From THE DEPARTMENT OF MICROBIOLOGY, TUMOR AND CELL BIOLOGY
Karolinska Institutet, Stockholm, Sweden

BIOINFORMATIC ANALYSES OF THE STRUCTURAL AND FUNCTIONAL COMPLEXITY IN CHROMOSOMAL INTERACTOMES

Alejandro Fernández Woodbridge

Stockholm 2015

Published by Karolinska Institutet.
Printed by **AJ E-print AB**

ISBN 978-91-7676-083-3

Bioinformatic analyses of the structural and functional complexity in chromosomal interactomes
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Alejandro Fernández Woodbridge

Principal Supervisor:
Anita Göndör
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Co-supervisor(s):
Rolf Ohlsson
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Erik Aurell
Royal Institute of Technology
Department of School of Computer Science and Communication

Opponent:
Albin Sandelin
Copenhagen University
Department of Biology and BRIC

Examination Board:
Ann-Kristin Östlund Farrants
Stockholm University
Department of Molecular Biosciences

Jan Komorowski
Uppsala Universitet
Department of Cell and Molecular Biology

Sten Linnarsson
Karolinska Institutet
Department of Medical Biochemistry and Biophysics

TO THOSE THAT I HAVE LOST, AND TO THOSE THAT FOUGHT AND WON FOR REMINDING ME WHY I STAND HERE TODAY

"THERE ARE MEN THAT FIGHT FOR A DAY, AND THEY ARE GOOD. THERE ARE OTHERS THAT FIGHT FOR A YEAR AND THEY ARE BETTER. THERE ARE MEN THAT FIGHT MANY YEARS, AND THEY ARE VERY GOOD. BUT THERE ARE MEN THAT FIGHT THEIR ENTIRE LIVES.... THOSE, THOSE ARE THE ONES THE WORLD CANNOT DO WITHOUT." BERTOLT BRECHT

ABSTRACT

Evolution requires information storage systems with different demands with respect to persistence. While the genome provides a mechanism for long term, static and accurate information storage, it is incapable of mediating adaptation to short term changes in the environment. Chromatin, however, constitutes a dynamic, reprogrammable memory with different levels of persistence. Moreover, chromatin states carry information not only in 2D, i.e. in the structure of the primary chromatin fibre, but also in the 3D organization of the genome in the nuclear space. The following thesis delves into the new bioinformatic and wet lab protocols developed to map, quantitative and functionally analyze the 3D architecture of chromatin.

The chromatin insulator protein CTCF is a major factor underlying the 3D organization of the epigenome. We have uncovered, however, that CTCF binding sites within a regulatory region have multiple functions that are influenced by the chromatin environment and possibly the combinatorial usage of the 11 Zn-fingers of CTCF (Paper I). This observation exemplifies that understanding the function of dynamic and transient chromatin fibre interactions requires novel technology that enables the detection of 3D chromatin folding with high resolution in single cells and in small cell populations. We therefore set out to devise a novel method for the visualization of higher order chromatin structures by combining the strengths of both DNA Fluorescent *In Situ* Hybridization (FISH) and *In Situ* Proximity Ligation Assay (ISPLA) technologies (Paper II). The resulting Chromatin *in Situ* Proximity (ChrISP) assay thus takes advantage of the direct contact detection of ISPLA and the locus-specific nature of FISH and uncovered the existence of compact chromatin structures at the nuclear envelope with unprecedented resolution. To complement ChrISP with a high throughput method capable of quantitatively recovering chromatin fibre contacts in small cell populations, we furthermore innovated the Nodewalk assay (Paper III). The protocol builds on existing ligation based chromosome conformation capture methods, but features significant reduction in the random ligation event frequency, inclusion of negative and positive ligation controls, iterative template resampling, increased signal to noise ratio and improved sensitivity. Using this technique, we have uncovered a cancer cell-specific, productive chromatin fibre interactome connecting the promoter and enhancer of *c-MYC* to a network of enhancers and super-enhancers. Underpinning this new protocol, I have developed the Nodewalk Analysis Pipeline (NAP) (Paper IV). This suite of tools consists of preprocessing, analysis and post-processing modules designed specifically for the rapid and efficient analysis of Nodewalk datasets through an interactive and user-friendly web based interface.

Overall the work described in this thesis advances our understanding of the role of CTCF in nuclear organization and provides innovative wet lab techniques along with specialized software tools. Moreover, this work is an example of an emerging trend where the challenge of understanding chromatin dynamics within the 3D nuclear architecture demands a close synergistic collaboration between the fields of biology, biotechnology and bioinformatics.

LIST OF SCIENTIFIC PAPERS

I. Guibert, S.; Zhao, Z.; Sjölinder, M.; Göndör, A.; **Fernandez, A.;** Pant, V.; Ohlsson, R*. CTCF-binding sites within the H19 ICR differentially regulate local chromatin structures and cis-acting functions. Epigenetics. 2012 Apr;7(4):361-9. PMID: 22415163

II. Chen, X.; Shi, C.; Yammine, S.; Göndör, A.; Rönnlund, D.; **Fernandez-Woodbridge, A.;** Sumida, N.; Widengren, J.; Ohlsson, R.* Chromatin in situ proximity (ChrISP): single-cell analysis of chromatin proximities at a high resolution. Biotechniques. 2014 Mar 1;56(3):117-8, 120-4.PMID: 24641475

III. Sumida, N.*; **Fernandez-Woodbridge, A.;** Göndör, A.; Ohlsson, R.* Nodewalk, an ultra-sensitive technique to quantitatively analyze stochastic chromatin interactomes, identifies long range-acting super-clusters of productive enhancers. Manuscript

IV. **Fernandez-Woodbridge, A.*;** Göndör, A.; Sumida, N. Nodewalk Analysis Pipeline: a bioinformatic analysis platform for pre-processing, analysis and post-processing of Nodewalk libraries.

Related papers not included in the thesis:

I. Göndör A, **Woodbridge AF**, Shi C, Aurell E, Imreh M, Ohlsson R*. Window into the complexities of chromosome interactomes. Cold Spring Harb Symp Quant Biol., April 5, 2011. PMID: 21467146

II. Stantic M, Sakil HA, Zirath H, Fang T, Sanz G, **Fernandez-Woodbridge A**, Marin A, Susanto E, Mak TW, Arsenian Henriksson M, Wilhelm MT*. TAp73 suppresses tumor angiogenesis through repression of proangiogenic cytokines and HIF-1α activity. PNAS, November 19, 2014. PMID: 25535357

III. Moro C*, **Fernandez-Woodbridge A**, Allister M, Zhang Q, Kandaswamy V, Bozóky B, Danielsson O, Catalano P, Isaksson B, Bozóky B. Integrative Immunohistochemical Classification of Adenocarcinomas of the Pancreatobiliary System (Submitted, Sept 2015)

IV. **Fernandez-Woodbridge A**, A., Shahin Varnoosfaderani, F., Mallet de Lima, C.; Sumida, N., Ronnegren L., A.; Millan Arino, L., Chen, X., Imreh, M., Göndör, A*. Role of TGFBeta in the regulation of the circadian rhythm via modulation of the chromatin architecture. *Manuscript*

V. **Fernandez-Woodbridge A***, A., Ronnegren L., A.; Zhao, H., Shahin Varnoosfaderani, F., Biswas, M., Chen, X., Shi, C., Göndör, A. 3DFishAssist: A tool for fast and accurate quantification of 3D Fish signals at subpixel resolution. *Manuscript*

