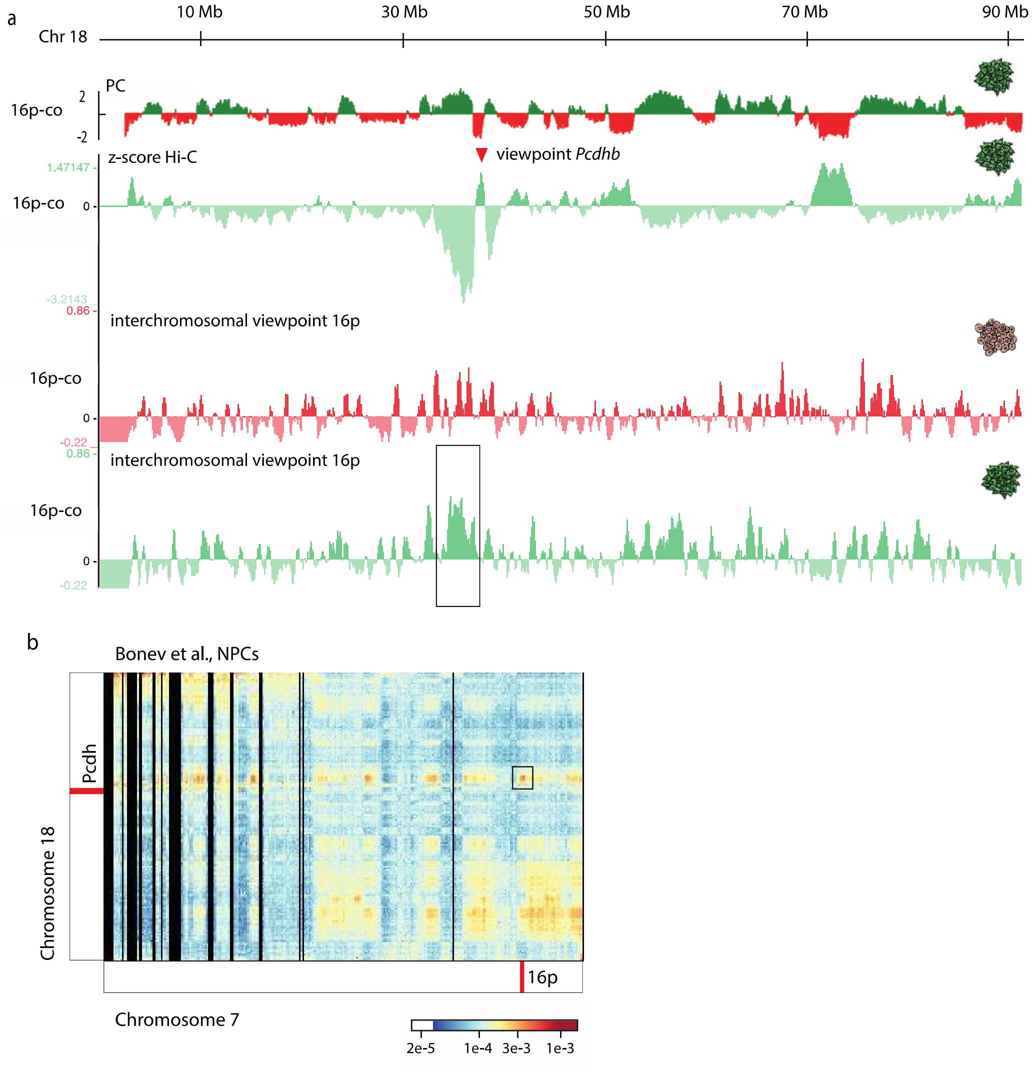


Figure 5. 13: Inter-chromosomal contacts between the 16p11.2 locus and chromosome 18
(a) from top to bottom: Principal component analysis (PCA) of chromosome 18 in 16p-co NPCs, z-score normalised Hi-C contacts (log) of the *Pcdhb* cluster with chromosome 18 in 16p-co NPCs, z-score normalised Hi-C contacts (log) of 16p with chromosome 18 in 16p-co ESCs, z-score normalised Hi-C contacts (log) of 16p with chromosome 18 in 16p-co NPCs. (b) Inter-chromosomal Hi-C contact frequencies between chromosome 7 and 18 in NPCs of cortical neurons from Bonev et al. (2017). Boxes in (a) and (b) indicate the position of the *trans* contact between 16p11.2 and the *Pcdh*-adjacent region.

5.14 Inter-chromosomal contacts between chromosome 7 and chromosome 18 change upon deletion of 16p11.2, specifically between two potential super-enhancers

There has been growing evidence that genomic regions containing super-enhancers can contact each other over large genomic distances (Beagrie et al., 2017; Quinodoz et al., 2018). Further, super-enhancers often associate with splicing speckles (Chen et al., 2018b), where inter-chromosomal contacts occur frequently (Quinodoz et al., 2018). To explore whether super-enhancers could be involved in the *trans* contacts between chromosome 7 and 18, we examined the two chromosomes for known super-enhancers. Since super-enhancer locations are not available for our cell type of interest, we chose the closest match to NPCs where super-enhancers have been identified, in mouse embryonic day E14.5 brain (Khan and Zhang, 2016). In fact, not the region comprising the 16p11.2 deletion itself but the genomic region immediately upstream of 16p11.2, as well as the *Pcdh*-adjacent region contain super-enhancers in the developing brain (positions indicated in Figures 5.8 and 5.14). From explorations of the intrachromosomal contacts of chromosome 7, we had seen that the regions upstream and downstream of 16p11.2 are engaged in the same long-range contacts as 16p11.2 (Figure 5.9). For inter-chromosomal contacts, the same trend can be observed. Both neighbouring genomic regions of 16p11.2 (+/- 500 kb) interact with the *Pcdh*-adjacent region on chromosome 18 in NPCs (Figure 5.14a). The presumable presence of super-enhancers at both sites of the inter-chromosomal contact provides a potential mechanism for contacts between chromosomes that could involve splicing speckles.

When comparing the z-score normalised contact frequencies of the 16p11.2-adjacent regions in the control and the 16p-d and 16p-dd NPCs, clear changes can be observed. While the contact patch anchoring the 16p11.2 locus to chromosome 18 is still present, the signal inside the patch is redistributed in 16p-d and 16p-dd (Figure 5.14 a). To explore this change in higher resolution with respect to the positions of the two super-enhancers on chromosome 7 and 18, I plotted the contacts of the 16p-adjacent region containing the chromosome 7 super-enhancer (16p – 500 kb) over the genomic region containing the chromosome 18 super-enhancer inside the inter-chromosomal contact patch (Figure 5.14 b). In NPCs, the contact between the two potential super-enhancer regions in 16p-co is lost upon deletion of 16p11.2 in both 16p-d and 16p-dd. In ESCs, this contact is not present in any of the three cell lines.

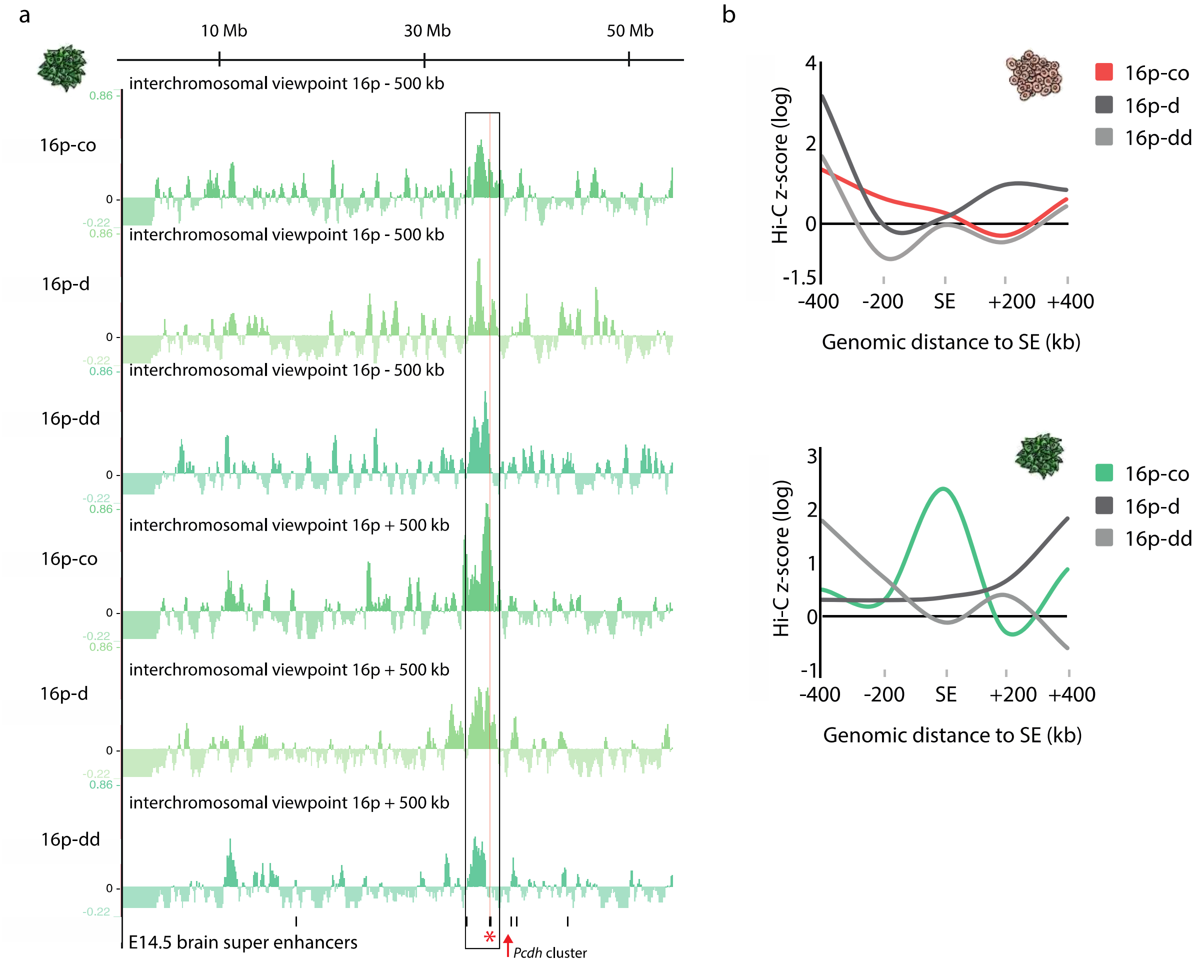


Figure 5. 14: The contacts of the 16p11.2 locus with a potential super-enhancer on chromosome 18 change when introducing the 16p11.2 deletion.

(a) Z-score normalised Hi-C contacts (log) with a ~60 Mb region on chromosome 18 in 16p-co, 16p-d, and 16p-dd NPCs. Top three viewpoints are located 500 kb upstream of 16p11.2, bottom three viewpoints are located 500 kb downstream of 16p11.2. Below, the positions of super-enhancers in embryonic day E14.5 mouse brain (black) are shown. Red asterisk indicates the chromosome 18 super-enhancer contacting the 16p11.2 locus. (b) Contacts of the 16p11.2-adjacent region containing the chromosome 7 super-enhancer (16p – 500 kb) with the genomic region containing the chromosome 18 super-enhancer marked in (a). Contacts with the 200 kb region containing the potential super-enhancer are shown plus 400 kb up- and downstream of the enhancer region. At the top the contacts of 16p-co, 16p-d, and 16p-dd ESCs are shown, below 16p-co, 16p-d, and 16p-dd NPCs are shown.

5.15 Moderate changes in contact frequencies between the *Pcdh* clusters and the adjacent potential super-enhancer on chromosome 18

The question remains how the loss of an inter-chromosomal contact upstream of the *Pcdh* clusters affects the gene expression of the *Pcdhb* genes. The *Pcdhb* locus does not interact with its upstream TAD or the potential super-enhancer within this region, in fact, all three *Pcdh* clusters are strongly depleted of contacts with the potential super-enhancer on chromosome 18 (Figure 5.15a). However, few contacts between the super-enhancer and the *Pcdhg* cluster emerge in 16p-d and 16p-dd, noticeable in a reduced depletion of contacts (Figure 5.15a, red arrow). Noteworthy, when analysing the differential Hi-C contacts between 16p-co and both 16p-d and 16p-dd of chromosome 18, the change in contact frequencies between the *Pcdh* clusters and the potential super-enhancer upstream of the clusters is one of the most prominent significant contact changes on chromosome 18. Thus, the change in contact depletion in 16p-d and 16p-dd NPCs might indicate the emergence of ectopic contacts between the *Pcdh* clusters and the potential super-enhancer in a subset of the cell population, which might be sufficient to alter gene expression of the *Pcdhb* cluster.

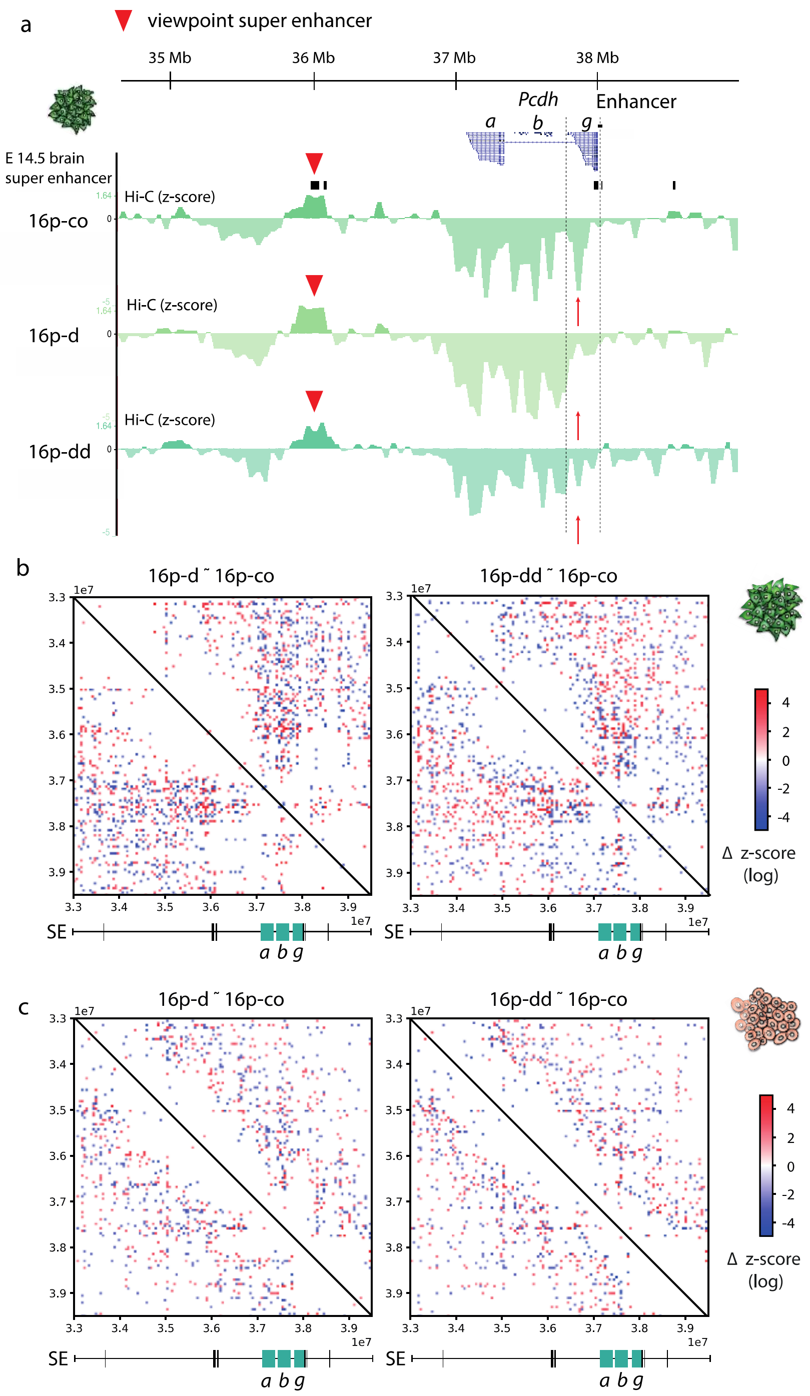


Figure 5. 15: Moderate changes in chromatin contact frequencies between the *Pcdh* clusters and the upstream region interacting with 16p11.2

(a) Z-score normalised Hi-C contacts (log) of the potential super-enhancer upstream of the *Pcdh* TAD with the *Pcdh* locus in 16p-co, 16p-d, and 16p-dd NPCs. Super-enhancer positions (black) are indicated above. Arrows indicate region with changed contact frequencies. (b) Delta z-score of Hi-C contacts in NPCs for 16p-d minus 16p-co (left), and 16p-dd minus 16p-co (right). (c) Delta z-score of Hi-C contacts in ESCs for 16p-d minus 16p-co (left), and 16p-dd minus 16p-co (right).

5.16 Summary and conclusions

In this chapter, we investigated chromosome folding in the context of congenital disease associated with a large genomic deletion at the human locus 16p11.2. The heterozygous deletion contains 27 protein-coding genes and is linked to a high prevalence of autism spectrum disorder (ASD) in humans. When introducing this deletion in mice, it phenocopies the human disease. We differentiated murine embryonic stem cells (ESCs) carrying the 16p11.2 deletion (16p-d) into neuronal precursor cells (NPCs) and further into premature dopaminergic neurons, alongside with an ESC line carrying the 16p11.2 deletion on one allele, and the respective duplication on the other allele for the purpose of dosage-compensation of the genes inside the deletion. This dosage-compensated control (16p-dd) provided the possibility to focus our investigation on the direct effects of the genomic deletion on genome topology, disentangled from indirect effects of the reduced gene dosage of the 27 protein-coding genes inside the 16p11.2 deletion. We could confirm the dosage compensation of genes located at the 16p11.2 region in the 16p-dd line in ESCs and NPCs, but not in neurons, where gene expression was altered possibly an effect of the two adjacent copies in one allele. This suggests potential regulatory effects of introducing the duplication at the 16p11.2 locus, a topic which we chose not to address in this study, but may have relevance to understand 16p11.2 duplication syndromes.

In ESCs, gene expression differences between the wildtype (16p-co) and 16p-d or 16p-dd were marginal, suggesting that the effects of the 16p11.2 deletion do not affect pluripotency but only manifest later in development. Thus, we focused on NPCs, where most common gene expression differences were found between 16p-d and 16p-dd with respect to the wildtype. In NPCs, the 3D chromatin folding of the 16p11.2 locus, a genomic region on the murine chromosome 7, revealed that the 16p11.2 deletion comprised a TAD inside the A compartment of chromosome 7. The chromatin contacts of the 16p11.2 locus change only marginally between wildtype and 16p-d or 16p-dd, with no effect on the association of 16p11.2 with the A compartment or the surrounding TAD boundaries, nor with a direct link to the gene expression differences observed in proximity to the 16p11.2 deletion.

However, we noticed that the 16p11.2 locus engaged in long-range contacts with other, but not all, A compartments along the entire length of chromosome 7, while being depleted of contacts with the remaining chromosome. These long-range contact patches were found exclusively for the genomic region of the 16p11.2 deletion and the immediate upstream and downstream regions (+/- 500 kb). Regions of the A compartment further away from the

16p11.2 locus did not engage in specific contact patches, but showed broadly distributed contacts along the chromosome. This unique contact pattern of the 16p11.2 locus suggests an affiliation of 16p11.2 in long-range chromatin contact hubs, which have been described for example for active genes or super-enhancers that engage in multiway contacts over large genomic distances (Beagrie et al., 2017; Quinodoz et al., 2018). The genes on chromosome 7 with significant expression changes in 16p-d or 16p-dd did not necessarily interact with the 16p11.2 locus, only a small subset of those genes were found inside a contact patch of 16p11.2, and further, only half of those genes were associated with the A compartment, suggesting that those gene expression differences are caused by indirect effects of the deletion, or alternatively result from rewiring of the A compartments that may affect the organisation of the adjacent B compartments bearing differentially expressed genes. Noteworthy, none of the genes on chromosome 7 changed their expression in both 16p-d and 16p-dd.

To our surprise, the most striking difference in gene expression common to 16p-d and 16p-dd was the upregulation of a protocadherin gene cluster on chromosome 18 in NPCs. The upregulation of the *Pcdhb* cluster was accompanied by downregulation of the *Pcdhg* cluster in 16p-d and upregulation of the *Pcdha* cluster in 16p-dd. All three cluster are within a ~1 Mb sized TAD inside the B compartment of chromosome 18, with possible sub-TAD structures. While the boundaries of the *Pcdh* TAD are maintained between wildtype, 16p-d, and 16p-dd, contacts within this TAD are rearranged in the mutants, consistent with a more open chromatin conformation of the *Pcdhb* cluster in 16p-d and 16p-dd, as well as ectopic contacts between the *Pcdhb* and the *Pcdhg* cluster. Further, the *Pcdhb* cluster revealed a slight increase in contact frequencies with a described enhancer, the cluster control region (CCR) of *Pcdhb*.

An exploration of the inter-chromosomal contact frequencies between chromosome 7 and 18 showed that indeed the 16p11.2 locus interacts with a region on chromosome 18, immediately upstream of the *Pcdh* clusters. Both, the 16p11.2 locus, more precisely the upstream neighbour of the 16p11.2 deletion, and the *Pcdh*-adjacent region involved in the inter-chromosomal contact contain super-enhancers in the developing mouse brain, indicating towards a possible mechanism for contacts between chromosomes involving super-enhancers. Combined evidence from SPRITE and TSA-seq suggests that super-enhancers from different chromosomes can associate with the same splicing speckle, thereby generating inter-chromosomal contacts (Chen et al., 2018b; Quinodoz et al., 2018). Thus, we propose that the contact of 16p11.2 with chromosome 18 detected by applying a z-score normalisation on our Hi-C data is a candidate contact between chromosomes, potentially mediated by super-enhancers. The presence of this contact was confirmed in Hi-C data from NPCs of cortical neurons, which was sequenced to higher depth than our Hi-C data (Bonev et al., 2017). The inter-chromosomal contact between the two E14.5 brain super-enhancer is present in NPCs, but not in ESCs, and is lost in 16p-d and 16p-dd. Finally, we showed that the *Pcdh* clusters are depleted of contacts with the upstream super-enhancer. Upon introduction of the 16p11.2 deletion, this depletion is diminished, potentially due to sporadic contacts between the super-enhancer and *Pcdhg* in a subset of the cell population. We propose that the moderate, but significant difference in contact frequencies between the brain super-enhancer and the *Pcdh* cluster might trigger the upregulation of the *Pcdhb* cluster in the absence of the 16p11.2 locus.

6. Discussion

6.1 Detecting chromatin contacts with GAM

Chromatin contacts and 3D genome topology play important roles in nuclear functions, such as gene regulation and replication. Different techniques can be used to map 3D genome folding and detect pairwise or multiway interactions between genomic regions or with nuclear compartments, such as nuclear bodies or the lamina. Pairwise chromatin contacts within the range of DNA loops and TADs can now be mapped with unprecedented resolution, with advances in 3C-based assays, such as high-depth Hi-C, that push the achievable genomic resolution to obtain contacts at 1 kb resolution or higher (Hsieh et al., 2019; Rao et al., 2014). With the development of ligation-free techniques, such as genome architecture mapping (GAM) (Beagrie et al., 2017), or split-pool recognition of interactions by tag extension (SPRITE) (Quinodoz et al., 2018), complex multi-way contacts have been revealed that extend over large genomic distances and connect super-enhancers with each other, enhancer with active genes, or that bring together active genes. In the first GAM dataset, produced in mouse embryonic stem cells (mESCs), pairwise and multiway contacts were studied by collecting and analysing 400 nuclear profiles (Beagrie et al., 2017). This data allowed the exploration of pairwise contacts at 30 kb resolution, and higher-order contacts between topological domains. Contacts between chromosomes were observed, but not analysed in this first small dataset, due to insufficient reproducibility of inter-chromosomal contacts when considering halves of the dataset, therefore requiring the expansion from GAM datasets with 400 NPs to larger datasets with more than 1000 NPs.

6.1.1 Optimisations of the GAM protocol

With the aim to explore chromatin contacts at higher resolution and with increased statistical power, I collected a larger GAM dataset comprising several thousand nuclear profiles. To generate a dataset of this magnitude, I first optimised GAM to reduce collection time and costs. In this thesis, I presented a greatly improved version of GAM that is five times faster (40 to 9 days to collect a 400-NP dataset) and cheaper (~3x) than the published method (Figure 3.2). Most importantly, this protocol includes a new method for whole genome amplification (WGA) that is independent from commercially available kits. This protocol will be of value in other genomic applications that require amplification of minute amounts of DNA. The work presented here allow the wider use of GAM, and the collection of larger datasets to investigate not only pairwise, but higher-order contacts, radial positioning, compaction and inter-chromosomal contacts, which are all features that GAM can detect, but in the first dataset with 400 NPs had not reached their full potential. Further, I showed that GAM can be combined with immunofluorescence, a major development that allows the targeted selection of specific cell types in a mixed cell population or tissue bringing the opportunity to map chromatin contacts in vivo, directly in tissue without disturbing the tissue context of the cells of interest (Figure 3.8). With the optimisations I present in this thesis, large GAM datasets can be collected to study genome folding in great detail in any cell type of interest (e.g. pluripotent cells in the epiblast, G. Loof, unpublished work; glutamatergic neurons in the hippocampus, I. Harabula, unpublished work).

While the GAM protocol presented in the beginning of chapter 3 drastically reduces time and costs of the experiment, the change from a commercially available kit for WGA to the in-house WGA protocol was accompanied by a slight decrease in data quality of nuclear profiles during most of the collection of the data presented in this thesis. Careful assessment of the quality of nuclear profiles collected with the new protocol helped us identify their most predictive quality criterium, which is the fraction of positive genomic windows with neighbouring positive windows (windows without neighbours were termed 'orphan windows'). The new protocol generated nuclear profiles with slightly higher percentages of orphan windows compared to the published protocol performed with the commercial WGA kit (WGA4, Sigma) sold before 2017. The discovery of this parameter helped us identify low quality samples in our dataset which we could remove based on the percentage of orphan windows, a quality cut-off that was found to be redundant with percentage of mapped reads in Beagrie et al. (2017), but improved the overall quality of the dataset compared to the cut-off applied previously (>15 % mapped reads).

The slight differences in the quality of nuclear profiles also affected the efficiency of detection of genomic regions in the dataset (i.e. the proportion of false negatives). Genome-wide comparisons of data produced with the new version of GAM or the published protocol revealed a decrease in DNA detection efficiency of 28 % in the new dataset (Figure 3.21). Although the efficiency of DNA detection is a good metric to help optimise protocol parameters and assess the quality of a dataset, it can only be computed after a relatively large number of samples is collected (e.g. at least 400 as in Beagrie et al. 2017). Further exploration of GAM performance using in silico polymers (Nicodemi laboratory) showed that efficiencies of window detection as low as 40 % are well tolerated in GAM, and result in similar data.

Furthermore, lower detection efficiencies within 40-100 % only marginally influence the number of nuclear profiles that are required to extract significant and reproducible contacts from GAM data (Figure 3.21). Fortunately, the GAM data collected with the new protocol still reached an efficiency of 70 % at 50 kb resolution, which allowed further exploration of the new GAM data without further compromises due to the efficiency.

Data comparisons of the large dataset produced here also showed bias in the sampling of eu- and heterochromatin, in which DNA from heterochromatic regions, such as lamina-associated domains, is detected in fewer nuclear profiles than euchromatic regions. Some of this bias has biological meaning due to the natural compaction of heterochromatin and can be found as well in the previously published GAM data. The slight difference in detection frequency in the published dataset was in fact used to identify the level of compaction of genomic regions (Beagrie et al., 2017). In the new dataset, this difference was more pronounced and a result of lower efficiency of extraction of DNA from heterochromatin.

Explorations of all tests performed in the past to optimise the GAM protocol helped us identify a small set of final optimisations that on their own had no or only minor effects on the performance of the technique, but when combined in one single experiment corrected for the observed bias, and extracted DNA with high efficiency from both eu- and heterochromatin (Figure 3.27). Therefore, a fully revised GAM protocol is now available which delivers high quality GAM data with a fully in-house approach, which is affordable, tuneable and much faster.

To be able to work with the datasets presented here that have uneven detection frequencies in eu- and heterochromatin, we applied data normalisation using a pointwise mutual information approach, and showed that the normalisation effectively removed biases from the data. This is a promising result as it is likely that the implementation of GAM, in other labs and other samples, may initially give data with lower efficiency, but that the downstream analyses can be still performed if care is taken to correctly normalise the data and track down the biases. Our work also led to a batch of QC metrics that will be implemented in GAMtools (Beagrie and Schueler, 2017), for the broader use of GAM data in other laboratories. After normalisation, chromatin contacts in the new dataset, collected in the mESC line F123, largely resembled the chromosome topologies mapped in Beagrie et al. (2017; 46C mESCs). Interestingly, for many genomic regions with noticeable differences between both GAM datasets, the F123 GAM data showed high similarity with Hi-C data from F123 (Figure 3.28), suggesting biological rather than technical differences as the origin of the observed differences in chromatin contact frequencies between the two GAM datasets. Different cell culture conditions or genetic backgrounds from different mouse strains have been shown to influence pluripotency of mESCs (Iijima et al., 2010; Kawase et al., 1994), which could also explain differences in chromosome topology detected in the two GAM datasets.

Nonetheless, for future applications of GAM, we strongly recommend the use of the further optimised protocol presented in Figure 3.27, which allows the application of the in-house WGA protocol and all other advantages that our optimisations provide in comparison to the published GAM protocol, without reduced efficiency in DNA extraction. Summarising, the most important improvements of GAM presented in this thesis are a 4-fold reduction of the experimental time, a reduction of the sequencing time due to elimination of the need to use customised sequencing runs with dark cycles, a 3-fold cost reduction per nuclear profile, the implementation of an easy cellular staining that improves visibility of nuclear profiles during laser microdissection, automation of all pipetting steps in the protocol except DNA purification, safe stopping points that help optimise collection times, and the ability to select cells during laser microdissection with immunofluorescence.

6.1.2 Reproducibility of GAM data generated with different collection modes

An open question was the reproducibility of chromatin contacts in GAM data from 1NP and 3NP datasets. Here, I collected a GAM dataset with 1123 1NPs, and a dataset with 847 x 3NPs using the multiplexed version of GAM with three nuclear profiles per tube. I showed that datasets are highly similar in terms of quality and chromatin topologies, such as A and B compartments (Figures 4.6 and 4.8), which confirmed our theoretical predictions and reassures that 3NP collection does not change the majority of detectable contacts in a GAM dataset. Further, visual comparisons of chromatin contact maps showed large similarities between the normalised co-segregation frequencies of 1NP and 3NP data. The 3NP data could in principle contain long-range contacts that are not only due to co-segregation of spatially close genomic regions, but that could be an artefact of multiplexing. Interestingly, I did not observe increased noise at large genomic distances in 3NP data. On the contrary, visual inspections of 3NP contact maps showed in many cases less contacts spanning large genomic distances than 1NP data, a difference which yet has to be explored, as it could indicate either reduced noise or loss of some long-range contacts in 3NPs. Here, investigation of the composition of contacts in both datasets with gene expression and chromatin features, such as histone marks for active and inactive chromatin will help understand which contacts are different between collection modes, and whether they contain expected pairs of genomic loci, such as active genes with other active genes, or unexpected combinations, such as active genes with inactive genes (Beagrie et al., 2017). Further, Beagrie et al. (2017) presented a mathematic model for determining contact probabilities from co-segregation frequencies. Statistical inference of co-segregation (SLICE) was applied to identify the most significant contacts in the data. Since then, SLICE has been developed further to extract contact probabilities from 3NP GAM data (Carlo Annunziatella, Antonio Scialdone, Mario Nicodemi; unpublished). Applying SLICE to 1NP and 3NP datasets will help identify which contacts are signal and which are noise, and ultimately help identify the same significant chromatin contacts in 1NP and 3NP data.

6.1.3 Genomic resolution of GAM data

One of the aims of this thesis was to improve the achievable resolution of GAM. In Beagrie et al. (2017), the collection of 400 single NPs allowed mapping of chromatin contacts at 30 kb resolution. Here, we showed that the combined dataset of 1123 1NPs and 847 3NPs detected all mappable window pairs (i.e. at all intrachromosomal distances) at least once at 20 kb resolution, a measure for estimating the saturation of a GAM dataset. While full saturation was not achieved at 10 kb resolution, visual exploration of chromatin contacts shows good detection of genomic windows at 10 kb, with only subtle differences in genome sampling compared to 20 kb resolution (Figure 4.8). These observations promise that detection of chromatin contact at 10 kb can be achieved with our current GAM datasets from the mESCs F123 line, namely with the SLICE model, by inferring contacts using deep learning, as it has been shown for improving resolution in Hi-C data (Zhang et al., 2018b), and potentially by continuing to add more nuclear profiles to the dataset, since the same biological samples are stored in liquid nitrogen and available for further data collection. In contrast, Hi-C datasets have achieved resolutions of 1 kb using very deep sequencing (Bonev et al., 2017; Rao et al., 2014), a goal that GAM has not accomplished yet. The main difference to GAM is that the resolution of proximity ligation assays depends almost entirely on the sequencing depth and, depending on the DNA fractionation step, even nucleosome resolution can be obtained with high depth sequencing (Hsieh et al., 2019). In GAM, the genomic resolution depends foremost on the number of nuclear profiles. Even with reduced DNA detection efficiencies GAM can be used at high resolution if enough nuclear profiles are collected. Down to a detection efficiency of 50 % the number of nuclear profiles required for obtaining chromatin contacts remains unaffected, as shown with modelling of GAM data (Figure 3.21). Thus, to obtain a resolution at the level of high depth sequencing Hi-C datasets, eventually GAM will require less sequencing, but the time commitment of the experimentalist might be the limiting factor of the highest resolution that GAM will meet with the currently available protocol. Further optimisations of GAM to reduce costs and workload, namely the automation of the DNA purification steps and especially of the laser microdissection, will help achieving high depth datasets and show whether there is a limit to the genomic resolution that can be achieved with GAM.