VI. **Fernandez-Woodbridge A***, M. Barrientos, J. Ulate, Barrantes M. ChipSeqNav: A realtime user friendly visualization of 800+ ChipSeq Experiments *Manuscript*. http://www.chipseqtools.org/chipseqnav.html

VII. **Fernandez-Woodbridge A***, Moro C, Göndör A, Fredlund E. NetClusViz: Visualizing large cohorts with multi-dimentional ngs datasets. *Manuscript*

CONTENTS

LIST OF ABBREVIATIONS

3C	Chromatin Conformation Capture
3D	3 dimensions
4C	circular chromatin conformation capture
5C	chromosome conformation capture carbon-copy
5-caC	5-carboxylcytosine
5-fC	5-formylcytosine
5-hmC	5-hydroxymethylcytosine
5-mC	5-methylcytosine
6C	combined 3C-ChIP-cloning
BAC	bacterial artificial chromosome
BER	Base Excision Repair
Blast	Basic Local Alignment Search Tool
Blat	BLAST-like alignment tool
BWA	Burrows-Wheeler
C	cytosine
Cap3C	Chromatin Conformation Capture + Sequence Capture techniques
ChiaPet	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin immunoprecipitation
ChrISP	Chromatin *in situ* proximity
C-MYC	V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
CTCF	ccctc binding factor
DamID	DNA adenine methyltransferase identification
DMR	differentially methylated region
DNA	deoxyribonucleic acid
DNMT1/3A/3B	DNA (Cytosine-5-)-Methyltransferase
emPCR	emulsion polymerase chain reaction
eRNA	enhancer ribonucleic acid
FACS	Fluorescence-activated cell sorting
FISH	fluorescence *in situ* hybridization
HAT	Histone acetyltransferases
HDAC	histone deacetylase
HMT	Histone methyltransferases
ICR	Imprinting control region
Igf2	Insulin-like growth factor 2 (mouse)

ISPLA	in-situ proximity ligation assay
KDM	Histone lysine demethylase
LAD	Lamin-associated domains
MapQ	mapping quality
NAP	Nodewalk Analysis Pipeline
NL	Nuclear Lamina
PCR	Polymerase chain reaction
PolII	DNA polymerase II
PRMT	Protein arginine methyltransferases
RCA	rolling circle amplification
RNA	Ribonucleic acid
STAR	Spliced Transcripts Alignment to a Reference
TET	Ten-eleven translocation
TF	Transcription Factor
UMI	Unique molecular identifier
WG3C	whole genome 3C

1 INTRODUCTION

The cell is, at its core a massively parallel, information processing system [1]. The first acknowledgement of the underlying heritable "information layer" at the center of every living organism can be traced back to Darwin. In his 1859 book "The Origin of Species" [2], Darwin describes a process of evolution that inherently and unavoidably requires a mechanism of individual information storage. The functional and mechanistic role of this information layer continued to unfold through the works of Mendel and Bovery. Almost a century later, the mechanistic model of the living information system was finally proposed by Kolstov in 1927 [3] and demonstrated through the work of Watson, Crick and Franklin 1953 [4,5]. The structure of DNA molecule is now known to support the reliable and robust storage of information for millions of years providing the main molecular mechanism underlying the evolution of species.

The reliability and robustness of DNA allows organisms to evolve and adapt to the environment through thousands and millions of years, yet for the same reasons, it is unable to respond to environmental changes that take place during smaller time scales, such as 100 year drought, migration from one ecosystem to another, seasonal or daily changes in light/dark cycles. Furthermore, multicellular organisms such as mammals, are composed of various cell-types displaying different phenotypes yet sharing the same genome. These two assertions suggest that there must be alternative, short range information storage mechanism in the cell [6,7].

The genome of a cell is akin to the read-only basic input/output system (BIOS) memory in a modern computer. The genome, is not the actual software of the cell but a boot up control sequence containing the core instruction set. The real biological operating system of a human cell is encoded in part in its chromatome[1], the global structure of chromatin - a material consisting of DNA, RNA and protein. The chromatome is a rewritable information layer which can respond to contextual information and dictates which basic instructions (genes) are to be executed and in what order. In contrast to the genome, the chromatome can be reprogrammed in shorter time scales and its changes can persist from several generations to a single cell division, and therefore it can metaphorically be called as the hard disk drive of the cell. The

[1] The term chromatome [8] is used to encompass trans-meiotic (trans-generational), trans-mitotic and transient (which are mostly lost during mitosis) chromatin marks, which encode regulatory information. The term epigenome, strictly speaking, should only be associated with mitotically or meiotically heritable chromatin marks, yet much remains to be learnt about the stringency of heritability, the persistence and the context specificity of most chromatin marks. Perhaps this uncertainty has propelled the rich and complex evolution of the definition of epigenetics [7,9]. The term was originally coined by developmental biologists to explain how cells containing the same genetic information manage to differentiate and "remember" their identity during the lifetime of the organism [6]. Since the 1950's, with the discovery of new cellular information substrates, the term has sometimes been used to describe a plethora of phenomena beyond that of strictly mitotically inheritable systems [6] to include all chromatin marks regardless of their heritability.

different mechanisms that enable this short term memory have begun to unfold in recent years partly because of the synergistic effect of emerging novel technologies.

In general, chromatin not only packages and compacts the human genome, but also encodes information about its function[2]. Histones and DNA provide physical substrates for numerous post-translational and other covalent modifications with different persistence and dynamics. Currently, ongoing research has mainly focused on 3 different types of chromatin modifications, namely: DNA Methylation, histone modifications and the presence of non-histone, architectural chromatin proteins. Work during the past decades, however, also uncovered that information can also be stored in the 3D organization of chromatin structure. The following work describes the development of tools and studies that pry into the complex structure and function of the 3D nuclear architecture, with the aim of understanding why and how the 3D chromatin organization functions as a short term memory in the cell.

At the beginning of this work, the available bioinformatics and wet lab tools used to study these complex 3D structures were extremely limited. This encouraged us to develop new approaches, which required interdisciplinary collaborations spanning biology, biotechnology and bioinformatics. In the first project we study how multiple binding sites of a chromatin architectural protein within a regulatory element guide the context-dependent formation of 3D chromatin loops linked with the regulation of different nuclear functions. In the second project we present a new microscopy method designed to measure different aspects of 3D nuclear architecture and chromatin fibre proximity in single cells with high resolution. In the final two projects, we present a novel high throughput method designed to examine the interaction patterns of multiple specific loci throughout the genome term Nodewalk, and a suite of bioinformatics tools adapted for its analysis.

[2] Although chromatin is the most widely studied epigenetic vehicle, there are at least some reports of cytoplasmic epigenetics where information is embedded in cytoplasmic proteins such as the pryon and steady state protein pool concentrations. [10,11]

1.1 CHROMATOMICS: PROGRAMMABLE DATA STORAGE WITH VARIABLE PERSISTENCE

The human genome encodes the instructions necessary to build the human being, it does so through long chains of nucleic acids ordered in a precise linear sequence, much like the string of 1's and 0's in a computer's binary code. There are 4 types of nucleotides which are linked linearly to form a 3 billion quaternary code written in a stable and robust double helix molecule, the DNA. The double helix structure first and famously described by Watson, Crick and Franklin [4,5] enables the molecule to function as a long term memory of the cell by reliably and robustly reproducing its sequence during millions of replications with minimal number of errors [12]. The reliability and robustness, however, imposes a restriction on how fast DNA can be naturally reprogrammed within an organism to respond to changes in its environment. Consequently, the newly found mechanism inadvertently posed a new question: How do genetically identical cells produce such a wide variation of phenotypes found in different lineages?