6.1.4 Further improvements of the GAM protocol

As discussed above, the power and achievable resolution of GAM is closely dependent on the number of nuclear profiles in the dataset. However, with an increasing number of nuclear profiles costs and workload increase drastically. Thus, we already had implemented a multiplexing step in the GAM protocol to reduce the number of samples that has to be processed. By collecting three nuclear profiles into one tube, we reduced the workload for collecting many samples, however, we lost the single cell information that is inherent to a 1NP dataset. Multiplexing was therefore limited to a relatively small number of NPs per tube that still allows to distinguish interacting from non-interacting loci in the resulting GAM data (Figure 3.3). An important modification of the GAM technique would be the use of barcodes as early in the protocol as possible to be able to combine many nuclear profiles before producing sequencing libraries. One possibility to add barcodes to DNA presents itself during the whole genome amplification of nuclear profiles. Here, DNA is amplified with random hexamer primers linked to a common adapter sequence, which is amplified further to generate enough copies of the input material for processing and library preparation. By using indexed adapters, we could multiplex nuclear profiles from up to 96 tubes into one reaction for the subsequent library preparation. The use of indexed primers may result in reduced efficiency of DNA amplification, which should be monitored carefully when introducing this step to the protocol. The implementation of a barcoded whole genome amplification reaction would reduce the costs and especially the workload per dataset.

6.1.5 Determining contacts probabilities with SLICE

Beagrie et al. (2017) showed the successful implementation of SLICE, a mathematical model to determine contact probabilities and distinguish interacting from non-interacting loci in

GAM. Pairwise and multiway interactions were analysed in the mESC genome with SLICE, and found preferential enriched for contacts involving active genes and enhancers at 30 kb resolution, and for active regions and super-enhancers at the scale of interactions between topological domains (Beagrie et al., 2017). The use of SLICE in the F123 dataset is therefore essential not only to reproduce and further explore these contacts at higher resolution, but also to extend the analysis with SLICE to inter-chromosomal chromatin contacts, which were not studied in Beagrie et al. due to the small size of the dataset. We initially applied SLICE to analyse the 1NP F123 dataset, using the same approach as in Beagrie et al (2017). However, we found that the bias of increased detection of euchromatic regions resulted in biased contact probabilities, leading to overestimated contacts between active regions, and produced contact probabilities that largely resembled the raw co-segregation frequencies in F123 (results not shown). On the contrary, in 46C, where this detection bias is not present, SLICE detected significant contacts independent of the chromatin state of the underlying genomic region. While I did not show this work in the thesis, the discovery of unbalanced detection of eu- and heterochromatic regions in our latest GAM datasets has now motivated further improvements to the SLICE model, to be able to consider DNA detection efficiencies as an input to the model. Currently, a new version of SLICE is under development, that calculates and normalises for the detection efficiency of every genomic locus (Francesco Musalli, Nicodemi lab, University of Naples, Italy). SLICE with data normalisation will not only be useful to analyse the F123 datasets and other GAM data that have DNA detection biases due to experimental conditions, it will also help to work with datasets from samples with non-optimal DNA preservation, for example post-mortem human biopsies, which are often subject to DNA degradation (Zsikla et al., 2004).

Further, the application of GAM in F123 showed its power to identify chromatin contacts with allele specificity. The consistency of allele-specific co-segregation frequencies depends on the uniform distribution of sequence variants in the genome. Thus, some genomic regions are under detected in allele-specific data because they have a lower SNP density and can be phased less effectively. Those regions would appear as never contacting in the current version of SLICE and consequently bias SLICE for regions with high sequence variability between the alleles. Here, normalisation of SLICE based on DNA detection efficiencies would allow the unbiased identification of allele-specific contact probabilities.

6.1.6 Future analysis of chromatin topology in F123

GAM has the ability to detect pairwise and multiway contacts (Beagrie et al., 2017). While pairwise contacts can be investigated using normalised co-segregation frequencies, higher-order contacts have so far only been studied using SLICE. Analysis of contact probabilities from both pairwise and multiway contact in the F123 genome with a version of SLICE that takes into account different detection efficiencies within a dataset will help compare the published GAM dataset with the here produced data in F123 mESCs. Concordant with the larger number of NPs in the F123 dataset, SLICE should be able to extract significant contacts at higher resolution than previously achieved. Especially for multiway contacts, this opportunity becomes very relevant. In Beagrie et al. (2017), 3-way contacts were analysed at the resolution of entire TADs (500 kb – 1 Mb), which does not allow for the discovery of higher-order contacts at the single gene level, for example between several enhancers with a single target promoter. A high-resolution analysis of multiway contact opens exciting possibilities to study gene regulation in 3D with GAM data.

Several studies have reported gene networks of genes with similar transcriptional activity, which suggests that genes come together for coordinated gene expression. Cluster of genes have been described for genes bound by the same transcription factors (Apostolou et al., 2013; Denholtz et al., 2013; Schoenfelder et al., 2010b; Wei et al., 2013). Furthermore, it has been suggested that different genes associate in different cell types (Bonev et al., 2017; Schoenfelder et al., 2015a) and that repressed genes also cluster in Polycomb bodies for coordinated gene silencing (Schoenfelder et al., 2015b). Deletions of genes from a gene cluster can affect the expression of other genes in the cluster (Engreitz et al., 2016; Fanucchi et al., 2013). However, attempts to image multiple genes inside a gene cluster showed low percentage of cells with multiway contacts, even for genes with coordinated response to the deletion of a gene from the cluster (Fanucchi et al., 2013). While most studies using 3C-based assays claim to detect chromatin hubs, none of them has actually shown the simultaneous association of multiple genes in a cluster. Polycomb-target genes for example have been shown to contact each other in pairwise contacts, but it remains unclear how often more than two Polycomb-target genes come together in the same nuclear body at the same time. Quinodoz et al. (2018) used SPRITE, a ligation-free assay to detect higher-order contacts, and showed that indeed multiple genes come together at splicing speckles or at the nucleolus. However, imaging of those contacts suggests relatively large spatial distances between the genes inside the same cluster (Quinodoz et al., 2018), larger than seen for the super-enhancer contacts detected by GAM and validated by cryo-FISH (Beagrie et al., 2017). Thus, SPRITE maps genes in spatial proximity but may not be directly sensitive to physical distance as in GAM, and therefore have some difficulty in distinguishing associations within a physically large chromatin hub from protein-mediated interactions with close spatial distances. GAM analyses contacts with respect to their spatial distance, independently of the network of proteins (or other molecules) that lead to the spatial proximity. SLICE applies a distance threshold for interacting loci, in the first use of SLICE the investigated spatial distance was limited to 100 nm. Thus, with the application of SLICE in F123 mESCs, we will be able to identify multiway interactions within close spatial distances between genes. We expect to gain insight into the function of multiway contacts in gene regulation, and reveal regulatory regions and genes that interact with each other simultaneously in the same nuclear hub. Further, we can use this information to learn more about the transcription factors involved in enhancer-gene and gene-gene contacts. While previous work has shown the involvement of architectural proteins, such as cohesion and mediator complex, and pluripotency factors, such as Klf4, Sox2 and Nanog, in the formation of gene regulatory networks in ESCs, those studies often start with a set of candidate proteins and test their enrichment in a set of chromatin contacts (Apostolou et al., 2013; Denholtz et al., 2013; Schoenfelder et al., 2010b; Wei et al., 2013). Additionally, contacts investigated with 3C-based assays are typically enriched for pairwise chromatin contacts, which might not involve the same transcription factors as multiway contacts. For an unbiased search of potential transcription factors involved in the establishment of multiway gene networks, higher-order contacts between genes could be mined for transcription factor motifs.

Another interesting open question is the abundance and specificity of inter-chromosomal contacts in the mESC genome. While there are several studies reporting inter-chromosomal contacts (Cairns et al., 2016; Fanucchi et al., 2013; Hacisuleyman et al., 2014; Lomvardas et al., 2006; Mifsud et al., 2015; Nagano et al., 2017; Schoenfelder et al., 2010b; Spilianakis et al., 2005), they are often understudied in 3C-based work due to their low abundance in proximity ligation assays. GAM offers the possibility to explore contacts between chromosomes and determine their contact probabilities using SLICE. Knowing how many cells engage in specific inter-chromosomal contacts is relevant, and might help clarify contradicting results from Hi-C and imaging (Maass et al., 2018a). Here, an interesting exploration would be to determine at which physical distances inter-chromosomal contacts occur predominantly. At the moment, we use SLICE with a distance threshold for interacting loci of 100 nm. Live cell imaging showed that inter-chromosomal contacts can have larger spatial distances than intrachromosomal contacts (Maass et al., 2018a), thus applying different distance thresholds could help explore different kinds of contacts. An example for inter-chromosomal contacts is the association of active genes from different chromosomes at splicing speckles (Quinodoz et al., 2018). Explorations of different spatial distances could help distinguish inter-chromosomal interactions with different functions, or even identify the contacts that are involved in direct, and presumable protein-mediated interactions between genomic regions from those that associate with the same nuclear body without direct interaction between the genomic regions. This exploration would lead to genes or super-enhancers that are interacting with different spatial distances, which can then be tested for their effect on transcription of nearby genes using genome editing. The deletion of a super-enhancer with a spatial distance of 90 nm might have different effects than the deletion of an enhancer within 300 nm distance and could provide systematic insights into the importance of spatial distance between regulatory elements and target regions within the nucleus, as well as direct evidence for the functionality of inter-chromosomal or very long-range contacts. The exploration of transcription factors involved in these contacts would also be interesting, especially as different transcription factors might contribute to contacts of different spatial distances, which could depend on the size or abundance of proteins connecting genomic regions. Defined questions about chromatin contacts based on spatial distance have so far not been addressed on a genome-wide scale and could provide unique information about genome folding that is currently not considered with any other genome-wide chromatin contact assay.

6.2 Chromosome topology of the parental alleles

6.2.1 GAM can be used to study allele-specific chromosome folding

The work presented in this thesis showed for the first time that GAM can map chromatin contacts with allele specificity. Using a hybrid mESC line with a high sequence variance between the parental alleles, we were able to phase and assign 80 % of the genomic regions detected in GAM data to the maternal or the paternal genome sequence. Other genome-wide techniques that have assigned chromatin contacts to their haplotypes usually phase a significantly lower fraction of their data. In Hi-C, phasing of chromatin contacts using a female mESC line with the same parental strains as F123, and with the same SNP density, resulted in a phasing efficiency of 26 to 38 % (Giorgetti et al., 2016). This large difference in efficiency can be explained by technical differences between the techniques. Hi-C interaction frequencies are measured by counting how many unique read pairs are found connecting two restriction fragments both of which contain a SNP. The number of phased interactions in a Hi-C dataset is the number of phased unique read pairs. For Hi-C and GAM, the percentage of phased reads is very similar, in Giorgetti et al. (2016) up to 38 % phased reads were reported, here we report 37 % phased reads. However, the strength of phasing contacts with GAM lies in the next step. Chromatin contacts in GAM are not based on single read information but on the co-segregation frequency of positive windows, the detection of which results from many reads. Using a relatively small proportion of phased reads, we were then able to assign positive GAM windows to their parental alleles, as the majority of them contained phased reads from only one, the maternal or the paternal allele. This increased the phased data from 37 % of reads to 80% of positive windows (Figure 4.9). Further, we compared our results with observations made in cryoFISH, which showed that about 10 % of nuclear profiles contain both alleles of a genomic region covered by DNA-FISH probes (Lavitas, 2011). We could reproduce this percentage in the phased GAM data, and showed that in ~8 % of profiles we could positively identify the two alleles. This is also interesting as it may allow to search for correlations between the contacts established by the two alleles in the same cell; this will of course be limited by the low percentage of samples containing both maternal and paternal alleles. The high efficiency of phasing allowed us to study chromatin contacts at 50 kb resolution in the 1123 NP dataset, with the potential to increase the resolution in the larger dataset that combines nuclear profiles from 1NP and 3NP datasets.

6.2.2 The abundance of allele-specific chromatin contacts in the genome

Using our 1NP GAM dataset, we found that chromatin contacts of the parental alleles are different at many regions in the F123 genome. On its own this is a remarkable observation, as several studies using Hi-C have so far suggested otherwise. Except for contact differences on the X chromosome of female cells due to X inactivation (Giorgetti et al., 2016), few differences between alleles have been reported. Exceptions can be found at some imprinted genes. For instance, several studies in human cells suggest allele-specific CTCF-mediated loops at the imprinted genes *H19* and *IGF2* (Dixon et al., 2015; Rao et al., 2014; Tan et al., 2018b). In mice, a recent study finds allele-specific contacts using 4C at the murine *H19/Igf2* locus in ESCs, accompanied by allele-specific DNA methylation and CTCF binding at the imprinting control region (ICR) immediately upstream of *H19* (Llères et al., 2019). Other loci have been described to display allele-specific contacts, such as the immunoglobulin heavy chain locus (Holwerda et al., 2013) and other imprinted loci, such as the *Dlk1-Dio3* locus

(Llères et al., 2019). Reassessment of allele-specific Hi-C data has recently shown a generally higher likelihood for imprinted genes to be found at differential chromatin loops (Greenwald et al., 2019). Overall, only a few structural changes between the parental alleles have been reported in genome-wide comparisons of allele-specific chromatin contacts. While Dixon et al. (2015) and Rao et al. (2014) find large structural homogeneity between the alleles, Greenwald et al. (2019) report slight differences in chromatin loop intensities between alleles, that were moderately connected to allele-specific expression. Chen et al. (2017) find few differences in A/B compartments between the alleles (0.6 % of the genome had allele-specific A or B compartments). Another study in mESCs suggests more changes between the compartments of the maternal and paternal allele, although the percentage of the genome with monoallelic compartments is not reported (Rivera-Mulia et al., 2018). The differences between alleles observed in this latter study were not found connected to parental origin (maternal / paternal), but rather to the genetic background of the allele. Further, while differences in A/B compartments between alleles did not correlate with allele-specific expression, they correlated with asynchronous replication timing between alleles. Asynchronous replication timing occurs in 12 % of the mESC genome with the genetic background of F123. Upon differentiation to neuronal precursor cells (NPCs) differences between alleles are reduced to 1 % of the genome, and alongside many regions that previously had allele-specific differences in replication timing and compartments are now synchronised. However, the supplementary material in Rivera-Mulia et al. suggests that the actual number of allele-specific compartments does not change drastically between ESCs and NPCs. This suggests that there are regions in the genome with altered chromosome topologies between the alleles that cannot be explained with asynchronous replication timing alone.

With GAM, we found that 25 % of the genome has allele-specific A/B compartments in F123 mESCs (Figure 4.11). This percentage is higher than what has been reported before, and has not been replicated yet. To confirm these structural changes between the alleles, the analysis of allele-specific chromatin contacts first has to be reproduced in the 3-NP GAM dataset presented in this thesis. Some differences in A/B compartments were also found between the unphased 1NP and 3NP F123 datasets and although those occurred less often than differences between alleles, it indicates a certain variability in A/B compartments that is not only inherent to allele-specific compartments. Truly different compartments between the alleles should therefore appear reproducibly in the phased 1-NP and 3-NP dataset. Nonetheless, in line with Rivera-Mulia et al. (2018), a first exploration of those monoallelic compartments showed no direct correlation with gene expression differences between alleles. However, the connection between allele-specific gene expression and mono-allelic compartments could be limited to genomic regions with certain features; for instance, gene-poor compartments might be affected by the allele-specific expression of one gene inside the compartment, while gene-dense regions are indifferent to this expression difference. Thus, a refined analysis of allele-specific expression and compartments might identify regions where indeed the two features are connected. Generally, the expression of genes in allele-specific compartments differs from biallelic compartments. While biallelic A compartments primarily contain active genes, and biallelic B compartments tend to entail inactive genes, the allele-specific compartments have the same percentage of active and inactive genes (Figure 4.12). This could suggest that genomic regions with equal amounts of active and inactive genes are less strict with their association to the A or B compartment. Further, previous work identified asynchronous replication timing in F123 mESCs (Rivera-Mulia et al., 2018). The data is publicly available and could be used to identify compartment changes that are connected to replication timing differences in the future analyses of the F123 GAM datasets produced here.

6.2.3 Chromatin contacts at the imprinted *H19/Igf2* locus

To begin exploring specific chromatin contact differences between alleles, I chose to investigate the topology of the previously described *H19/Igf2* locus in F123 mESCs. The imprinted control region immediately next to *H19* has been described to form maternal-specific interactions in mESCs with a downstream region called the *H19/Igf2* distal anchor domain (HIDAD) within the borders of the *H19/Igf2* TAD (Llères et al., 2019; Rao et al., 2014; Tan et al., 2018b). This interaction leads to maternal-specific expression of *H19* (Llères et al., 2019). On the paternal allele, HIDAD interacts with the promoter of *Igf2* which leads to paternal-specific expression of *Igf2*. *H19* and *Igf2* are located 35 kb apart from each other, which does not allow the clear identification of the allele-specific contacts of *H19* and *Ifg2* with HIDAD in the 50 kb GAM data. However, the chromatin conformation of the larger *H19/Igf2* locus is clearly different between the two alleles. Our GAM data shows a clear directionality of contacts of the *H19/Igf2* locus to the downstream imprinted gene cluster containing many maternally active genes. These interactions are specific for the maternal allele. On the paternal allele, interactions with other imprinted genes are rare, and the *H19/Igf2* locus specifically interacts with upstream regions. These changes are clearly visible in the allele-specific contact maps, where the entire locus displays allele-specific domain structures, suggesting a reorganisation of TADs at the *H19/Igf2* locus depending on which genes are active or silent in each allele. The observed contact differences also appear amongst the top 5 % of allele-specific contacts.

Previous work described differential binding of CTCF at the *H19/Igf2* locus, specifically at the ICR next to *H19* (Llères et al., 2019). It is possible that the differential binding of CTCF can alter the chromatin contacts of the ICR not only locally (Llères et al., 2019), but might have larger effects on chromosome topology, even to the extent of changing TAD structures at the locus. CTCF has been shown to be important for the manifestation of TAD borders at a number of disease-associated gene loci (Hnisz et al., 2016; Ji et al., 2016). Genome-wide assays have shown that CTCF is enriched at the anchors of loops and the boundaries of TADs (Dixon et al., 2012; Rao et al., 2014). Depletion of cohesin, which binds to CTCF to facilitate loop formation, has been shown to alter the positions of TAD borders in single cells (Bintu et al., 2018). The observations at the *H19/Igf2* locus are intriguing, as it suggests that different TAD borders can also occur at different alleles in the same cell. The loss of monoallelic DNA methylation at this CTCF site could also be studied in the case of two congenital diseases associated with gain or loss of methylation at the ICR of *H19* and *Igf2* (Nativio et al., 2011). In Silver-Russel syndrome (SRS) and Beckwith-Wiedemann syndrome (BWS), *H19* and *Igf2* lose their allele-specific expression, respectively. One important question is why these allele-specific domains and long-range contacts have not been captured in previous studies using Hi-C or other 3C-based techniques at the locus. Thus, experiments to validate the presence of the paternal *H19/Igf2* TAD boundary in the GAM data should be done, for example by phasing the 3NP data to replicate this conformation in a second GAM dataset and further by DNA-FISH experiments with allele specificity.

6.2.3 The regulation of allele-specific gene expression

Monoallelic expression is the preferential expression of genes on only one of the parental alleles. In mammals, there are different forms of monoallelic expression. Random monoallelic expression describes the process of cell type- or tissue-specific expression of a gene with allele specificity. Here, we found 0.5 % of all transcribed, protein-coding genes have significant and strong allele-specific expression, in line with published observations (Eckersley-Maslin et al., 2014). Other forms of monoallelic expression include X-chromosome inactivation in females, genomic imprinting, and allelic exclusion. Male ESCs do not have X-inactivation or allelic exclusion, but show imprinting, the process of inactivating or ‘imprinting’ one allele of a gene in all cells and tissues. Imprinting usually involves the differential DNA methylation of the parental alleles, either at the promoter of the imprinted gene, or at *cis*-regulatory elements in proximity to the gene (Weaver and Bartolomei, 2014). At some loci, differential methylation is connected to differential chromatin looping between the alleles (Llères et al., 2019; Rao et al., 2014; Tan et al., 2018b). Here, we report that allele-specific chromatin contacts occur at almost all imprinted genes, and although we have not yet explored the functionality of these contacts, we showed an example of allele-specific TAD structures at the imprinted *H19/Igf2* locus, which suggested a role for monoallelic binding of CTCF in the formation of these contacts.

Not only imprinted genes, but also genes with random monoallelic expression, are found associated with allele-specific chromatin contacts in the GAM data. To begin exploring the function of these contacts, we investigated the presence of active chromatin features. Here, preliminary explorations detected no difference in the occupancy of active histone marks or CTCF at regions contacting genes with mono- or biallelic gene expression. In line with our results, Chen et al. (2017) report no difference in enrichment for active chromatin marks and CTCF, when examining genes with allele-specific expression and their immediate genomic surrounding. However, we can find examples with allele-specific enhancer contacts, identified with H3K27ac, at monoallelic genes, which could explain some of the allele-specific expression in F123 (Figure 4.15).

For further investigation of allele-specific gene regulation and chromatin contacts, several considerations have to be made. First, DNA methylation of the *cis*-regulatory elements, including enhancers and promoter of a gene or of nearby CTCF sites could be considered. In mammals, DNA methylation occurs at cytosine residues and is found primarily at CpG sites (cytosine followed by guanine). Its allele-specificity has been shown for imprinted loci (Weaver and Bartolomei, 2014). An exploration of these regions with respect to chromosome topologies and expression is interesting, however the number of allele-specific DMRs is low (Tomizawa et al., 2011), and though important for imprinting, a direct role for DNA methylation in the establishment of random mono-allelic expression is yet missing (Eckersley-Maslin et al., 2014). Nevertheless, an exploration of reported DMRs could identify allele-specific genes and chromatin contacts connected to DNA methylation, which will provide insight into the impact of this repressive mark on allele-specific topologies and expression.

Not only DNA methylation has repressive functions, but also the Polycomb Group proteins. Polycomb repressive complexes (PRCs) are known to bind to and repress genes throughout the ESC genome, in many cases to silence developmental genes (Azuara et al., 2006; Boyer et al., 2006; Jorgensen et al., 2006). One proposed function of PRCs in allele-specific expression is the inactivation of genes on one of the X chromosomes in females involving Xist-mediated silencing (Kohlmaier et al., 2004; Mak et al., 2004). In imprinting, Polycomb complexes are recruited allele-specifically to some imprinted loci, for example at the *Kcnq1* gene, where Polycomb mediated silencing is thought to be mediated by a ncRNA (*Kcnq1ot1*) (Mancini-Dinardo et al., 2006; Terranova et al., 2008). A role for Polycomb in random monoallelic expression has not been described yet. In contrast, the histone marks H3K4me2/3 and H3K9me3 have been found at the active and inactive alleles of genes with monoallelic expression, respectively (Eckersley-Maslin et al., 2014). A first comparison of genes with allele-specific expression and described states of gene promoters in mESCs (Ferrai et al., 2017) showed that 18% of promoters of allele-specific genes is marked by H3K27me3 (Figure 4.5). While differences between mESC lines and cell culture conditions have to be considered, this association suggest that Polycomb does not play a general role at defining the repressed states of promoters of allele-specific genes. Polycomb regulation may also drive allele-specific expression, for instance via monoallelic silencing of enhancers, an interesting topic that has not been addressed in the literature.

6.2.4 Nuclear positioning of allele-specific genes

Another layer of gene regulation is the positioning of genes inside the nucleus. The heterochromatic environment of the nuclear periphery can have repressive effects on genes (Finlan et al., 2008; Reddy et al., 2008), and association with the nuclear lamina is dynamic throughout differentiation (Peric-Hupkes et al., 2010). The two alleles of a gene can have different radial positions inside the nucleus. While some examples show that indeed monoallelic expression of a gene is associated to its allele-specific positioning inside the nucleus (Takizawa et al., 2008), the genome-wide prevalence of this mechanism is poorly studied. In GAM, radial positioning of genomic loci can be obtained via the genome coverage of nuclear profiles containing the locus of interest. Nuclear profiles with the largest genome coverage tend to come from the centre of the nucleus, while profiles with lowest genome coverage come from the apical parts of the nucleus (Beagrie et al., 2017). Here, we have the possibility to study the allele-specific radial positioning of genes with monoallelic expression genome-wide and also combine the information of their radial position with the knowledge of other chromosome topologies, such as association with A and B compartments, or the specific chromatin contacts of a gene. This could address questions about the dependencies of chromatin topologies, for instance if an allele has different associations with enhancers depending in its radial position inside the nucleus.

6.2.5 Chromatin contacts mediated by transcription factors

Another approach to understand chromatin contacts and what drives them is to study the clustering of transcription factors (TFs) in 3D space. Regulated genes and *cis*-regulatory regions are often bound by cell-type specific TFs (Whyte et al., 2013). In ESCs, core pluripotency factors, including Oct4, Nanog, Sox2, and Klf4 drive the expression of key cellular identity genes in the ESC genome, while they repress developmental genes to prevent ESCs from differentiating (reviewed in (Young, 2011)). Interestingly, genes and their regulatory elements, but also groups of genes with similar transcription levels, contact each other when bound by the same TFs (Apostolou et al., 2013; de Wit et al., 2013; Denholtz et al., 2013; Wei et al., 2013). Genomic regions that are marked by co-occupancy of pluripotency factors tend to preferentially contact each other, suggesting a critical role for pluripotency factors to shape the nuclear architecture of ESCs. The clustering of genomic regions can be observed for DNA occupied by the same TFs, as well as for combinations of different TFs, that are thought to bring together different parts of the genome via physical protein-protein interactions (Ma et al., 2018b).

Here, we explored specific chromatin contacts in mESCs that are common to the alleles and contacts that are specific to the paternal or the maternal allele. We searched for enriched transcription factor motifs in open chromatin regions involved in specific chromatin contacts, and determined which combinations of those motifs are enriched in specific contact of the alleles or contacts that are strong on both alleles. The first observation was that allele-specific and common contacts share the same TF motif pairs. Interestingly, the most prominent combinations of TF motifs at allele-specific binding sites were from the architectural protein CTCF, and several transcriptional regulators, namely Sall1, Ubp1, Maz, and further Smad3, Klf15 and Klf5 (Table 4.3). Some of these TFs have known functions in the maintenance of pluripotency, in particular Sall1. Sall1 is part of the pluripotency regulatory network and many of its genomic targets are also bound by Nanog (Karantzali et al., 2011). Binding of Sall1 leads to transcriptional repression of developmental genes, and specifically supresses mesodermal and ectodermal differentiation in ESCs. Others, such as Maz or Klf15 are expressed in several tissues. Maz is a ubiquitously expressed TF and a regulator of several genes in different cell types, including *c-Myc* (Bossone et al., 1992). Further, Maz can function both as activator and repressor and bind to the same *cis*-regulatory elements as Sp1 (Song et al., 2003), which is directly interacting with Nanog and Oct4 (Wang et al., 2006). UBIP1/LBP-1 was studied extensively in its role as repressor of human immunodeficiency virus type 1 (HIV-1) transcription, however its functions in ESCs are largely unknown. The combination of several transcriptional regulators with CTCF suggests that bi- and monoallelic chromatin contacts bring together specific gene-regulatory networks, for example developmental genes that are silenced in pluripotent cells. Further analysis of chromatin contacts associated to specific TFs could include the exploration of genes in those contacts. Are they indeed targets of the TF in question and, if data is available, do they change expression in knockout experiments of the respective TF? Further, ChIP-seq experiments might have been performed in mESCs for some of these TFs, which could help identify which binding sites are occupied by the TF. For CTCF, ChIP-seq data is available in F123 mESCs, which not only allows us to investigate the presence of CTCF at its binding sites, but further helps explore allele-specific differences in TF occupancy. Since we found that monoallelic contacts are enriched for the same binding sites for TFs than biallelic contacts, albeit at different frequency, the allele specificity of a given interaction might be connected to the actual presence of the TF at its binding site. Allele-specific binding has been described for CTCF in the context of imprinting, and indeed we identified large-scale structural changes at a described monoallelic CTCF site. With the SNP information from F123, we should be able to phase the CTCF ChIP-seq data and explore the extent of this phenomenon. Further, CTCF has been proposed to be connected to differential gene expression, not only of imprinted genes, but of genes with random monoallelic expression (Greenwald et al., 2019). At many of those genes, loops between CTCF sites had mild differences in contact frequencies of the alleles. However, these differences were almost unnoticeable and have not been reported in the first publication of this data, suggesting that either the assay (Hi-C) used to identify the differential loops was not suited for the purpose, or that other mechanisms play more important roles in the manifestation of monoallelic expression. Furthermore, an enrichment for CTCF binding sites with sequence variants at and around genes with allele-specific expression has been shown (Chen et al., 2017). This opens the possibility that the same TFs regulate mono- and biallelic chromatin contacts, however, SNPs within their binding sites could influence the ability of the TF to bind to the respective site and thus alter the chromatin contact. One approach to explore this possibility without collecting additional ChIP-seq data would be to search the TF binding sites identified here for sequence variance. This could indicate if allele-specificity of contacts is connected to differential binding of TFs to the alleles.

6.3 The effects of structural rearrangements at the 16p11.2 locus

The compartmentalisation of chromosomes into domains, such as TADs or DNA loops, play important roles in gene regulation and, if disrupted, can lead to disease (Spielmann et al., 2018). Chromosomes themselves are organised in chromosome territories, with varying proportions of intermingling (Branco and Pombo, 2006). While specific and regulatory contacts within chromosomes are the subject of many studies, few specific contacts between chromosomes have been reported and their functions are still debated. An interesting example of inter-chromosomal contacts with regulatory functions are the interactions of an enhancer on chromosome 14 with olfactory receptor genes on different chromosomes, which regulate the expression of individual olfactory receptors in different sensory neurons (Lomvardas et al., 2006). While evidence of contacts between chromosomes increases (Cairns et al., 2016; Fanucchi et al., 2013; Hacisuleyman et al., 2014; Horta et al., 2018; Mifsud et al., 2015; Nagano et al., 2017; Schoenfelder et al., 2010b; Spilianakis et al., 2005), little is known about the impact of inter-chromosomal contacts in disease or how structural rearrangements of the genome affect these contacts. In the context of several structural rearrangements changes in inter-chromosomal contact frequencies have been reported, for example for of the human 2q37 deletion (Maass et al., 2018b), the 22q11.2 and 1q21.1 CNVs (Zhang et al., 2018a), or 16p11.2 CNVs (Loviglio et al., 2016). A role for changes in inter-chromosomal contacts in the manifestation of disease has been proposed for some of these CNVs, such as the 2q37 deletion (Maass et al., 2018b). Here, we show that a deletion at the 16p11.2 locus, associated with autism spectrum disorder (ASD), results in the rearrangement and loss of specific inter-chromosomal contacts between the 16p11.2 locus and chromosome 18. We propose a role for these changes of inter-chromosomal contacts in the upregulation of the nearby *Pcdhb* gene cluster, which comprises protocadherin genes with important functions in neuronal development.

6.3.1 Upregulation of protocadherin genes in the context of the 16p11.2 deletion

CNVs at the human 16p11.2 locus are associated with various phenotypes, most importantly with a high prevalence of ASD, as well as an increased risk for intellectual disabilities and psychiatric disorders and further, abnormal head circumference (Kumar et al., 2008; Niarchou et al., 2019; Steinman et al., 2016). We investigated the most common human deletion at the

16p11.2 locus, using as model system a previously described mouse line carrying the 16p11.2 deletion that phenocopies the human disease (Horev et al., 2011). When studying gene expression changes in the context of the deletion and its dosage-compensated control, a cell line carrying the 16p11.2 deletion on one allele and the respective duplication on the other allele, we discovered the significant upregulation of a cluster of protocadherin genes in neuronal precursor cells (NPCs), but not in ESCs (Figure 5.11). Most protocadherin genes are organised in clusters in the genome, and have unique functions in the development of the nervous system. Protocadherins allow self-recognition of neurons and thus prevent adhesion of dendrites and axons from the same cell (Rubinstein et al., 2015). Self-avoidance is achieved by expressing diverse combinations of protocadherins at the cell surface. Further, unique combinations of protocadherins at the cell surface regulate the recognition and cell-to-cell adhesion of neurons (Thu et al., 2014). The discovery of the misexpression of *Pcdh* genes in the context of the 16p11.2 deletion is exciting, as earlier studies have proposed that *Pcdh* genes could be connected to the autism phenotype. Deficits in synaptic development and impaired neuronal connectivity have been linked to ASD (Belmonte et al., 2004). Altered connections between neurons can be explained by the misexpression of protocadherins, for example, if the stochasticity of their expression is lost, the recognition of different neurons or even self-recognition might be impaired. Further, examination of disease-associated SNPs identified ASD-related SNPs in *Pcdha* genes (Anitha et al., 2013).