Insights into possible models started to emerge through the process of X-Inactivation [13]. In this biological phenomena, one of the two X chromosomes is silenced in female mammals to compensate for the extra chromosome. Importantly, it was found that this was not the product of genomic alterations. Riggs [14] and Holliday [15] proposed a model in which the guanine preceding the cytosine base in the DNA sequence could be methylated, and this mark could then be replicated using the palindromic nature of the CpG structure. Not only has the existence of this modification been shown, but there are currently experimental evidence on how the cell manages to perform the basic operations of information storage: copy, read and write [16]. Each of these mechanisms employs complex cascades of enzymes, which essentially allow the cell to store information without modifying its genetic core [16].

A second mechanism of information storage was found as studies began the dive into the in vivo form of DNA. In the nucleus of cells, the DNA molecule is wrapped around molecular spools forming the fundamental chromatin unit: the nucleosome [17]. At the lowest level of compaction, these nucleosomes space out regulated intervals forming the so-called "beads on a string" [18] chromatin organization that has been observed and confirmed through electron microscopy. The protruding tails of each one of the different histones forming these spools provide ample opportunities for regulation through modifications of the exposed amino acids. Over 70 different of these post-translational modifications have been described [19]. Although large international efforts such as ENCODE [20], RoadMap Epigenetics [21] initiative have provided invaluable initial insights into the field, the overwhelming majority of modifications have yet to be described.

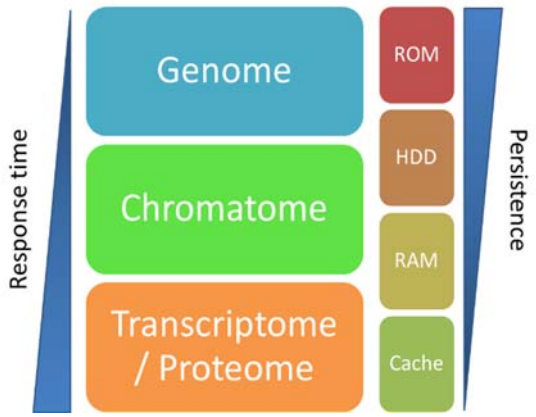

Figure 1 Cellular information systems show parallels with man-made storage systems, highlighting that different tasks require different response times and persistence characteristics.

Contrary to the DNA data medium, these modifications are not stable enough as to provide storage mechanism for the million year scale required in the process of natural selection. Nevertheless, this very same transient nature enables them to provide a "reprogrammable" memory to the cell at shorter time scales. As new cellular storage mechanism are decoded, an emerging parallel between the artificial and biological storage mechanisms becomes clearer as shown in Figure 1 [22]. It would seem that just like human made storage systems, cellular systems require data storage mechanisms that have different persistence and response time. The aforementioned cytosine and histone modifications are examples of chromatin modifications as a volatile short-term memory.

Replication is a mechanism where the DNA and chromatin are duplicated. While proofreading and mitotic checkpoints produce ultrahigh fidelity copies of the DNA sequence, these systems are not known to exist for other chromatin features. Although it is known that chromatin features are actively re-established during S phase (for example DNA and histone methylation at certain genomic locations [16]), much of this process occurs with unknown accuracy. Furthermore, many dynamic chromatin modifications persist only for short periods of time and/or are not passed on to the daughter cells during cell division, and therefore cannot be called heritable (epigenetic). Here lies the fundamental difference between the genome and the chromatome: replication inherently requires the robust duplication of every single base pair, while chromatin marks employ additional and context dependent mechanisms. The levels of the persistence of chromatin marks are thus regulated in a context dependent fashion.

Finally, some chromatin marks can survive the whole genome reset triggered during meiosis and are inherited from one generation to the next [23–25]. Such trans-generational chromatin features allow the cell to propagate information from parent to offspring, enabling for example the fine tuning of metabolism in the offspring in response to the environmental cues of the mother/grandmother [23,26].

The scope, function and importantly the evolutionary processes that created them differ greatly between chromatin modifications that exist within different time frames. While the trans-generational marks may have the highest impact (since they are present from early developmental stage) they also possess the least information. Trans-mitotic epigenetic features may very well be the key to understanding phenotypical differences between cell lineages and transient chromatin marks might ultimately explain stochastic heterogeneity in cell populations. Unfortunately, the lifespan of most of these features remains unknown due to the

fact that chromatomics is a relatively new field with most of the research only recently being made possible through recent technological advances. Within these poorly studied systems, perhaps the best known modifications are those impinging on cytosine methylation and histone modifications.

1.1.1 Modifications of the Primary Chromatin Fibre: Cytosine Methylation and Histone Modifications

1.1.1.1 CpG methylation and its role in the regulation of gene expression

Cytosine Methylation is perhaps the best understood chromatin modification. This mechanism allows the cell to covalently modify the cytosine base in the DNA strand. This modification is not only reversible but is actively copied during cell division providing the cell with a "short term" memory. Furthermore, once modified the base might undergo a complex cycle of modifications(C > 5-mC > 5-hmC > 5-fC > 5-caC > C) until finally being transformed into a mismatched base that is excised and repaired [16]. Recent studies suggest that at least some of the different intermediate chemical states may be interpreted differently by the cell, thereby providing a large range of variation [27]. The epigenetic information medium is established, maintained and remodeled by different molecular mechanisms. For example, *de novo* melthyltransferases (DNMT3A and 3B) deposit the methyl marks (writing), the maintenance melthyltransferase enzyme (DNMT1) copies the established methyl marks to the newly synthesized DNA (copying) and a complex interplay between different enzymes, such as TET enzymes and members of the base excision repair (BER), can achieve de-methylation (erasing/writing) [16]. Finally it is known that many transcription factors, CTCF for example, are not only able to recognize methylated sites and bind preferentially to unmethylated sites (reading), but also protect their binding sites from *de novo* methylation [28].

The function of DNA methylation has been extensively studied at several cis regulatory elements, most notably the CpG islands of promoters [16]. Interestingly, CpG-rich promoters are usually free of DNA methylation and are found in the promoters of housekeeping genes [29]. The methylation of promoters moderately enriched in CpGs has been shown to impact gene expression by preventing the binding of transcription factors, thereby effectively silencing the gene. CpG-poor promoters, on the other hand, tend to be unaffected by DNA methylation [30].

1.1.1.2 Reprogramming of DNA methylation patterns during development

The chromatin of stem cells is generally unmethylated, and gradually gains DNA methylation as different cell types emerge during development [31]. Generation of induced pluripotent cells on the other hand requires the removal of the mature cell type-specific DNA methylation pattern, although this erasure is not always complete [32]. There are two major reprogramming events during development, where DNA methylation patterns are erased and re-programmed. The first wave of de-methylation takes place during gametogenesis, when all acquired epigenetic marks must be rolled back [33]. It is during this process that the male and

female gamete-specific chromatin marks are established, among them methylation patterns of imprinting control regions that control the parent of origin specific expression of imprinted genes. The second wave takes place after fertilization, this event serves to establish the totipotent state in the zygote. This epigenetic reprogramming, however, does not affect the established parental specific marks at imprinted genes [28].

There are some regions of the genome that have developed the ability to escape the epigenetic reset during germline development. These trans-generational epigenetic marks have the potential to carry information from parent to child. Studies have shown that although these features are rare, they have strong effects on offspring. An early study for example, showed a predisposition to develop diabetes in the grandchildren of women who experienced long periods of hunger due to long winters in isolated towns in the north of Sweden [23,26]. One plausible evolutionary interpretation of this is that long periods of scarcity may trigger epigenetic changes in the germ line aimed at fine tuning the metabolism of the offspring as to make it more suited to the environment experienced by the mother.