Clustered protocadherins are arranged in tandem and are located on chromosome 18 in mice and chromosome 5 in humans (Wu and Maniatis, 1999). The *Pcdha*, *Pcdhb*, and *Pcdhg* clusters each have 18 to 22 exons encoding for different protocadherins, which in the case of *Pcdha* and *Pcdhg* are combined with a common promoter for each cluster, while *Pcdhb* exons are transcribed each from their own promoter (Tasic et al., 2002). To achieve unique combinations of protocadherins in single cells, different genes of the *Pcdh* clusters are expressed stochastically in individual neurons (Esumi et al., 2005; Kaneko et al., 2006). Variable exon usage, alternative splicing, and monoallelic expression allow the expression of a large diversity of protocadherins, a process that is regulated by different epigenetic mechanisms, such as DNA methylation (Toyoda et al., 2014).

In NPCs carrying the heterozygous 16p11.2 deletion, the *Pcdhb* cluster is upregulated, accompanied by downregulation of the *Pcdhg* cluster. In the dosage compensated deletion, the *Pcdhb* and the *Pcdha* cluster are upregulated. All three clusters have been described to be regulated by different enhancer elements (Guo et al., 2012; Ribich et al., 2006; Yokota et al., 2011). Here, we focused on the upregulation of the *Pcdhb* cluster and found that indeed *Pcdhb* interacts with its known *cis*-regulatory element in NPCs, and further that this interaction occurs more frequently in cells where *Pcdhb* is upregulated (Figure 5.12). The deletion of this *Pcdhb* enhancer leads to downregulation of *Pcdhb* genes, and also some isoforms of the *Pcdhg* genes (Yokota et al., 2011), which suggests a role of this enhancer in the observed phenotype. The interaction of the enhancer and the genes in the *Pcdhb* cluster is mediated by CTCF, however the enhancer and the promoters of *Pcdh* genes are also negatively regulated to allow stochastic and monoallelic expression in different neurons (Gendrel et al., 2013; Hu, 2016). The monoallelic expression of *Pcdha* and *Pcdhb* is regulated by Smchd1, a protein required for the methylation of the inactive X chromosome and several imprinted loci, including genes at the Prader-Willi syndrome locus on chromosome 7 (Gendrel et al., 2013). Knockout of *Smchd1*, similar to the observed phenotype in the 16p11.2 mutants, leads to upregulation of genes in the *Pcdha* and *Pcdhb* cluster, but not the *Pcdhg* cluster. These observations suggest that Smchd1 could be involved in the upregulation of the *Pcdhb* cluster, however its expression was not significantly altered in NPCs carrying the 16p11.2 deletion. It can be speculated whether Smchd1 was misexpressed at an earlier stage in differentiation than the timepoint of our investigation (NPCs, day 5). It has been shown that the expression of *Pcdh* genes is regulated early during neuronal development, and once established does not change in response to depletion of Smchd1 in later timepoints (Hu, 2016).

6.3.2 Inter-chromosomal contact of the 16p11.2 locus

While the modulation of transcription by *trans*-acting elements, such as transcription factors is one possible scenario for the effects observed at the *Pcdh* gene clusters, we also explored the possibility of reorganisation of chromosome topologies in a connection to the expression differences of protocadherins. An exploration of inter-chromosomal contacts in Hi-C data from NPCs revealed specific inter-chromosomal contacts between the larger 16p11.2 locus on chromosome 7 and an upstream region of the *Pcdh* clusters on chromosome 18. The regions on both chromosomes that contact each other specifically in wildtype NPCs but lose their interaction in cells with the deletion of the 16p11.2 locus (Figure 5.14) contain super-enhancers in the developing mouse brain. This observation could indicate that also in NPCs these regions are super-enhancers. Super-enhancers are clustered enhancer regions with a high occupancy of transcription factors, mediator, and H3K27ac (Whyte et al., 2013), and have been shown to preferentially contact with other super-enhancers in higher-order contacts (Beagrie et al., 2017). The association of multiple super-enhancers has been proposed to occur predominantly in condensates, stable compartments in the nucleus formed by liquid-liquid phase separation. Several proteins have been shown to form condensates associated with super-enhancers, such as mediator and RNA polymerase II (Cho et al., 2018) or the transcriptional coactivators BRD4 and MED1 (Sabari et al., 2018). Phase separation has further been suggested to play a role in the formation of many of the membrane-free compartments inside the nucleus, such as splicing speckles. At splicing speckles, active regions from different chromosomes engage in multiway contacts (Quinodoz et al., 2018), and further a large proportion of super-enhancers can be found at speckles (Chen et al., 2018b). The strong association of the two regions from chromosome 7 and 18 containing potential super-enhancers could indicate the association with the same nuclear condensate, potentially the splicing speckle. It can be speculated how exactly the removal of the 16p11.2 locus from an inter-chromosomal contact hub would affect other regions inside the hub. One possible scenario could be that in the absence of the 16p11.2 locus, the corresponding region on chromosome 18 changes its localisation with respect to other chromosomes and is more accessible for regions within its own chromosome territory, such as the *Pcdh* gene clusters immediately downstream of the potential super-enhancer. In fact, the *Pcdh* clusters are strongly depleted of contacts with the super-enhancer region (Figure 5.15). The depletion is not lost upon deletion of 16p11.2, however it is slightly but significantly reduced. The observed difference could be explained by stochastic contacts of the *Pcdh* clusters with the inter-chromosomal hub in a small proportion of the cell population. Super-enhancers have been shown to be strong transcriptional activators that are usually insulated from unrelated nearby genes (Dowen et al., 2014; Wang et al., 2014). Their targets are often found in insulated gene deserts, which have been termed super-enhancer domains and the borders of these super-enhancer domains coincide with TAD boundaries (Dowen et al., 2014; Wang et al., 2014). The disruption of these domains can lead to the activation of neighbouring genes (Dowen et al., 2014; Ji et al., 2016), similar to what we observe at the *Pcdhb* cluster.

6.3.3 Future directions and clarifying experiments

The exploration of the effects of introducing a deletion at the 16p11.2 locus associated with ASD have led to a hypothesis that explains gene expression differences of protocadherin genes on chromosome 18 with changes in inter-chromosomal contact frequencies, presumable between super-enhancers. To confirm this potentially disease-associated mechanisms, a few follow up experiments would be required. First, we were able to detect inter-chromosomal interactions between the 16p11.2 locus and chromosome 18 in our Hi-C data and in published data of NPCs, which already showed the reproducibility of this contact (Figure 5.13). The loss of this interaction upon 16p11.2 deletion however should ideally be confirmed with an orthogonal method to detect chromatin contacts, such as DNA-FISH. Inter-chromosomal contacts have been shown in some cases to be captured better with imaging than with 3C-based techniques (Maass et al., 2018a), which promises clearer results, and at the same time would tell us the prevalence of the 16p11.2 inter-chromosomal contact in the cell population. Furthermore, we suggested that super-enhancers might be involved in the formation of this inter-chromosomal contact. As a proxy for the presence of super-enhancers we used identified super-enhancer positions in the mouse brain (E14.5), which does not guarantee the presence of super-enhancers in NPC of our differentiation system. Therefore, identification of super-enhancers in NPCs is important to clarify this point. Here, ChIP-seq for an active enhancer marks, such as H3K27ac could be used to identify super-enhancers. We proposed that upregulation of *Pcdh* genes might be the result of spurious interactions, or the loss of insulation, between a nearby super-enhancer and the *Pcdh* gene clusters. Therefore, a fascinating experiment would be to engineer a contact between the *Pcdh* clusters and this nearby super-enhancer. A technique to enforce interactions between genomic regions has been described recently (Kim et al., 2019). In the light-activated-dynamic-looping (LADL) system two synthetic architectural proteins are introduced into the genome, which heterodimerise upon activation with blue light. Here, an excellent anchor point for an engineered loop with the *Pcdh* gene clusters would be the enhancer of *Pcdhb*, which showed increase contact frequencies with *Pcdhb* genes upon deletion of 16p11.2, and is able to activate all genes in the cluster (Yokota et al., 2011).

7. References

Ahmed, K., Dehghani, H., Rugg-Gunn, P., Fussner, E., Rossant, J., and Bazett-Jones, D.P. (2010). Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. PLoS One *5*, e10531-10543.

Aitchison, J.D., and Rout, M.P. (2015). The interactome challenge. The Journal of Cell Biology *211*, 729.

Allahyar, A., Vermeulen, C., Bouwman, B.A.M., Krijger, P.H.L., Verstegen, M.J.A.M., Geeven, G., van Kranenburg, M., Pieterse, M., Straver, R., Haarhuis, J.H.I.*, et al.* (2018). Enhancer hubs and loop collisions identified from single-allele topologies. Nature Genetics *50*, 1151-1160.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England) *31*, 166-169.

Anderson, B.M., Schnetz-Boutaud, N., Bartlett, J., Wright, H.H., Abramson, R.K., Cuccaro, M.L., Gilbert, J.R., Pericak-Vance, M.A., and Haines, J.L. (2008). Examination of association to autism of common genetic variationin genes related to dopamine. Autism Res *1*, 364-369.

Andrey, G., Schopflin, R., Jerkovic, I., Heinrich, V., Ibrahim, D.M., Paliou, C., Hochradel, M., Timmermann, B., Haas, S., Vingron, M.*, et al.* (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. Genome Res *27*, 223-233.

Anitha, A., Thanseem, I., Nakamura, K., Yamada, K., Iwayama, Y., Toyota, T., Iwata, Y., Suzuki, K., Sugiyama, T., Tsujii, M.*, et al.* (2013). Protocadherin α (PCDHA) as a novel susceptibility gene for autism. J Psychiatry Neurosci *38*, 192-198.

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L.*, et al.* (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell *12*, 699-712.

Apostolou, E., and Thanos, D. (2008). Virus Infection Induces NF-kappaB-dependent interchromosomal associations mediating monoallelic IFN-beta gene expression. Cell *134*, 85-96.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jorgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M.*, et al.* (2006). Chromatin signatures of pluripotent cell lines. Nat Cell Biol *8*, 532-538.

Banani, S.F., Rice, A.M., Peeples, W.B., Lin, Y., Jain, S., Parker, R., and Rosen, M.K. (2016). Compositional Control of Phase-Separated Cellular Bodies. Cell *166*, 651-663.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299-308.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell *144*, 214-226.

Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., de Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat Struct Mol Biol *24*, 515-524.

7. References

Bariselli, S., Hörnberg, H., Prévost-Solié, C., Musardo, S., Hatstatt-Burklé, L., Scheiffele, P., and Bellone, C. (2018). Role of VTA dopamine neurons and neuroligin 3 in sociability traits related to nonfamiliar conspecific interaction. Nature Communications *9*, 3173.

Barlow, D.P., Stoger, R., Herrmann, B.G., Saito, K., and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature *349*, 84-87.

Bartolomei, M.S. (2009). Genomic imprinting: employing and avoiding epigenetic processes. Genes Dev *23*, 2124-2133.

Bartolomei, M.S., Zemel, S., and Tilghman, S.M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153-155.

Barutcu, A.R., Lajoie, B.R., McCord, R.P., Tye, C.E., Hong, D., Messier, T.L., Browne, G., van Wijnen, A.J., Lian, J.B., Stein, J.L., *et al.* (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol *16*, 214.

Barutcu, A.R., Maass, P.G., Lewandowski, J.P., Weiner, C.L., and Rinn, J.L. (2018). A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. Nat Commun *9*, 1444-1455.

Beagrie, R.A., and Schueler, M. (2017). GAMtools: an automated pipeline for analysis of Genome Architecture Mapping data. bioRxiv, 114710.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.M., Branco, M.R., *et al.* (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519-524.

Belaghzal, H., Dekker, J., and Gibcus, J.H. (2017). Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods *123*, 56-65.

Beliveau, B.J., Boettiger, A.N., Avendano, M.S., Jungmann, R., McCole, R.B., Joyce, E.F., Kim-Kiselak, C., Bantignies, F., Fonseka, C.Y., Erceg, J., *et al.* (2015). Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. Nat Commun *6*, 7147-7160.

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., *et al.* (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc Natl Acad Sci U S A *109*, 21301-21306.

Belmont, A.S., and Straight, A.F. (1998). In vivo visualization of chromosomes using lac operator-repressor binding. Trends Cell Biol *8*, 121-124.

Belmonte, M.K., Allen, G., Beckel-Mitchener, A., Boulanger, L.M., Carper, R.A., and Webb, S.J. (2004). Autism and abnormal development of brain connectivity. The Journal of neuroscience : the official journal of the Society for Neuroscience *24*, 9228-9231.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods *58*, 268-276.

Bintu, B., Mateo, L.J., Su, J.-H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., and Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science *362*, eaau1783-1791.

7. References

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A.S., Yu, T., Marie-Nelly, H., McSwiggen, D.T., Kokic, G., Dailey, G.M., Cramer, P., *et al.* (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. Nat Struct Mol Biol *25*, 833-840.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.T., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418-422.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., *et al.* (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557-572 e524.

Bossone, S.A., Asselin, C., Patel, A.J., and Marcu, K.B. (1992). MAZ, a zinc finger protein, binds to c-MYC and C2 gene sequences regulating transcriptional initiation and termination. Proceedings of the National Academy of Sciences *89*, 7452.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Boyle, S., Rodesch, M.J., Halvensleben, H.A., Jeddeloh, J.A., and Bickmore, W.A. (2011). Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res *19*, 901-909.

Branco, M.R., Branco, T., Ramirez, F., and Pombo, A. (2008). Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. Chromosome Res *16*, 413-426.

Branco, M.R., and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol *4*, e1380780-1380788.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., *et al.* (2012). Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell *10*, 157-170.

Brown, J.M., Green, J., das Neves, R.P., Wallace, H.A., Smith, A.J., Hughes, J., Gray, N., Taylor, S., Wood, W.G., Higgs, D.R., *et al.* (2008). Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. J Cell Biol *182*, 1083-1097.

Brown, J.M., Roberts, N.A., Graham, B., Waithe, D., Lagerholm, C., Telenius, J.M., De Ornellas, S., Oudelaar, A.M., Scott, C., Szczerbal, I., *et al.* (2018). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. Nature Communications *9*, 3849-3863.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Varnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.M., Osborne, C., *et al.* (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol *17*, 127-143.

Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biology *16*, 195.

Chapman, A.R., He, Z., Lu, S., Yong, J., Tan, L., Tang, F., and Xie, X.S. (2015). Single Cell Transcriptome Amplification with MALBAC. PLOS ONE *10*, e0120889.

7. References

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S.*, et al.* (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479-1491.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018a). Dynamic interplay between enhancer-promoter topology and gene activity. Nat Genet *50*, 1296-1303.

Chen, H., Liu, S., Seaman, L., Najarian, C., Wu, W., Ljungman, M., Higgins, G., Hero, A., Wicha, M., and Rajapakse, I. (2017). Parental allele-specific genome architecture and transcription during the cell cycle. bioRxiv, 201715.

Chen, W.V., and Maniatis, T. (2013). Clustered protocadherins. Development (Cambridge, England) *140*, 3297-3302.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E.K., Adam, S.A., Goldman, R., van Steensel, B., Ma, J., and Belmont, A.S. (2018b). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. J Cell Biol *270*, 4025-4048.

Chetverina, D., Aoki, T., Erokhin, M., Georgiev, P., and Schedl, P. (2014). Making connections: insulators organize eukaryotic chromosomes into independent cis-regulatory networks. Bioessays *36*, 163-172.

Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, II (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science *361*, 412-415.

Cremer, T., and Cremer, M. (2010). Chromosome territories. Cold Spring Harb Perspect Biol *2*, a003889-003910.

Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.C., Sanborn, A.L., Machol, I., Shamim, M.*, et al.* (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. Proc Natl Acad Sci U S A *113*, E4504-4512.

Davies, J.O., Oudelaar, A.M., Higgs, D.R., and Hughes, J.R. (2017). How best to identify chromosomal interactions: a comparison of approaches. Nat Methods *14*, 125-134.

Davies, J.O., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R., and Hughes, J.R. (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat Methods *13*, 74-80.

de Wit, E., Bouwman, B.A.M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J.A.M., Krijger, P.H.L., Festuccia, N., Nora, E.P., Welling, M.*, et al.* (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227-231.

de Wit, E., Vos, E.S., Holwerda, S.J., Valdes-Quezada, C., Verstegen, M.J., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Mol Cell *60*, 676-684.

DeAngelis, M.M., Wang, D.G., and Hawkins, T.L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. Nucleic acids research *23*, 4742-4743.

DeChiara, T.M., Robertson, E.J., and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. Cell *64*, 849-859.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

7. References

Deng, W., Rupon, J.W., Krivega, I., Breda, L., Motta, I., Jahn, K.S., Reik, A., Gregory, P.D., Rivella, S., Dean, A.*, et al.* (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. Cell *158*, 849-860.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. Cell Stem Cell *13*, 602-616.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W.*, et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardimci, G.G., Chakraborty, A., Bann, D.V., Wang, Y.*, et al.* (2018). Integrative detection and analysis of structural variation in cancer genomes. Nat Genet *50*, 1388-1398.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dostie, J., and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nature Protocols *2*, 988–1002.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C.*, et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res *16*, 1299-1309.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schujiers, J., Lee, T.I., Zhao, K.*, et al.* (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374-387.