1.1.1.3 Genomic Imprinting

Importantly, the reprogramming process differs significantly between the maternal and the paternal germlines. The spatial separation of parental genomes provides opportunity to establish and store parent of origin specific information at imprinting control regions. One such locus is represented by the *H19* imprinting control region (ICR), a small 7 kbp sub-telomeric regulatory element on the short arm of the human chromosome 11[34]. On the maternal allele, this region maintains its unmethylated status, while on the paternal allele the locus is methylated. The information encoded in this region is very unique, as the differential epigenetic states established during male and female germ line development are maintained in the somatic cells of the offspring throughout development. This allows the paternal and maternal loci to have different behavior. For example, the manifestation of the imprint results in the parent of origin specific monoallelic expression of two imprinted genes: the *H19* non coding RNA and *IGF2 [35]*. As opposed to random monoallelic expression that does not differentiate between parental alleles, *H19* is always expressed from the maternal allele whereas *IGF2* is expressed from the paternal allele.

The imprinted expression of the *H19* and *IGF2* genes is regulated by the binding of CTCF to the *H19* ICR. CTCF binds to the unmethylated maternal allele of the ICR during female germline development, thereby protecting this locus from acquiring DNA methylation in the somatic cells of the offspring. Furthermore, the maternally bound CTCF acts as an insulator protein that prevents the enhancer to active promoter of the *IGF2* gene on the maternal allele. On the paternal allele, DNA methylation marks acquired during male germline development prevent CTCF binding and spread onto the promoter of the paternal *H19* gene, leading to its repression. In the absence of CTCF, the enhancer is now capable of activating *IGF2* expression.

An interesting evolutionary theory states that the evolution of imprinted gene expression can be traced back to ancient polygamous marsupials. Hence, as females would bare the offspring of multiple males simultaneously, natural selection would inevitably push the paternal chromosomes into an arms race as to induce the expression of growth promoting genes in the offspring, resulting in competition between offsprings of different fathers [36]. It would be easy to see how such an arms race could become detrimental to the mother and therefore not evolutionarily viable. This scenario leads to a parental conflict between the genomes and results in the expression of growth promoting imprinted genes from the paternal chromosomes and growth inhibitory imprinted genes from the maternal genome [36]. This could explain why the *H19* ICR mediates the paternal expression of *IGF2* during development [37]. To provide insights into this and other evolutionary perspectives we further investigate the mechanism underlying this interesting regulatory element in the first paper.

1.1.1.4 Histone modifications

While cytosine methylation is encoded directly in the DNA backbone, the nucleic acid chain is not the only substrate for information. *In vivo* the DNA backbone is embedded in a wide variety of structural and regulatory proteins, which provide further opportunities for modification. Of these, perhaps the most studied are the modifications of the nucleosome, the basic unit of chromatin. As mentioned earlier, nucleosomes are molecular spools which coil DNA into compact structures critically packing a 2 meter long human genome into a miniscule 7-20 μm diameter nucleus [38]. The nucleosome is composed of an octameric complex containing 2 copies of one of the 4 core histone types (H2A, H2B, H3, H4), whereas the linker region between nucleosomes binds in some cases a stabilizing H1 histone [39]. The highly conserved structure of the nucleosome core coils DNA (147 nucleotides or 1.5 turns) leaving 8 protruding histone "tails". Once formed, the nucleosome can also be stabilized by the H1 linker histone that binds both the incoming and outgoing strands as well as the core itself providing greater stability to the entire complex [39]. Critically, each one of these 5 histones present protruding tails which remain accessible for posttranslational modifications. Currently, over 10 different chemical modifications [40] have been identified in different amino-acids in these tails; just in H1 for example, 48 different modified locations have been observed through mass spectrometry [39]. In some cases, the enzymes that perform these modifications have already been characterized such as histone acetyltransferases (HAT), histone methyltransferase (HMT) and protein arginine methyltransferase (PRMT). Furthermore, enzymes that can erase these modifications such as histone deacetylases (HDACs) and histone lysine demethylases (KDMs) and some of the factors that can recognize the changes are already known (genes with bromo, chromo and Tudor domains [41] have also been characterized).

Initial studies have shown that some repressive histone marks, like the H3K9 di- and tri-methylations are inheritable in yeast [42], yet the mechanism by which this is achieved is still a matter of debate despite ongoing research [43]. There are currently 2 different models. The semi-conservative model proposes that nucleosomes are split into histone tetramers or dimers which are then complemented with naïve "halves". After replication, the inherited

histones present information, which can then be used as template to copy the information to the naïve half. In the second model (the random model) the histone octamer is not split, but is randomly assigned to either of the DNA strands [44]. In both models, copying enzymes might re-establish the modifications using the original histone octamere/tetramer, the same strategy employed in DNA methylation where the methylated C serves as a template to methylate the unmethylated complement of the G in the CpG after replication.

Recent technological advances have enabled the study of a few of these marks although currently most studies overwhelmingly focus on modifications of the H3 tail and to a lesser extent of the H4 [20,21]. These studies have revealed a rich variation between cell types and initial identification of the general patterns surrounding poised (H3K4me1) and active (H3K4me1, H3K27ac) enhancers, active (H3K27ac, H3K4me3) and inactive promoters (H3K27me3 or H3K9me2/3). From these and numerous other studies it is clear that histone modifications compose an information layer and actively participate in the regulation of gene expression yet causality of this relationship is remains largely unknown [45–49]

1.1.1.5 Replication timing

The replication of the genome follows a temporal sequence, which both influences and is influenced by marks of the primary chromatin fibre. Hence, active genes and regions packaged in open chromatin configuration tend to replicate early, whereas repressed domains and silenced genes tend to replicate late during S phase [50]. Importantly, the timing of replication can also instruct the deposition of chromatin modifications on the newly replicated DNA. On one hand, early replication promotes the establishment of open chromatin structure that increases the probability of gene expression in the next cell cycle. Late replication timing, on the other hand, promotes the establishment of compact repressed chromatin. Due to the two-way relationship between chromatin and replication timing, the timing of replication at a given locus is a heritable feature, providing a vehicle for epigenetic inheritance [50]. Interestingly, genomic loci that display random or stable monoallelic expression, such as imprinted genes, tend to acquire asynchronous replication timing during development. Apart from chromatin features, an important regulator of replication timing is the transcription factor CTCF. Hence it has been shown that mutation of the CTCF binding sites in the mouse maternal *H19* ICR leads to synchronous early replication timing of both alleles. The context-specific role of individual CTCF binding sites within the mouse *H19* ICR will be further explored in the first paper.

1.1.2 Spatial Chromatome: 3-Dimensional Nuclear Architecture

Although chromatin marks can affect the accessibility of the underlying DNA code and serve as docking sites for various factors, an emerging view is that beyond the local, “linear” effects of these modifications there lies a richer environment of 3-dimensional configurations that can themselves hold information. However, beyond the basic “beads on a string” structure of the primary chromatin fibre, we are still lacking a validated model describing the folding of the genome within the nuclear space. A few models have been proposed, such as the 30 nm

chromatin fiber [51], as well as the 120 nm and 170 nm structures [52], the fractal globule model [53], the chromosome territory / interchromatin compartment [54] and the interchromatin network model of chromatin [55], yet direct in *vivo* demonstration of these models is still a matter of debate. The ambiguity in the field is partly due to that most of the existing microscopy methods that can validate the models work only in fixed cells, whereas methods that operate in *vivo* do not possess the resolution to decipher the detailed 3D structure and dynamics of chromatin. To overcome these limitations, new emerging technologies are rapidly advancing to resolve these structures [56]. One of these technologies is presented as part of this thesis and will be discussed in the following chapters.

What is clear, and has been since the first observations by Heitz in 1928 [57], is that interphase chromatin contains both highly condensed structures (which he called heterochromatin) and relaxed regions (later called euchromatin). Euchromatin has been identified as transcriptionally permissive chromatin containing genes, whereas heterochromatin has been mainly correlated to inactive genes/gene poor regions. These structures have also been extensively correlated with different histone modifications - suggesting that these 3D configurations can themselves be regulated and therefore may perform regulatory functions making them a plausible candidate for information storage.

1.1.2.1 Organization principles in 3D genome organization: spatial separation of active and inactive chromatin

To understand the information encoded in the 3D nuclear architecture it is important to understand the organizational principles that regulate the arrangement of the genome in the nucleus. In interphase nuclei, each chromosome occupies a relatively confined space called chromosome territory [58]. Although the existence of chromosome territories has been well documented, gene-rich chromatin fibres are highly dynamic and may loop out from their corresponding chromosome territory to mingle with chromatin fibres of other chromosomes [55].

The "macro" organization of the chromatin includes the radial orientation of chromosomes: the nuclear periphery (close to the nuclear membrane) is transcriptionally repressive, and contains regions of the genome that are constitutively or developmentally repressed [50]. The interior of the nucleus is, on the other hand, transcriptionally permissive and contains gene-rich regions and active genes. With the advent of new technologies it has been made possible to identify substructures within these macro structures [58] that are tightly linked to the establishment of cell type specific gene expression patterns and cellular memories [59].

An important organizer of the peripheral localization of repressed chromatin domains is the nuclear lamina [60–64]. The lamina (NL) is a fiber mesh attached to the surface of the inner nuclear membrane, and is composed of A type and B type lamins as well as other proteins [65]. It has been shown that this layer provides a rich functional interphase and a docking site for regions of different chromosomes called lamina-associated domains (LADs). While most

of the LADs overlap with heterochromatic, gene-poor regions, many LADs contain developmentally repressed genes that are positioned to the nuclear periphery in a cell-type specific manner. In line with the cell-type specific chromatin-lamina interactions, the constituents of the lamina are also known to be developmentally regulated [66]. For example, while Lamin-B is constitutively expressed throughout many cell types, Lamin-A appears to be absent at the initial developmental stages [66] and other nuclear envelope transmembrane proteins are also expressed in a cell-type specific manner [67]. Chromatin-lamina interactions during development are regulated by sequence-specific transcription factors and histone modifications, such as H3K27me3 and H3K9me2 [68]. In summary, chromosomes not only cluster at the nuclear periphery, but this colocalization also leads to functional interaction between the NL and the LADs [69,70] *via* lamin and core histones communication [71].

Together, these results start to elucidate how epigenetic regulation of the histone marks can potentially propagate into higher order chromatin conformation changes. In turn, the spatial separation of transcriptionally permissive and repressive environments is considered to promote the stability of cell-type specific epigenetic memories and thereby the cell-type specific expression patterns. Despite its significance, the study of LADs, their structure and function is hampered by the technical limitations [59].

Understanding the mechanism of chromatin-lamina interactions requires the development of novel methods that can overcome the technical difficulties posed by these compact, difficult-to-digest and often repetitive regions of the genome, which makes it difficult to analyze LADs by chromatin immunoprecipitation (ChIP). To this date, only a handful of datasets are available that describe chromatin-lamina contacts in a high throughput manner [61–63]. These datasets were generated using an assay called DamID, which is based on a fusion protein between lamin B1 and bacterial adenine methyltransferase. The introduction of this system leads to the methylation of sequences that are in close proximity to the lamina. By detecting the methylation marks the assay provides indirect information about the identity of LADs, although very little is known of the artifacts and biases of the technique. In the second paper we present a new assay, called Chromatin *In Situ* Proximity, that can assist in the study of LADs in single cells and their functional and structural implications in higher order chromatin structures.

1.1.2.2 Chromatin crosstalk between regulatory elements

While LADs are typically associated with long-range repression, other higher order features, such as loops, can be associated with activation or repression of specific genes. Within this larger context, the loop structure is central to the subject of this thesis. Loop regulation enables for example the mouse *H19* ICR locus to pleiotropically regulate the replication timing of other imprinted loci located within different chromosomes during germline development [34]. In addition, it also allows promoters to colocalize with long distance enhancers [72].

Enhancers control the emergence of cell-type specific differential gene expression during development and have been shown to play key roles in diseases [72]. The factors that

define enhancer identity are the subject of intensive investigation. Hence, enhancer elements are short DNA sequences that contain a high concentration of binding sites for transcription factors that recruit chromatin modifiers, the mediator complex and multiple components of the Pol II machinery. By looping the enhancer-bound complex and placing it in direct contact with the gene promoter, enhancers activate the expression of "nearby" genes [72].

This mechanical colocation between enhancers and promoters is mediated by multiple factors. First, the general open or close configuration of the chromatin may facilitate or block TFs from binding to enhancer elements. Open chromatin that marks active enhancers contains H3K4me1 and is rich in H3K27ac modifications [40]. Decommissioning of enhancers, on the other hand, is linked with histone demethylase and deacetylase activity [73]. A third category of enhancers may reside in a poised state marked by H3K4me1 without the simultaneous presence of H3K27ac marks, which enables the cell to execute quick and specific responses to environmental or developmental cues [74]. Second, enhancer-promoter communication is also promoted by the transcription of the so-called enhancer RNAs (eRNAs), that promote the recruitment and kinase activity of the mediator complex [75]. Finally, loop structure and dynamics is also influenced by functional interactions between chromatin architectural proteins, such as the interplay between cohesin and CTCF [76] .

To stabilize the loop, the cohesin complex links the 2 DNA strands together while CTCF acts as a positional element restricting the movement of the bound strand, thereby aligning the enhancer to the promoter [72,77]. In this complex interplay, CTCF effectively helps align or prevent the alignment of the promoter-enhancer loci by different methods. It can for example prevent the alignment by forming alternative loops [78,79] or it can directly bind to cohesin thereby fixing its movement on the DNA strand. For this reason, although CTCF received the name "insulator protein", it is clear that the mechanism of insulation is achieved by coordinating the interactions between regulatory elements in 3D. Hence, depending on the chromatin context, CTCF can facilitate both the alignment and the misalignment the distant loci, explaining why CTCF may act sometimes as an insulator, while at other times as a facilitator of gene expression. The different context-dependent roles of CTCF are further explored in Paper 1.

Recently, a unique category of enhancers has been discovered that is deeply influenced by the 3D architecture of chromatin. These so-called "super-enhancers" differ significantly from regular enhancers in several features. Firstly, super-enhancers are typically composed by large clusters of linearly dense enhancer elements, possess high mediator occupancy, are decorated with H3K27ac and present significant levels of eRNA expression [80]. Second, these large clusters have been shown to regulate lineage-specific and cell fate determining genes in isolated chromatin loops [80]. Third, the multiple enhancer subunits within super-enhancers have been shown to integrate signals from different cell fate determining pathways in combinatorial patterns, ensuring high probability of transcription at target genes [81]. While these hubs may very well represent linear organization regulating upstream and downstream genes, it is still unknown if under certain circumstances such clusters can also form in 3D by

simultaneously collocating enhancers in *trans*. In the third paper we provide a fresh new look at the 3D interactome of super-enhancers opening the possibility to new configurations that are not bound by linear clustering.