Dundr, M., and Misteli, T. (2010). Biogenesis of nuclear bodies. Cold Spring Harb Perspect Biol *2*, a000711-000725.

Eckersley-Maslin, M.A., Thybert, D., Bergmann, J.H., Marioni, J.C., Flicek, P., and Spector, D.L. (2014). Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev Cell *28*, 351-365.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452-455.

Esumi, S., Kakazu, N., Taguchi, Y., Hirayama, T., Sasaki, A., Hirabayashi, T., Koide, T., Kitsukawa, T., Hamada, S., and Yagi, T. (2005). Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. Nat Genet *37*, 171-176.

Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A.D., and Ren, B. (2016). Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. Cell Res *26*, 1345-1348.

Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M.S., and Mhlanga, M.M. (2013). Chromosomal contact permits transcription between coregulated genes. Cell *155*, 606-620.

Ferguson-Smith, A.C., Cattanach, B.M., Barton, S.C., Beechey, C.V., and Surani, M.A. (1991). Embryological and molecular investigations of parental imprinting on mouse chromosome 7. Nature *351*, 667-670.

Ferrai, C., Torlai Triglia, E., Risner-Janiczek, J.R., Rito, T., Rackham, O.J.L., de Santiago, I., Kukalev, A., Nicodemi, M., Akalin, A., Li, M.*, et al.* (2017). RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation. Molecular Systems Biology *13*, 946.

Ferrai, C., Xie, S.Q., Luraghi, P., Munari, D., Ramirez, F., Branco, M.R., Pombo, A., and Crippa, M.P. (2010). Poised transcription factories prime silent uPA gene prior to activation. PLoS Biol *8*, e1000270-1000281.

Ferreira, J., Paolella, G., Ramos, C., and Lamond, A.I. (1997). Spatial organization of large-scale chromatin domains in the nucleus: a magnified view of single chromosome territories. J Cell Biol *139*, 1597-1610.

Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet *4*, e1000039-1000051.

Finn, E.H., Pegoraro, G., Brandao, H.B., Valton, A.L., Oomen, M.E., Dekker, J., Mirny, L., and Misteli, T. (2019). Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. Cell *176*, 1502-1515.e1510.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandao, H.B., Ulianov, S.V., Abdennur, N., Razin, S.V., Mirny, L.A., and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature *544*, 110-114.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schopflin, R., Kraft, K., Kempfer, R., Jerkovic, I., Chan, W.L.*, et al.* (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265-269.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S.*, et al.* (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol *11*, 852-865.

Fu, Y., Rocha, P.P., Luo, V.M., Raviram, R., Deng, Y., Mazzoni, E.O., and Skok, J.A. (2016). CRISPR-dCas9 and sgRNA scaffolds enable dual-colour live imaging of satellite sequences and repeat-enriched individual loci. Nat Commun *7*, 11707-11714.

Fudenberg, G., and Imakaev, M. (2017). FISH-ing for captured contacts: towards reconciling FISH and 3C. Nat Methods *14*, 673-678.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep *15*, 2038-2049.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H.*, et al.* (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58-64.

Gadow, K.D., Roohi, J., DeVincent, C.J., and Hatchwell, E. (2008). Association of ADHD, tics, and anxiety with dopamine transporter (DAT1) genotype in autism spectrum disorder. J Child Psychol Psychiatry *49*, 1331-1338.

Gall, J.G., and Pardue, M.L. (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc Natl Acad Sci U S A *63*, 378-383.

Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, II, Iarovaia, O.V., and Razin, S.V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res *41*, 3563-3575.

Gendrel, A.-V., Tang, Y.A., Suzuki, M., Godwin, J., Nesterova, T.B., Greally, J.M., Heard, E., and Brockdorff, N. (2013). Epigenetic Functions of Smchd1 Repress Gene Clusters on the Inactive X Chromosome and on Autosomes. Molecular and Cellular Biology *33*, 3150.

Germier, T., Sylvain, A., Silvia, K., David, L., and Kerstin, B. (2018). Real-time imaging of specific genomic loci in eukaryotic cells using the ANCHOR DNA labelling system. Methods *142*, 16-23.

Gibcus, J.H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M., Xie, L., Paulson, J.R., Earnshaw, W.C.*, et al.* (2018). A pathway for mitotic chromosome formation. Science (New York, NY) *359*, eaao6135-6146.

Giorgetti, L., and Heard, E. (2016). Closing the loop: 3C versus DNA FISH. Genome Biol *17*, 215-223.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E.*, et al.* (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575-579.

Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C.*, et al.* (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). Human molecular genetics *24*, 3143-3154.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C.*, et al.* (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol *27*, 182-189.

Gomez-Marin, C., Tena, J.J., Acemel, R.D., Lopez-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E.*, et al.* (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci U S A *112*, 7542-7547.

Greenwald, W.W., Li, H., Benaglio, P., Jakubosky, D., Matsui, H., Schmitt, A., Selvaraj, S., D’Antonio, M., D’Antonio-Chronowska, A., Smith, E.N.*, et al.* (2019). Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nature Communications *10*, 1054.

Grewal, S.I., and Elgin, S.C. (2007). Transcription and RNA interference in the formation of heterochromatin. Nature *447*, 399-406.

Gribnau, J., Hochedlinger, K., Hata, K., Li, E., and Jaenisch, R. (2003). Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. Genes Dev *17*, 759-773.

Gu, B., Swigut, T., Spencley, A., Bauer, M.R., Chung, M., Meyer, T., and Wysocka, J. (2018). Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. Science *359*, 1050-1055.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W.*, et al.* (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948-951.

Guillot, P.V., Xie, S.Q., Hollinshead, M., and Pombo, A. (2004). Fixation-induced redistribution of hyperphosphorylated RNA polymerase II in the nucleus of human cells. Exp Cell Res *295*, 460-468.

7. References

Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T., and Wu, Q. (2012). CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. Proceedings of the National Academy of Sciences *109*, 21081.

Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R.*, et al.* (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat Struct Mol Biol *21*, 198-206.

Hakim, O., Sung, M.H., Voss, T.C., Splinter, E., John, S., Sabo, P.J., Thurman, R.E., Stamatoyannopoulos, J.A., de Laat, W., and Hager, G.L. (2011). Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. Genome Res *21*, 697-706.

Hatano, S.-y., Tada, M., Kimura, H., Yamaguchi, S., Kono, T., Nakano, T., Suemori, H., Nakatsuji, N., and Tada, T. (2005). Pluripotential competence of cells associated with Nanog activity. Mechanisms of Development *122*, 67-79.

Hellings, J.A., Arnold, L.E., and Han, J.C. (2017). Dopamine antagonists for treatment resistance in autism spectrum disorders: review and focus on BDNF stimulators loxapine and amitriptyline. Expert opinion on pharmacotherapy *18*, 581-588.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.W., Lyou, Y., Townes, T.M., Schubeler, D., and Gilbert, D.M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol *6*, e245 2220-2236.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A.*, et al.* (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454-1458.

Holwerda, S.J.B., van de Werken, H.J.G., Ribeiro de Almeida, C., Bergen, I.M., de Bruijn, M.J.W., Verstegen, M.J.A.M., Simonis, M., Splinter, E., Wijchers, P.J., Hendriks, R.W.*, et al.* (2013). Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. Nucleic Acids Research *41*, 6905-6916.

Horev, G., Ellegood, J., Lerch, J.P., Son, Y.E., Muthuswamy, L., Vogel, H., Krieger, A.M., Buja, A., Henkelman, R.M., Wigler, M.*, et al.* (2011). Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. Proc Natl Acad Sci U S A *108*, 17076-17081.

Horike, S., Cai, S., Miyano, M., Cheng, J.F., and Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet *37*, 31-40.

Horta, A., Monahan, K., Bashkirova, E., and Lomvardas, S. (2018). Cell type-specific interchromosomal interactions as a mechanism for transcriptional diversity. bioRxiv, 287532.

Hsieh, T.-H.S., Slobodyanyuk, E., Hansen, A.S., Cattoglio, C., Rando, O.J., Tjian, R., and Darzacq, X. (2019). Resolving the 3D landscape of transcription-linked mammalian chromatin folding. bioRxiv, 638775.

Hsieh, T.-Han S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, Oliver J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108-119.

Hsieh, T.S., Fudenberg, G., Goloborodko, A., and Rando, O.J. (2016). Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. Nat Methods *13*, 1009-1011.

Hu, J. (2016). Investigating the role of Smchd1 in control of clustered protocadherin expression (Queensland University of Technology).

7. References

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet *46*, 205-212.

Hyon, C., Chantot-Bastaraud, S., Harbuz, R., Bhouri, R., Perrot, N., Peycelon, M., Sibony, M., Rojo, S., Piguel, X., Bilan, F*., et al.* (2015). Refining the regulatory region upstream of SOX9 associated with 46,XX testicular disorders of Sex Development (DSD). American Journal of Medical Genetics Part A *167*, 1851-1858.

Iborra, F.J., Pombo, A., Jackson, D.A., and Cook, P.R. (1996). Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. Journal of cell science *109 (Pt 6)*, 1427-1436.

Iijima, S., Tanimoto, Y., Mizuno, S., Daitoku, Y., Kunita, S., Sugiyama, F., and Yagami, K. (2010). Effect of different culture conditions on establishment of embryonic stem cells from BALB/cAJ and NZB/BINJ mice. Cellular reprogramming *12*, 679-688.

Jackson, D.A., Bartlett, J., and Cook, P.R. (1996). Sequences attaching loops of nuclear and mitochondrial DNA to underlying structures in human cells: the role of transcription units. Nucleic Acids Res *24*, 1212-1219.

Jackson, D.A., and Pombo, A. (1998). Replicon Clusters Are Stable Units of Chromosome Structure: Evidence That Nuclear Organization Contributes to the Efficient Activation and Propagation of S Phase in Human Cells. The Journal of Cell Biology *140*, 1285-1295.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J*., et al.* (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384 e1319.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I*., et al.* (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262-275.

Jo, J., Xiao, Y., Sun, A.X., Cukuroglu, E., Tran, H.D., Goke, J., Tan, Z.Y., Saw, T.Y., Tan, C.P., Lokman, H*., et al.* (2016). Midbrain-like Organoids from Human Pluripotent Stem Cells Contain Functional Dopaminergic and Neuromelanin-Producing Neurons. Cell Stem Cell *19*, 248-257.

Jorgensen, H.F., Giadrossi, S., Casanova, M., Endoh, M., Koseki, H., Brockdorff, N., and Fisher, A.G. (2006). Stem cells primed for action: polycomb repressive complexes restrain the expression of lineage-specific regulators in embryonic stem cells. Cell cycle (Georgetown, Tex) *5*, 1411-1414.

Juric, I., Yu, M., Abnousi, A., Raviram, R., Fang, R., Zhao, Y., Zhang, Y., Qiu, Y., Yang, Y., Li, Y*., et al.* (2019). MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. PLOS Computational Biology *15*, e1006982.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol *30*, 90-98.

Kaneko, R., Kato, H., Kawamura, Y., Esumi, S., Hirayama, T., Hirabayashi, T., and Yagi, T. (2006). Allelic gene regulation of Pcdh-alpha and Pcdh-gamma clusters involving both monoallelic and biallelic expression in single Purkinje cells. J Biol Chem *281*, 30551-30560.

Karantzali, E., Lekakis, V., Ioannou, M., Hadjimichael, C., Papamatheakis, J., and Kretsovali, A. (2011). Sall1 regulates embryonic stem cell differentiation in association with nanog. J Biol Chem *286*, 1037-1045.

7. References

Kawase, E., Suemori, H., Takahashi, N., Okazaki, K., Hashimoto, K., and Nakatsuji, N. (1994). Strain difference in establishment of mouse embryonic stem (ES) cell lines. The International journal of developmental biology *38*, 385-390.

Kelsey, G., and Bartolomei, M.S. (2012). Imprinted genes ... and the number is? PLoS Genet *8*, e1002601.

Kempfer, R., and Pombo, A. (2019). Methods for mapping 3D chromosome architecture. Nature Reviews Genetics.

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J.M., Ouellette, S.B., Azhir, A., Kumar, N.*, et al.* (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol *19*, 125-136.

Khan, A., and Zhang, X. (2016). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research *44*, D164-D171.

Kim, J.H., Rege, M., Valeri, J., Dunagin, M.C., Metzger, A., Titus, K.R., Gilgenast, T.G., Gong, W., Beagan, J.A., Raj, A.*, et al.* (2019). LADL: light-activated dynamic looping for endogenous gene expression control. Nature Methods *16*, 633–639.

Kim, J.H., Titus, K.R., Gong, W., Beagan, J.A., Cao, Z., and Phillips-Cremins, J.E. (2018). 5C-ID: Increased resolution Chromosome-Conformation-Capture-Carbon-Copy with in situ 3C and double alternating primer design. Methods *142*, 39-46.

Kim, S., Liachko, I., Brickner, D.G., Cook, K., Noble, W.S., Brickner, J.H., Shendure, J., and Dunham, M.J. (2017). The dynamic three-dimensional organization of the diploid yeast genome. Elife *6*, 23623-23645.

Kohlmaier, A., Savarese, F., Lachner, M., Martens, J., Jenuwein, T., and Wutz, A. (2004). A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. PLoS Biol *2*, E171.

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

Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Liu, T., Ye, Z., Li, B., Dixon, J.R.*, et al.* (2017). Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. bioRxiv, 118737.

Kumar, R.A., KaraMohamed, S., Sudi, J., Conrad, D.F., Brune, C., Badner, J.A., Gilliam, T.C., Nowak, N.J., Cook, E.H., Jr., Dobyns, W.B.*, et al.* (2008). Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet *17*, 628-638.

Kumaran, R.I., and Spector, D.L. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. J Cell Biol *180*, 51-65.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2017). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. Mol Cell *65*, 432-446 e435.

Lamble, S., Batty, E., Attar, M., Buck, D., Bowden, R., Lunter, G., Crook, D., El-Fahmawi, B., and Piazza, P. (2013). Improved workflows for high throughput library preparation using the transposome-based Nextera system. BMC Biotechnol *13*, 104-104.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

7. References

Lavitas, M.-L. (2011). Nuclear architecture and genome function in mammalian nuclei (London, UK: Imperial College London).

LeClerc, S., and Easley, D. (2015). Pharmacological therapies for autism spectrum disorder: a review. P T *40*, 389-397.

Lettice, L.A. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics *12*, 1725-1735.

Lettice, L.A., Daniels, S., Sweeney, E., Venkataraman, S., Devenney, P.S., Gautier, P., Morrison, H., Fantes, J., Hill, R.E., and FitzPatrick, D.R. (2011). Enhancer-adoption as a mechanism of human developmental disease. Human mutation *32*, 1492-1499.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Current biology : CB *20*, R754-763.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, T., Jia, L., Cao, Y., Chen, Q., and Li, C. (2018). OCEAN-C: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks. Genome Biol *19*, 54-68.

Li, X., Luo, O.J., Wang, P., Zheng, M., Wang, D., Piecuch, E., Zhu, J.J., Tian, S.Z., Tang, Z., Li, G.*, et al.* (2017a). Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. Nature Protocols *12*, 899–915.

Li, Y., Almassalha, L.M., Chandler, J.E., Zhou, X., Stypula-Cyrus, Y.E., Hujsak, K.A., Roth, E.W., Bleher, R., Subramanian, H., Szleifer, I.*, et al.* (2017b). The effects of chemical fixation on the cellular nanostructure. Exp Cell Res *358*, 253-259.

Li, Y., Rivera, C.M., Ishii, H., Jin, F., Selvaraj, S., Lee, A.Y., Dixon, J.R., and Ren, B. (2014). CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. PLoS One *9*, e114485-114501.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Liu, X., Zhang, Y., Chen, Y., Li, M., Zhou, F., Li, K., Cao, H., Ni, M., Liu, Y., Gu, Z.*, et al.* (2017). In Situ Capture of Chromatin Interactions by Biotinylated dCas9. Cell *170*, 1028-1043 e1019.

Liu, Z., Legant, W.R., Chen, B.C., Li, L., Grimm, J.B., Lavis, L.D., Betzig, E., and Tjian, R. (2014). 3D imaging of Sox2 enhancer clusters in embryonic stem cells. Elife *3*, e04236.

Llères, D., Moindrot, B., Pathak, R., Piras, V., Matelot, M., Pignard, B., Marchand, A., Poncelet, M., Perrin, A., Tellier, V.*, et al.* (2019). CTCF controls imprinted gene activity at the mouse &lt;em&gt;Dlk1-Dio3&lt;/em&gt; and &lt;em&gt;Igf2-H19&lt;/em&gt; domains by modulating allele-specific sub-TAD structure. bioRxiv, 633065.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. Cell *126*, 403-413.