1.2 BIOTECHNOLOGY: OBSERVING NUCLEAR ARCHITECTURE THROUGH THE MICROSCOPE/SEQUENCER

3D chromatin structures can be explored with multiple techniques that can be grossly grouped into 2 major categories: microscopy and sequencing. Both techniques are able to measure common features, such as proximity and interaction, yet these measurements differ greatly in spatial and temporal resolution. Microscopy provided the first observations of the 3D chromatin structure, and is still used for macro overviews of the general organizational characteristics of chromatin at the single cell level. Microscopy is the only spatially aware method in which it is possible to quantify both proximity (average distance) and potential for interaction (% of the time in which 2 elements are in direct contact) in the context of the nuclear organization (i.e. interaction at the periphery, proximity to the nucleolus, proximity to nuclear pores). Finally, these techniques also provide ways to assess population heterogeneity, as each cell is measured independently.

Sequencing techniques are also able to measure proximity and interaction, however, they are typically aimed at observing the average configuration of chromatin conformations in large cell populations, in a high throughput manner [53,82–84]. Although low cell count techniques are also emerging [85], the vast majority of these assays measure different aspects of the nuclear architecture typically in large cell populations but at much higher genomic resolution than microscopy-based techniques. Some techniques target transient, yet functional interaction events, such as triggering of transcription through enhancer-promoter interaction [86], or loop locking by CTCF [87]. Others aim to describe the general localization of chromosome territories in single cells [58]. Yet another category of techniques describes the spatio-temporal organization through time in large cell populations [53].

As with any technique, each methodology also possesses different accuracy, sensitivity and spatio-temporal resolution. The differences between the strengths and weaknesses of these techniques provide ample room for innovation, giving rise to new tools such as the ChrISP technique described in Paper II and the Nodewalk assay described in Papers III and IV.

1.2.1 Microscopy based techniques for the analysis of 3D Nuclear Architecture

Fluorescent *In Situ* Hybridization (FISH) is perhaps the most widely adopted technique in 3D studies and is often used as a validation of sequencing based techniques. This protocol employs fluorescently labeled sequence-specific probes to bind to their complementary sequence in fixed nuclei. Employing large chromosome territory probes Cremer et al [58] show clear delineation of the core chromosome territories and other macro structures. Expressed regions on the other hand loop out from the core chromosome territory, creating a corona that intermingles extensively with other expressed regions located on different chromosomes [88]. Smaller BAC probes (~200kbp) pinpoint the location of specific genes and can be used to identify the proximity between these regions in small (100-10k) populations of cells. Using different probe sizes and targets (DNA/RNA/Protein), FISH enables the examination of small structures and chromatin fibre proximities in the single cell level.

FISH is limited, however, by the resolution of the underlying microscopy technique that varies depending on the optical system. In conventional confocal microscopy, the fluorophore diffraction limits the resolution of the depth dimension to ~300nm [89]. Enhancements in image processing and optics enable super resolution microscopy techniques to improve the resolution to 10-50 nm [89] - yet they still require very specific labeling, are limited by the number of "colors" they can detect and require specialized/costly hardware.

1.2.2 Next Generation Sequencing and chromatin conformation capture

While Microscopy techniques provide single cell measurements, they usually lack the resolution to score for tight interactions between specific regulatory elements, such as promoters, insulators, enhancers, differentially methylated regions. The advent of the next generation sequencing (NGS) provided a different approach to uncover 3D chromatin organization. DNA sequencing is the process of converting a physical DNA fragment into a digital code of its nucleic acid composition. In the context of the 3D nuclear architecture, sequencing allows us to decode the sequence of DNA fragments that were found in tight proximity of one another. Given that these fragments usually contain a uniquely identifiable sequence, it is possible to reverse engineer the point of origin of each fragment in the genome.

The techniques that apply sequencing to map 3D chromatin fibre interactions build on the chromosome conformation capture (3C) technique originally developed by Dekker et al [90,91], which was the first method to score for interactions between 2 defined loci in the genome. In this technique the nuclear architecture is fixed by formaldehyde crosslinking. These fixed cells present a snapshot where chromatin regions that were in close proximity in *vivo* are covalently crosslinked to each other by formaldehyde molecules. The treatment prevents interacting fragments from drifting apart when the nuclei are lysed and chromatin is digested by restriction enzymes. These enzymes are molecular scissors which cut chromatin containing a specific sequence. The digested 3C complexes are then isolated from each other by diluting the sample. Next, the open ends of the complex are ligated together by DNA ligase under dilute conditions. Once ligated, these chimeric products can be assayed through different methods to

determine the presence of ligation events which are then used to extrapolate interaction and proximity rates in the living cell.

The first 3C is a powerful technique to assay the proximity or interaction between 2 specific loci. However, it is not capable of uncovering interactions between distant loci without a prior knowledge or educated guess about their identity. This limitation is due to the fact that 3C employs specific PCR primers for the detection of ligation products in chimeric DNA. With the advent of microarrays and next generation sequencing it became possible to develop the 3C assay further into higher throughput techniques. Using a single bait, the circular chromosome conformation capture assay (4C)[34] employs a circularization step, that enables the use of inverse primer pairs to amplify unknown interacting fragments ligated to a known bait region. This technique has the advantage of being able to discover novel proximal and interacting regions of a known bait (one to many) but requires both ends of the bait fragment to ligate efficiently to its interactors. Critically, the different variants of the 4C assay [34,92] provide a thorough and focused view of the 3D landscape, but they are limited to a single locus. In contrast, the HiC[53], a whole genome 3C technique (WG3C), provides an overall view of the proximity between large regions of the genome, yet lacks the sensitivity to score for transient long-range interactions. The HiC suffers from an exponentially decaying resolution, and therefore the technique can only score for proximity between "small" elements found in the vicinity of each other on the linear chromosome [82]. To bridge the gap between high resolution yet focused sensitivity of the 4C and the macro overview of the Hi-C, Dekker et al introduced the carbon copy 3C protocol or 5C[93]. The method allowed for the multiplexing of hundreds of 3C experiments allowing for a "many to many" approach where any combination between a fixed set of loci is targeted by the primer combinations. Most recently, the combination of the HiC method and sequence capture technologies enabled further focusing of the HiC method [94–96]. These 3C Capture (Cap3C) methods provide a many to all assay. Despite the increase in sensitivity and bait segments, the "many" methods usually refer to hundreds of loci or less, yet some applications demand the screening of thousands of loci.

1.3 BIOINFORMATICS OF 3C NGS LIBRARIES

With the availability of cheaper, faster and high throughput next generation sequencing, the number of WG3C and Cap3C has grown considerably in the last years. As with most techniques, each protocol requires in turn a custom analysis pipeline designed to convert mountains of digitized DNA sequences into meaningful quantifications of the nuclear architecture. Typically, the interpretation of these digital DNA sequences is done through a process of preprocessing, mapping, filtering, summarizing, normalizing, annotating, screening and visualizing [97–107].

1.3.1 PreProcessing/Filtering

Pre-Processing is the process of generating a "sequencing independent" format. These tasks typically include base calling (converting the sequencing images and voltage readings (IonTorrent) to base pair representation (A,T,G,C)). This typically results in a fastq file. After this process the resulting reads are trimmed to remove low mapping quality at the ends (artifact of the Illumina platform), followed by the removal of sequencing adapters and multiplexing adapters. Additionally, in some cases it is necessary to eliminate PCR duplicates which can introduce false positive interactions in extreme cases [94].

1.3.2 3C Mapping: single end mapping of paired end libraries

After removing sequencing and library preparation related artifacts in pre-processing, most pipelines must then disambiguate the origin of the read in a process called mapping. This process is supported by a wide variety of well-known tools, such as Blast[108], MAQ[109], Blat[110], BowTie[111,112] and BWA[113] and most recently STAR[114], though most 3C datasets are usually mapped with BWA or BowTie due to their speed and accuracy. Mapping algorithms find the best match for a DNA sequence, by comparing each read to the reference genome. Due to the chimeric nature of 3C libraries, mapping is usually performed at only one end at the time, which significantly reduces the accuracy and sensitivity of the process. Since false mapping artifacts can introduce false positive signals it is critical to assess the accuracy, sensitivity and limitations of the mapping algorithms in the pipeline.

The mapping process assumes that the site of origin of each read in the reference genome displays the highest similarity to the observed read, and therefore by finding regions with the highest similarity to each sequenced read it is possible to pinpoint their genomic origin. This is done by scanning the reference genome for hits (regions that bear some similarity to the read), yet since there are bound to be differences between the reference and the observed sequence, mapping algorithms must also generate alignments. Critically, these alignments must compensate for insertions and deletions. These mutations are very common in organisms and introduce a high level of uncertainty in the mapping process. Furthermore, the 3C template undergoes digestion, ligation and PCR amplifications, which might introduce further differences in the reads. The mutations/artifacts force the mapping algorithms to test different ways of aligning a read to the same genomic location with different combination of insertions and deletions, rather than comparing each location base by base. This simple change immensely

increases the complexity of the task, as there is an almost infinite number of combinations for each read and for each position. While different algorithms can estimate these alignments under restricted conditions for a handful of reads, performing this analyzes for next generation sequencing with millions of reads is impractical. This limitation prompted the creation of mapping algorithms that employ heuristics, which rely on "shortcuts" and are able to dramatically decrease the time and resources required to map these massive libraries [110]. The cost for these heuristics is however that they are not always able to find the optimal alignment. To compensate for this, these algorithms employ probabilistic models, which estimate the probability of suboptimal mapping for each alignment, typically called the mapping quality (MapQ)[109]. The MapQ level represents the level of confidence that a reported alignment is the true alignment. This parameter is estimated in different ways using different models. In BWA long aligner for example, MapQ is calculated in relationship to the difference between the best and second best alignments. This accounts for scenarios where the read can have a nearly perfect alignment to one region of the reference genome as the top hit, yet the second best hit is only slightly worse. While this case will be assigned a poor MapQ value, a case where the second best hit has a very poor alignment would get very high MapQ. In this way the MapQ factors in two characteristics: the alignment score and the uniqueness of the alignment.

MapQ is not a universal assessment of sensitivity or accuracy, because this relationship is context dependent. Instead, it depends on read length, availability of paired- or single-end reads, complexity/size of the reference genome, sequencing/library preparation error rates and target search regions. Read length and the availability of paired- vs. single-end reads have the strongest and most visible effects on mapping efficiency. Longer reads provide the aligner with more information, than short reads and therefore have higher mapping efficiency. As shown in our benchmarking study[3] human 40bp reads with one mismatch represent 85% sensitivity and 99% specificity, while 20bp reads with one mismatch show 18% sensitivity and 80% specificity. Furthermore, this performance cannot be extrapolated to other genomes, such as yeast or bacteria, which have diploid or haploid genomes containing lower number of repeats, and a genome size several orders of magnitude smaller than that of the human genome.

1.3.3 Validation and Summarization

As the mapping process generates an equal or greater number of observations than the input, a critical step before progressing the analysis is to summarize the data to produce more manageable datasets. Typical examples of summarization include counting the number of reads overlapping a gene in RNA-Seq [115–117], finding enriched regions or peaks in ChipSeq datasets [118,119] or counting the number of interactions between different loci in 3C

[3] To select the best aligner for the Nodewalk pipeline, we performed a series of benchmarking experiments in order to profile the speed, sensitivity and specificity of the different algorithms. These results are available at: http://www.chipseqtools.org/benchmark-results.html

experiments [120]. This critical step significantly reduces the size of the data by several orders of magnitude. During the summarization process, it is common to introduce filters that discriminate different reads. For example some HiC pipelines filter out reads that do not map within a threshold distance from a restriction site [53].

1.3.4 Normalization

The resulting summary of the data is then analyzed to remove further artifacts. In contrast to the preprocessing, artifacts targeted in the normalization procedure typically arise in the wet-lab or library preparation steps preceding sequencing. These are extremely important in chromatin conformation capture techniques, which are plagued with biases introduced by restriction enzyme digestion/star activity, PCR artifacts, mapping artifacts, chromatin accessibility- and GC content-artifacts [98,100,121,122].

2 AIMS OF THE THESIS

The overall aim of the thesis was to investigate the structure and function of three-dimensional chromatin organizations in the mouse and human genomes. Furthermore, the studies also aimed at developing novel technologies to explore chromatin conformation in single cells at high resolution and in small cell populations. The thesis builds on four papers with the specific aims to:

1. Examine the context-dependent nature of CTCF binding sites in the regulation of chromatin fibre interactions at the mouse *H19* ICR (Paper I).
2. Develop a novel, microscopy-based assay that enables the visualization of chromatin compaction and chromatin fibre interactions in single cells, at high resolution (Paper II).
3. Develop a high throughput assay for the quantitative detection of chromatin fibre interactions between many baits and the rest of the genome in small cell populations, which we termed Nodewalk (Paper III).
4. Develop bioinformatic pipelines specifically tailored for the analysis of the results generated by Nodewalk assay (Paper IV).

3 RESULTS AND DISCUSSION

3.1 PAPER I: DECONSTRUCTING THE FUNCTIONAL ELEMENTS OF THE *H19* ICR INSULATOR

The chromatin insulator protein CTCF is a major factor underlying the domain organization of the epigenome, the formation of higher order chromatin structures and the regulation of chromatin crosstalk [78,87,123]. This factor is typically associated with insulator function in mammals due to its role in manifesting the function at the imprinting control region in the 5'-flank of the *H19* gene (*H19* ICR). This 4kb sub-telomeric region on the mouse chromosome 7, or 7 kbp region in human chromosome 11 is inherited in a CpG-methylated version from the father, while the maternal allele is generally methylation-free. This arrangement underlies the parent of origin-specific expression of the maternal *H19* and the paternal *Igf2* alleles, which are separated by approximately 100 kb. Thus, the CTCF binding sites within the *H19* ICR are occupied only at the maternal allele preventing the *H19* enhancers from communicating with the *Igf2* promoters, thereby repressing its expression. Conversely, the methylated status of the paternal *H19* ICR allele hinders CTCF binding, thereby enabling the *H19* enhancers to activate *Igf2* gene expression [124] Studies into the function of CTCF produced mouse strains in which these binding sites were knocked out [125].

Within these mutants, a particularly interesting strain termed 142 provided an opportunity to explore if the CTCF binding sites functioned as backup to each other or if they had different functions. This strain was generated by knocking out three of the four CTCF binding sites within the *H19* ICR. Interestingly the presence of the neomycin gene, which was used during the selection of the transfected embryonic stem cells, kept the maternally inherited, mutant 142 allele generally methylation-free. Importantly, site 2 remained occupied by CTCF, as determined by ChIP analysis, although this assay is not able to discriminate between CTCF occupancy at sites number 1 and 2 due to the spatial proximity between these two sites in the linear sequence. Surprisingly, this single unmutated CTCF binding site was able to functionally manifest the imprinted gene expression state of *Igf2* and *H19*. However, when the neomycin gene was removed by in *vivo* recombination, the maternal 142* allele (the floxed allele) gained DNA methylation and was unable to insulate the maternal *Igf2* allele from the enhancer any more. At the same time this strain displayed a concordantly lower expression of *H19*. Additionally, the reduced CTCF binding in the floxed and unfloxed versions of the maternal *H19* ICR shifted its replication timing in both strains to early replicating, suggesting that this feature is independent of the status of CTCF occupancy at CTCF binding sites 2 or 1/2.