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

Loviglio, M.N., Leleu, M., Männik, K., Passeggeri, M., Giannuzzi, G., van der Werf, I., Waszak, S.M., Zazhytska, M., Roberts-Caldeira, I., Gheldof, N.*, et al.* (2016). Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Molecular Psychiatry *22*, 836–849.

Lu, L., Liu, X., Peng, J., Li, Y., and Jin, F. (2018). Easy Hi-C: A simple efficient protocol for 3D genome mapping in small cell populations. bioRxiv.

Lucas, J.S., Zhang, Y., Dudko, O.K., and Murre, C. (2014). 3D trajectories adopted by coding and regulatory DNA elements: first-passage times for genomic interactions. Cell *158*, 339-352.

Luperchio, T.R., Sauria, M.E.G., Wong, X., Gaillard, M.-C., Tsang, P., Pekrun, K., Ach, R.A., Yamada, N.A., Taylor, J., and Reddy, K. (2017). Chromosome Conformation Paints Reveal The Role Of Lamina Association In Genome Organization And Regulation. bioRxiv.

Lupianez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R.*, et al.* (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012-1025.

Ma, H., Reyes-Gutierrez, P., and Pederson, T. (2013). Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. Proc Natl Acad Sci U S A *110*, 21048-21053.

Ma, H., Tu, L.C., Naseri, A., Huisman, M., Zhang, S., Grunwald, D., and Pederson, T. (2016). Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. Nat Biotechnol *34*, 528-530.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A.*, et al.* (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Methods *12*, 71-78.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A.*, et al.* (2018a). Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. Methods *142*, 59-73.

Ma, X., Ezer, D., Adryan, B., and Stevens, T.J. (2018b). Canonical and single-cell Hi-C reveal distinct chromatin interaction sub-networks of mammalian transcription factors. Genome Biology *19*, 174.

Maass, P.G., Barutcu, A.R., Weiner, C.L., and Rinn, J.L. (2018a). Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. Mol Cell *69*, 1039-1045.e1033.

Maass, P.G., Weise, A., Rittscher, K., Lichtenwald, J., Barutcu, A.R., Liehr, T., Aydin, A., Wefeld-Neuenfeld, Y., Polsler, L., Tinschert, S.*, et al.* (2018b). Reorganization of inter-chromosomal interactions in the 2q37-deletion syndrome. Embo j *37*, e96257-96270.

Mahamid, J., Pfeffer, S., Schaffer, M., Villa, E., Danev, R., Cuellar, L.K., Forster, F., Hyman, A.A., Plitzko, J.M., and Baumeister, W. (2016). Visualizing the molecular sociology at the HeLa cell nuclear periphery. Science *351*, 969-972.

Maharana, S., Iyer, K.V., Jain, N., Nagarajan, M., Wang, Y., and Shivashankar, G.V. (2016). Chromosome intermingling-the physical basis of chromosome organization in differentiated cells. Nucleic Acids Res *44*, 5148-5160.

7. References

Mak, W., Nesterova, T.B., de Napoles, M., Appanah, R., Yamanaka, S., Otte, A.P., and Brockdorff, N. (2004). Reactivation of the paternal X chromosome in early mouse embryos. Science *303*, 666-669.

Mancini-Dinardo, D., Steele, S.J., Levorse, J.M., Ingram, R.S., and Tilghman, S.M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev *20*, 1268-1282.

Mao, Y.S., Zhang, B., and Spector, D.L. (2011). Biogenesis and function of nuclear bodies. Trends Genet *27*, 295-306.

Markaki, Y., Smeets, D., Fiedler, S., Schmid, V.J., Schermelleh, L., Cremer, T., and Cremer, M. (2012). The potential of 3D-FISH and super-resolution structured illumination microscopy for studies of 3D nuclear architecture: 3D structured illumination microscopy of defined chromosomal structures visualized by 3D (immuno)-FISH opens new perspectives for studies of nuclear architecture. Bioessays *34*, 412-426.

Marshall, O.J., Southall, T.D., Cheetham, S.W., and Brand, A.H. (2016). Cell-type-specific profiling of protein-DNA interactions without cell isolation using targeted DamID with next-generation sequencing. Nat Protoc *11*, 1586-1598.

Mayer, R., Brero, A., von Hase, J., Schroeder, T., Cremer, T., and Dietzel, S. (2005). Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. BMC Cell Biol *6*, 44-65.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A.*, et al.* (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet *47*, 598-606.

Monahan, K., Horta, A., and Lomvardas, S. (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature *565*, 448-453.

Monneron, A., and Bernhard, W. (1969). Fine structural organization of the interphase nucleus in some mammalian cells. Journal of ultrastructure research *27*, 266-288.

Müller, S., Neusser, M., and Wienberg, J. (2002). Towards unlimited colors for fluorescence in-situhybridization (FISH). Chromosome Research *10*, 223-232.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods *13*, 919-922.

Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R.*, et al.* (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics *49*, 1602-1612.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59-64.

Nagano, T., Lubling, Y., Varnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., and Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature *547*, 61-67.

Nagano, T., Lubling, Y., Yaffe, E., Wingett, S.W., Dean, W., Tanay, A., and Fraser, P. (2015). Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. Nat Protoc *10*, 1986-2003.

Nakamura, K., Sekine, Y., Ouchi, Y., Tsujii, M., Yoshikawa, E., Futatsubashi, M., Tsuchiya, K.J., Sugihara, G., Iwata, Y., Suzuki, K*., et al.* (2010). Brain serotonin and dopamine transporter bindings in adults with high-functioning autism. Archives of general psychiatry *67*, 59-68.

Nativio, R., Sparago, A., Ito, Y., Weksberg, R., Riccio, A., and Murrell, A. (2011). Disruption of genomic neighbourhood at the imprinted IGF2-H19 locus in Beckwith-Wiedemann syndrome and Silver-Russell syndrome. Human molecular genetics *20*, 1363-1374.

Naumova, N., Smith, E.M., Zhan, Y., and Dekker, J. (2012). Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods (San Diego, Calif) *58*, 192-203.

Nemeth, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Peterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Langst, G. (2010). Initial genomics of the human nucleolus. PLoS Genet *6*, e1000889-1000899.

Nguyen, M., Roth, A., Kyzar, E.J., Poudel, M.K., Wong, K., Stewart, A.M., and Kalueff, A.V. (2014). Decoding the contribution of dopaminergic genes and pathways to autism spectrum disorder (ASD). Neurochemistry international *66*, 15-26.

Ni, Y., Cao, B., Ma, T., Niu, G., Huo, Y., Huang, J., Chen, D., Liu, Y., Yu, B., Zhang, M.Q*., et al.* (2017). Super-resolution imaging of a 2.5 kb non-repetitive DNA in situ in the nuclear genome using molecular beacon probes. Elife *6*, 21660-21622.

Niarchou, M., Chawner, S., Doherty, J.L., Maillard, A.M., Jacquemont, S., Chung, W.K., Green-Snyder, L., Bernier, R.A., Goin-Kochel, R.P., Hanson, E*., et al.* (2019). Psychiatric disorders in children with 16p11.2 deletion and duplication. Transl Psychiatry *9*, 8.

Nir, G., Farabella, I., Pérez Estrada, C., Ebeling, C.G., Beliveau, B.J., Sasaki, H.M., Lee, S.D., Nguyen, S.C., McCole, R.B., Chattoraj, S*., et al.* (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS genetics *14*, e1007872-e1007872.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413-413.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J*., et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.

Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., Pawson, T., Forman-Kay, J.D*., et al.* (2015). Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol Cell *57*, 936-947.

O'Sullivan, J.M., Hendy, M.D., Pichugina, T., Wake, G.C., and Langowski, J. (2013). The statistical-mechanics of chromosome conformation capture. Nucleus *4*, 390-398.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296-300.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics *12*, 283-293.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W*., et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet *36*, 1065-1071.

Osborne, C.S., and Eskiw, C.H. (2008). Where shall we meet? A role for genome organisation and nuclear sub-compartments in mediating interchromosomal interactions. J Cell Biochem *104*, 1553-1561.

Oudelaar, A.M., Davies, J.O.J., Downes, D.J., Higgs, D.R., and Hughes, J.R. (2017). Robust detection of chromosomal interactions from small numbers of cells using low-input Capture-C. Nucleic Acids Research *45*, e184-e192.

Oudelaar, A.M., Davies, J.O.J., Hanssen, L.L.P., Telenius, J.M., Schwessinger, R., Liu, Y., Brown, J.M., Downes, D.J., Chiariello, A.M., Bianco, S.*, et al.* (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. Nat Genet *50*, 1744-1751.

Parada, L., and Misteli, T. (2002). Chromosome positioning in the interphase nucleus. Trends Cell Biol *12*, 425-432.

Pederson, T. (2011). The nucleolus. Cold Spring Harb Perspect Biol *3*, a000638-000653.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M.*, et al.* (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell *38*, 603-613.

Pombo, A., and Branco, M.R. (2007). Functional organisation of the genome during interphase. Curr Opin Genet Dev *17*, 451-455.

Pombo, A., Jackson, D.A., Hollinshead, M., Wang, Z., Roeder, R.G., and Cook, P.R. (1999). Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. The EMBO Journal *18*, 2241-2253.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K.*, et al.* (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402-405.

Qin, Y., Kong, L.K., Poirier, C., Truong, C., Overbeek, P.A., and Bishop, C.E. (2004). Long-range activation of Sox9 in Odd Sex (Ods) mice. Hum Mol Genet *13*, 1213-1218.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y.*, et al.* (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744-757.e724.

Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nature methods *14*, 263-266.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D.*, et al.* (2017). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305-320.e324.

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

Redolfi, J., Zhan, Y., Valdes-Quezada, C., Kryzhanovska, M., Guerreiro, I., Iesmantavicius, V., Pollex, T., Grand, R.S., Mulugeta, E., Kind, J*., et al.* (2019). DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. Nature Structural & Molecular Biology *26*, 471-480.

Reik, W., and Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nat Rev Genet *2*, 21-32.

Reinius, B., and Sandberg, R. (2015). Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nature Reviews Genetics *16*, 653-664.

Ribich, S., Tasic, B., and Maniatis, T. (2006). Identification of long-range regulatory elements in the protocadherin-α gene cluster. Proceedings of the National Academy of Sciences *103*, 19719-19724.

Rivera-Mulia, J.C., Dimond, A., Vera, D., Trevilla-Garcia, C., Sasaki, T., Zimmerman, J., Dupont, C., Gribnau, J., Fraser, P., and Gilbert, D.M. (2018). Allele-specific control of replication timing and genome organization during development. Genome Res *28*, 800-811.

Robinett, C.C., Straight, A., Li, G., Willhelm, C., Sudlow, G., Murray, A., and Belmont, A.S. (1996). In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. J Cell Biol *135*, 1685-1700.

Rodley, C.D.M., Bertels, F., Jones, B., and O'Sullivan, J.M. (2009). Global identification of yeast chromosome interactions using Genome conformation capture. Fungal Genetics and Biology *46*, 879-886.

Rosin, L.F., Nguyen, S.C., and Joyce, E.F. (2018). Condensin II drives large-scale folding and spatial partitioning of interphase chromosomes in Drosophila nuclei. PLOS Genetics *14*, e1007393-1007418.

Rubinstein, R., Thu, Chan A., Goodman, Kerry M., Wolcott, Holly N., Bahna, F., Mannepalli, S., Ahlsen, G., Chevee, M., Halim, A., Clausen, H*., et al.* (2015). Molecular Logic of Neuronal Self-Recognition through Protocadherin Domain Interactions. Cell *163*, 629-642.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C*., et al.* (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*, aar3958-3967.

Sabbattini, P., Georgiou, A., Sinclair, C., and Dillon, N. (1999). Analysis of mice with single and multiple copies of transgenes reveals a novel arrangement for the lambda5-VpreB1 locus control region. Mol Cell Biol *19*, 671-679.

Sadoni, N., Langer, S., Fauth, C., Bernardi, G., Cremer, T., Turner, B.M., and Zink, D. (1999). Nuclear Organization of Mammalian Genomes. The Journal of Cell Biology *146*, 1211-1226.

Sambrook, J.F., and Russell, D. (2001). Molecular Cloning: A Laboratory Manual, Vol 1.

Schmitz, R., Ceribelli, M., Pittaluga, S., Wright, G., and Staudt, L.M. (2014). Oncogenic mechanisms in Burkitt lymphoma. Cold Spring Harbor perspectives in medicine *4*, a014282-014294.

Schoenfelder, S., Clay, I., and Fraser, P. (2010a). The transcriptional interactome: gene expression in 3D. Current Opinion in Genetics & Development *20*, 127-133.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W*., et al.* (2015a). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res *25*, 582-597.

Schoenfelder, S., Javierre, B.M., Furlan-Magaril, M., Wingett, S.W., and Fraser, P. (2018). Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. J Vis Exp, e57320-57336.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S*., et al.* (2010b). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet *42*, 53-61.

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M*., et al.* (2015b). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat Genet *47*, 1179-1186.

Schwaiger, M., Stadler, M.B., Bell, O., Kohler, H., Oakeley, E.J., and Schubeler, D. (2009). Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. Genes Dev *23*, 589-601.

Schwartzman, O., Mukamel, Z., Oded-Elkayam, N., Olivares-Chauvet, P., Lubling, Y., Landan, G., Izraeli, S., and Tanay, A. (2016). UMI-4C for quantitative and targeted chromosomal contact profiling. Nat Methods *13*, 685-691.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G.J., and Marti-Renom, M.A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLOS Computational Biology *13*, e1005665.

Servant, N., Varoquaux, N., Heard, E., Barillot, E., and Vert, J.-P. (2018). Effective normalization for copy number variation in Hi-C data. BMC bioinformatics *19*, 313-313.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458-472.

Shao, S., Zhang, W., Hu, H., Xue, B., Qin, J., Sun, C., Sun, Y., Wei, W., and Sun, Y. (2016). Long-term dual-color tracking of genomic loci by modified sgRNAs of the CRISPR/Cas9 system. Nucleic Acids Res *44*, e86-98.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V*., et al.* (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.

Shimizu, N., Maekawa, M., Asai, S., and Shimizu, Y. (2015). Multicolor FISHs for simultaneous detection of genes and DNA segments on human chromosomes. Chromosome Research *23*, 649-662.

Shopland, L.S., Johnson, C.V., Byron, M., McNeil, J., and Lawrence, J.B. (2003). Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: evidence for local euchromatic neighborhoods. J Cell Biol *162*, 981-990.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics *38*, 1348-1354.

Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S., and Cremer, T. (2002). Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Exp Cell Res *276*, 10-23.

Song, J., Ugai, H., Nakata-Tsutsui, H., Kishikawa, S., Suzuki, E., Murata, T., and Yokoyama, K.K. (2003). Transcriptional regulation by zinc-finger proteins Sp1 and MAZ involves interactions with the same cis-elements. International journal of molecular medicine *11*, 547-553.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb Perspect Biol *3*, a000646-000657.

Speicher, M.R., Gwyn Ballard, S., and Ward, D.C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat Genet *12*, 368-375.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nature Reviews Genetics *19*, 453-467.

Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., and Flavell, R.A. (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637-645.

Staal, W.G., de Krom, M., and de Jonge, M.V. (2012). Brief report: the dopamine-3-receptor gene (DRD3) is associated with specific repetitive behavior in autism spectrum disorder (ASD). J Autism Dev Disord *42*, 885-888.

Steinman, K.J., Spence, S.J., Ramocki, M.B., Proud, M.B., Kessler, S.K., Marco, E.J., Green Snyder, L., D'Angelo, D., Chen, Q., Chung, W.K*., et al.* (2016). 16p11.2 deletion and duplication: Characterizing neurologic phenotypes in a large clinically ascertained cohort. American journal of medical genetics Part A *170*, 2943-2955.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A*., et al.* (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59-64.

Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241-245.

Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., and Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. Dev Cell *39*, 529-543.

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res *24*, 390-400.

Takizawa, T., Gudla, P.R., Guo, L., Lockett, S., and Misteli, T. (2008). Allele-specific nuclear positioning of the monoallelically expressed astrocyte marker GFAP. Genes & development *22*, 489-498.

Tan, L., Xing, D., Chang, C.-H., Li, H., and Xie, X.S. (2018a). Three-dimensional genome structures of single diploid human cells. Science *361*, 924-928.

Tan, L., Xing, D., Chang, C.H., Li, H., and Xie, X.S. (2018b). Three-dimensional genome structures of single diploid human cells. Science *361*, 924-928.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B*., et al.* (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1611-1627.

Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., and Maniatis, T. (2002). Promoter Choice Determines Splice Site Selection in Protocadherin α and γ Pre-mRNA Splicing. Molecular Cell *10*, 21-33.

Terranova, R., Yokobayashi, S., Stadler, M.B., Otte, A.P., van Lohuizen, M., Orkin, S.H., and Peters, A.H. (2008). Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. Dev Cell *15*, 668-679.

Thu, C.A., Chen, W.V., Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between alpha, beta, and gamma protocadherins. Cell *158*, 1045-1059.

Tiwari, V.K., Cope, L., McGarvey, K.M., Ohm, J.E., and Baylin, S.B. (2008). A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. Genome Res *18*, 1171-1179.

Tokuyasu, K.T. (1973). A technique for ultracryotomy of cell suspensions and tissues. J Cell Biol *57*, 551-565.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell *10*, 1453-1465.

Tomizawa, S.-i., Kobayashi, H., Watanabe, T., Andrews, S., Hata, K., Kelsey, G., and Sasaki, H. (2011). Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. Development (Cambridge, England) *138*, 811-820.

Toyoda, S., Kawaguchi, M., Kobayashi, T., Tarusawa, E., Toyama, T., Okano, M., Oda, M., Nakauchi, H., Yoshimura, Y., Sanbo, M.*, et al.* (2014). Developmental Epigenetic Modification Regulates Stochastic Expression of Clustered Protocadherin Genes, Generating Single Neuron Diversity. Neuron *82*, 94-108.

van de Werken, H.J., de Vree, P.J., Splinter, E., Holwerda, S.J., Klous, P., de Wit, E., and de Laat, W. (2012a). 4C technology: protocols and data analysis. Methods in enzymology *513*, 89-112.

van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A.*, et al.* (2012b). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods *9*, 969-972.

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

Verschure, P.J., van der Kraan, I., Manders, E.M.M., and van Driel, R. (1999). Spatial Relationship between Transcription Sites and Chromosome Territories. The Journal of Cell Biology *147*, 13-24.

Vidal, E., le Dily, F., Quilez, J., Stadhouders, R., Cuartero, Y., Graf, T., Marti-Renom, M.A., Beato, M., and Filion, G.J. (2018). OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. Nucleic Acids Res *46*, e49-58.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep *10*, 1297-1309.

Visser, A.E., Eils, R., Jauch, A., Little, G., Bakker, P.J., Cremer, T., and Aten, J.A. (1998). Spatial distributions of early and late replicating chromatin in interphase chromosome territories. Exp Cell Res *243*, 398-407.

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

7. References

Wang, H., Xu, X., Nguyen, C.M., Liu, Y., Gao, Y., Lin, X., Daley, T., Kipniss, N.H., La Russa, M., and Qi, L.S. (2018). CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. Cell *175*, 1405-1417.e1414.

Wang, H., Zang, C., Taing, L., Arnett, K.L., Wong, Y.J., Pear, W.S., Blacklow, S.C., Liu, X.S., and Aster, J.C. (2014). NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. Proc Natl Acad Sci U S A *111*, 705-710.

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature *444*, 364-368.

Wang, S., Su, J.H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C.T., and Zhuang, X. (2016a). Spatial organization of chromatin domains and compartments in single chromosomes. Science *353*, 598-602.

Wang, S., Su, J.H., Zhang, F., and Zhuang, X. (2016b). An RNA-aptamer-based two-color CRISPR labeling system. Sci Rep *6*, 26857-26863.

Wang, X.T., Dong, P.F., Zhang, H.Y., and Peng, C. (2015). Structural heterogeneity and functional diversity of topologically associating domains in mammalian genomes. Nucleic Acids Res *43*, 7237-7246.

Watford, S.M., Grashow, R.G., De La Rosa, V.Y., Rudel, R.A., Friedman, K.P., and Martin, M.T. (2018). Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: Use case in breast carcinogenesis. Computational Toxicology *7*, 46-57.

Watson, C.M., Crinnion, L.A., Harrison, S.M., Lascelles, C., Antanaviciute, A., Carr, I.M., Bonthron, D.T., and Sheridan, E. (2016). A Chromosome 7 Pericentric Inversion Defined at Single-Nucleotide Resolution Using Diagnostic Whole Genome Sequencing in a Patient with Hand-Foot-Genital Syndrome. PLoS One *11*, e0157075.

Weaver, J.R., and Bartolomei, M.S. (2014). Chromatin regulators of genomic imprinting. Biochimica et biophysica acta *1839*, 169-177.

Wei, Z., Gao, F., Kim, S., Yang, H., Lyu, J., An, W., Wang, K., and Lu, W. (2013). Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. Cell Stem Cell *13*, 36-47.

Weinreb, C., and Raphael, B.J. (2016). Identification of hierarchical chromatin domains. Bioinformatics *32*, 1601-1609.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Wu, Q., and Maniatis, T. (1999). A Striking Organization of a Large Family of Human Neural Cadherin-like Cell Adhesion Genes. Cell *97*, 779-790.

Xie, S.Q., Lavitas, L.M., and Pombo, A. (2010). CryoFISH: fluorescence in situ hybridization on ultrathin cryosections. Methods Mol Biol *659*, 219-230.

Xie, S.Q., Martin, S., Guillot, P.V., Bentley, D.L., and Pombo, A. (2006). Splicing speckles are not reservoirs of RNA polymerase II, but contain an inactive form, phosphorylated on serine2 residues of the C-terminal domain. Mol Biol Cell *17*, 1723-1733.

Xing, H., Mo, Y., Liao, W., and Zhang, M.Q. (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. PLoS computational biology *8*, e1002613-e1002613.

Yang, T., Zhang, F., Yardımcı, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., and Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome research *27*, 1939-1949.

Ying, Q.L., Stavridis, M., Griffiths, D., Li, M., and Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat Biotechnol *21*, 183-186.

Yokota, S., Hirayama, T., Hirano, K., Kaneko, R., Toyoda, S., Kawamura, Y., Hirabayashi, M., Hirabayashi, T., and Yagi, T. (2011). Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster. J Biol Chem *286*, 31885-31895.

Young, Richard A. (2011). Control of the Embryonic Stem Cell State. Cell *144*, 940-954.

Zhang, X., Zhang, Y., Zhu, X., Purmann, C., Haney, M.S., Ward, T., Khechaduri, A., Yao, J., Weissman, S.M., and Urban, A.E. (2018a). Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. Nature Communications *9*, 5356-5370.

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W.J., Hu, M., Tang, J., and Yue, F. (2018b). Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat Commun *9*, 750-758.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908-921.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U.*, et al.* (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics *38*, 1341-1347.

Zheng, M., Tian, S.Z., Capurso, D., Kim, M., Maurya, R., Lee, B., Piecuch, E., Gong, L., Zhu, J.J., Li, Z.*, et al.* (2019). Multiplex chromatin interactions with single-molecule precision. Nature *566*, 558-562.

Zink, D., Bornfleth, H., Visser, A., Cremer, C., and Cremer, T. (1999). Organization of early and late replicating DNA in human chromosome territories. Exp Cell Res *247*, 176-188.

Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science *338*, 1622-1626.

Zsikla, V., Baumann, M., and Cathomas, G. (2004). Effect of buffered formalin on amplification of DNA from paraffin wax embedded small biopsies using real-time PCR. J Clin Pathol *57*, 654-656.

Zullo, J.M., Demarco, I.A., Pique-Regi, R., Gaffney, D.J., Epstein, C.B., Spooner, C.J., Luperchio, T.R., Bernstein, B.E., Pritchard, J.K., Reddy, K.L.*, et al.* (2012). DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell *149*, 1474-1487.

8. Appendix

Table 8. 1: Parameters tested in GAM to enable high throughput collection of nuclear profiles

Ease of protocol

WGA, cell	Target of optimisation	State of protocol before test	New implementation	Conclusions	Recommendation for GAM users
IH, 2	Storage of cryosections	No pause between cryosectioning and LMD	Store cryosections at -20°C before LMD	No difference in NP quality	Implement as optional save stopping point
S, 1	Storage of nuclear profiles	Freezing nuclear profiles during WGA	Freezing nuclear profiles after LMD	Improved NP quality	Implement in protocol
S, 1	Storage of nuclear profiles	No stopping points in WGA	Freezing samples during WGA protocol	No difference in NP quality	Implement as optional save stopping point
S, 1	Pooling of sequencing libraries	DNA quantification with Qubit	DNA quantification with PicoGreen	No difference in NP quality	Implement in protocol
S, 1	Sequencing machine	HiSeq 2000 for sequencing	Nextseq 500 for sequencing	No difference in NP quality	Implement in protocol

High-throughput, reducing costs per NP

WGA	Target of optimisation	State of protocol before test	New implementation	Conclusions	Recommendation for GAM users
IH, 2	Caps for collection of NPs (LMD)	8-well strip opaque filled LMD caps	8-well strip PCR caps	Slightly reduced NP quality	Not recommended
IH, 2	Volumes of Nextera library reagents	Volumes according to manufacturer	Scale down to 1/5 volume of all reagents	Identical performance	Implement in protocol
S, 1	DNA purification	AMPure XP beads	In-house SPRI beads	No difference in NP quality	Implement in protocol

High-throughput, reducing time per NP

WGA	Target of optimisation	State of protocol before test	New implementation	Conclusions	Recommendation for GAM users
S, IH, M, 1	Automation of WGA	Manual pipetting	Automated pipetting (with larger volumes)	No difference in NP quality	Implement in protocol
IH, 2	Library preparation protocol	Manual pipetting	Automated pipetting	Identical performance	Implement in protocol
S, 1	DNA purification	Qiagen MinElute columns	Qiagen MinElute 96 UF Plate	No difference in NP quality	Implement in protocol
S, 1	DNA purification	Qiagen MinElute 96 UF Plate	Ampure XP beads purification 0.8x	No difference in NP quality	Implement in protocol
S, 1	Library preparation	TruSeq DNA library preparation	Nextera XT DNA library preparation	No difference in NP quality	Implement in protocol
S, 1	Pooling of sequencing libraries	Manual normalisation	Beads-based normalisation	Unbalanced distribution of sequencing reads per library	Not recommended

46C = 1; F123 = 2; Sigma = S; in-house = IH; Malbac = M; other = O

Table 8. 2: Parameters for optimising DNA extractions from NPs

Efficiency of DNA extraction					
WGA	Target of optimisation	State of protocol before test	New implementation	Conclusions	Recommendation for GAM users
IH, 2	Washes on cryosections	30 min at RT in 1x PBS	Overnight at 4°C in 1x PBS	Slightly improved NP quality	Implement as optional save stopping point
IH, 2	Washes on cryosections	30 min at RT in 1x PBS	Add wash: 10 min 0.5 % Triton X-100	Human contamination in the test, inconclusive	Repeat test
S, 1	Orientation of LMD slide	Sample facing up in LMD	Sample facing down in LMD	No difference in NP quality	No recommendation
S, IH, 2	Orientation of section on LMD slide	Membrane side (sample below)	Metal frame side (sample on top)	No difference in NP quality	Chose side depending on cell type, some cells do not stick to both sides
S, 1, 2	LMD membrane slide	PEN membrane	PEN membrane on glass	No difference in NP quality, but less practical	No recommendation
S, IH, 1, 2	LMD membrane slide	PEN membrane	PPS membrane	Decreased NP quality	If membrane is needed, optimise temperature of incubation
S, 1	LMD membrane slide	PEN membrane	Pol membrane	Decreased NP quality	Not recommended
S, 1	Statics during LMD collection	LMD microscope without airfilter	Humidity chamber around LMD microscope	Improved NP quality	Implement, reduced statics make collection of NPs more efficient
S, 1	Visualize NPs	No staining	DAPI staining	NPs not visible	Not suitable for GAM
S, 1	Visualize NPs	No staining	HOECHST staining	NPs not visible	Not suitable for GAM
S, 1	Visualize NPs	No staining	Toto-3 staining	NPs not visible	Not suitable for GAM
S, 1	Visualize NPs	No staining	EvaGreen staining	NPs not visible	Not suitable for GAM
S, 1	Visualize NPs	No staining	Propidium iodine (PI) staining	Decreased NP quality	Dye interferes with WGA, not suitable for GAM
S, 1	Visualize NPs	No staining	SYBR Gold staining	Decreased NP quality	Dye interferes with WGA, not suitable for GAM
S, 1	Visualize NPs	No staining	SYTO RNA Select staining	Decreased NP quality	Dye interferes with WGA, not suitable for GAM
S, 1	Visualize NPs	No staining	Eosine Y staining	Decreased NP quality	Dye interferes with WGA, not suitable for GAM
S, 1	Visualize NPs	No staining	Crysel violet staining	Decreased NP quality	Dye interferes with WGA, not suitable for GAM
S, 1	Visualize NPs	No staining	Crystal violet staining	Decreased NP quality	Dye interferes with WGA, not suitable for GAM
S, 1	Visualize NPs	No staining	Cresyl violet staining	Cells visible, not nuclei, no difference in NP quality	Implement in protocol
S, IH, 1, 2	Visualize NPs	No staining, cresyl violet	Immunofluorescence staining	Nuclei visible, improved NP quality	Implement in protocol
S, 1	DNA contamination	DNA-free working rooms without air filter	Air filter in DNA-free working rooms	Improved NP quality, less contamination	Implement in protocol
S, 1	WGA protocol	WGA4 kit, size: 50x	WGA4 kit, size: 500 x	Decreased NP quality	Not recommended

46C = 1; F123 = 2; Sigma = S; in-house = IH; Malbac = M; other = O

S, 1	Cell lysis (WGA)	4 hours incubation	Overnight incubation	No difference in NP quality	Consider retesting, might not apply to every cell type
S, 1	Cell lysis (WGA)	1x Proteinase K	2x, 3x, 4x, 8x Proteinase K	Improved NP quality at 2x, 3x	Equal quality for 2x and 3x, use one of these concentrations
S, 1	Cell lysis (WGA)	Proteinase K	Streptomyces protease	Decreased NP quality	Not suitable for DNA extraction from fixed cryosections
S, 1	Cell lysis (WGA)	Proteinase K	Pancreas protease	Decreased NP quality	Not suitable for DNA extraction from fixed cryosections
S, 1	Number of PCR cycles in WGA	25 cycles	23 cycles, 20 cycles	Increased number of PCR cycles	Not recommended
O, 1	WGA protocol	Sigma WGA4 kit	Ampli-1 kit	Decreased NP quality, atypical sequencing tracks	Not suitable for GAM
O, 1	WGA protocol	Sigma WGA4 kit	Repli-G kit	Decreased NP quality, atypical sequencing tracks	Not suitable for GAM
M, 1	WGA protocol	Sigma WGA4 kit	Yikon Malbac kit	Good NP quality, typical sequencing tracks	Kit can be used for GAM
M, 1	Cell lysis (WGA)	50 min incubation	2 hours, 4 hours incubation	Improved NP quality at 4 hours	Start with 4h, consider testing longer incubation times
M, 1	Cell lysis (WGA)	1x protease	2x protease	No difference in NP quality	Lower sample number than usual, consider testing again
IH, 2	Volume of WGA reagents	Volumes according to initial protocol	Increase volumes of all reagents 2x	Slightly reduced NP quality	Not recommended
IH, 2	Cell lysis (WGA)	4 hours incubation	Overnight incubation	No difference in NP quality	No recommendation
IH, 2	Cell lysis (WGA)	55 C incubation temperature	50, 60, 65 C incubation temperature	Temperature influences genome coverage	Optimal at 60 C
IH, 2	Cell lysis (WGA)	1x protease	2x protease, 3x protease	Improved NP quality	Equal quality for 2x and 3x, use one of these concentrations
IH, 2	DNA amplification (WGA)	WGA with 3 min MDA and 4 min elongation	WGA with 5 min MDA and 5 min elongation	Slightly improved NP quality	Consider implementation, repeat test with larger sample numbers
IH, 2	WGA protocol	no MALBAC looping step	MALBAC looping step	High variability of NP quality	Repeat test with higher sample numbers
IH, 2	dNTP and Mg concentration in WGA	0.2 nM dNTP, 2mM Mg	0.4 mM dNTP, 4 mM Mg	Decreased NP quality	Not recommended
IH, 2	PCR primer annealing in WGA	annealing temperature 58 C	annealing temperature 65 C	Decreased NP quality	Not recommended
IH, 2	Number of PCR cycles in WGA	26 cycles	24 cycles, 22 cycles	higher variability in DNA yield after WGA	DNA yield below input recommendation for library prep, do not implement
S, 1	DNA purification	Ampure XP beads 0.8x	Ampure XP beads 1.7x	No difference in NP quality	Implement in protocol
S, IH, 1, 2	Number of PCR cycles in library prep	12 cycles	11 cycles	Reduced PCR duplicates, but does not work for small volume prep	Implement only for full volume library prep, otherwise not recommended
IH, 2	Sequencing	Single end s (1x 75 bp)	Paired end s (2x 75 bp)	Reads cover identical regions, no change in positive windows	No recommendation

46C = 1; F123 = 2; Sigma = S; in-house = IH; Malbac = M; other = O

8. Appendix