(f)x(m)	Total *H19*	Maternal Igf2	Rep. Timing
SD7 x 142	High	Inactive	Late
SD7 x 142*	High	Inactive	Late
142 x SD7	High	Low*	**Early**
142* x SD7	**Low**	**High**	**Early**

Table 1 Differential contribution of CTCF binding sites within the mouse H19 ICR to the regulation of replication timing and imprinted gene expression.

The phenotypic differences observed in replication timing and imprinting suggested that the chromatin structure around sites 1/2 could be very different from that of sites 3/4. To examine this possibility, a 4C experiment was performed using 2 different restriction enzymes, which separated sites 1/2 and 3/4 into different baits (referred to 5' and 3' respectively). The resulting patterns indeed documented that the 3' and 5' ends of the *H19* ICR have somewhat overlapping but also very distinct interaction patterns. Moreover, the interaction pattern of the 5'-end in the floxed *H19* ICR allele (142*) presents higher frequency of interactions than the wild type allele, suggesting that CTCF occupancy of sites 1/2 in the wild type *H19* ICR prevents the formation of extensive chromatin fibre interactions. Conversely, the 3'-end in the floxed *H19* ICR allele (142*) presented significantly lower number of interactions than the wild type.

We conclude that the CTCF binding sites at the 5'-region not only provide specificity to long-range contacts but also protect against chromatin compaction, whereas CTCF binding at the 3'-end is required for the establishment long-range interactions. Taken together these results uncovered that the different CTCF binding sites within the *H19* ICR perform different roles with respect to the regulation of replication timing and imprinted gene expression, as well as the establishment of chromatin fibre interactions. By extrapolation, the distribution of CTCF binding sites in the genome cannot automatically be translated into insulator functions. Moreover, their roles might be context-dependent to promote transcriptional activation [126], for example. These results further demonstrate CTCF's role as a master weaver of the genome [123] by showing how the insulator function may rely on the characteristics of the loop formed by CTCF and not CTCF itself.

3.2 PAPER II: VISUALIZING STRUCTURAL DENSITY AND PROXIMITY WITH CHRISP

While many techniques exist for analyzing nuclear architecture, its sheer complexity and range of scale makes it impossible for a single technology to encompass every aspect of these structures. Hence, different technologies manage only to capture individual aspects at very narrow temporal-spatial scales. For example, while the 3C-based methods asses proximity/interactions within large cell populations [34,53,94,96], they provide little insight into the heterogeneous and stochastic behavior observed in single cell protocols[127]. Alternatively small cell populations can be studied using DNA Fluorescent *In Situ* Hybridization (FISH), yet this method cannot score for the proximity of 2 loci beyond 250 nm in 3D space. Although this resolution is good enough for many applications, it is not possible to study the functional interactions between regulatory elements, which requires direct contact between them. The *In Situ* Proximity Ligation Assay (ISPLA) can detect close proximity (<40 nm) between 2 proteins [128], but the underlying Rolling Circle Amplification (RCA) step provides a very low spatial resolution. Moreover the amplification rate introduced by the RCA step hinders linear quantification, as opposed to DNA/RNA FISH. While a rapid RCA amplification can quickly saturate a particular signal localized in an open accessible compartment, interaction events localized in dense heterochromatin may provide more hostile microenvironments for the RCA process. Consequently, we set out to create a method, which would combine the strengths of both FISH and ISPLA technologies in order to effectively assess interaction between different loci / proteins in the nucleus. The resulting chromatin *in situ* proximity (ChrISP) combines the direct contact detection of ISPLA with the locus-specific nature of FISH.

The method employs the original ISPLA ligation system to score for interaction. This system is built on antibody pairs used to identify either biotin or digoxigenin. These targets can be integrated into both DNA/RNA probes and/or a primary antibody providing support for both DNA/RNA and protein targets respectively. The secondary antibodies carry a covalently bound DNA adapter that has a sequence specific to each antibody type. After the antibodies have been hybridized to their specific epitopes and the sample is washed to eliminate unbound antibodies, a fluorescently labeled splinter and a padlock probe are added to the sample. Each end of the splinter contains a complementary sequence that is specific to each adapter. The splinter therefore forms a bridge between the adapters of the distinct antibodies if they are found in tight proximity. These bridges are stabilized by the padlock probe that forms a more stable circular product between the adapters. Consecutive and carefully designed washes eliminate partially bound splinters leaving only splinters bound at both ends. Given that there is only one splinter present at each interaction, and that the splinter is located at the site of interaction, the technique provides a quantitative measure of interaction with the localization resolution of FISH but conditionally bound to the proximity restriction of ISPLA.

The unique characteristics of ChrISP enable it to measure new aspects of chromatin architecture. In the paper we employ ChrISP to score for the 3D compaction of repeat elements found in Cot1 DNA as well as unique sequences within chromosome 11 through the use of a

chromosome 11 (CT11) territory probe. In the case of CT11 territory probe, the technique visualized chromatin compaction in the vicinity of the nuclear membrane. An extensive array of negative controls along with the sequencing of the both the Cot1 and CT11 probes validate this interesting result proving that technique can produce new insights into chromatin organization. A follow-up application of ChRISP has documented that these compact structures are in proximity to H3K9me2 histone modifications, which likely represent the large organized chromatin K9 modifications previously detected in the epigenome by ChIP within LADs [129]. Recently chromatin compaction itself has been linked to peripheral localization as well as gene repression [130,131], underscoring the relevance of the visualization of compact, repressed structures in single cells.

3.3 PAPER III: FOCUSED, LOW COST, HIGH SENSITIVITY, ROBUST ASSESSMENT OF FUNCTIONAL INTERACTIONS WITH NODEWALK

In this paper we describe Nodewalk, a new chromatin conformation capture technique. Nodewalk is a low cost, many-to-all chromatin conformation capture assay aimed at high resolution/sensitivity inspection of specific loci. As mentioned earlier, existing WG3C and Cap3C techniques are able to assay large cell populations [53] at varying levels of resolution typically employing millions of cells and large/expensive sequencing coverage to assay large numbers of loci in a single experiment [53,82,85,96]. In this paper we propose an alternative way, one that works in small/inexpensive iterations and focuses on interactomes impinging on single loci.

Briefly, the technique starts with optimized steps of crosslinking, digestion and ligation presented in the original 3C protocol [90,91]. First, the chromatin is crosslinked with freshly prepared monomeric formaldehyde instead of formalin. The use of monomeric formaldehyde enables the selective crosslinking of DNA/protein targets that are in direct physical contact, as opposed to the proximal crosslinking generated by longer formaldehyde chains forming in formalin over time [132]. Second, the sample is digested by restriction enzymes to reach at least 90% digestion efficiency at the 5` and 3`ends of the baits. The digested chromatin is then mixed with a negative control, digested Drosophila chromatin in order to assess the spurious (or random) ligation rate by scoring for human-Drosophila chimeras. This step was inspired by the ChiaPet assay`s A/B primer strategy [133]. The sample undergoes 3C ligation at ultra low concentration enabled by the low input required by the technique. The complexes are then reverse crosslinked and DNA is extracted. Next, the DNA is digested using the Nextera transposon system. This step not only reduces the size of the template to insert sizes compatible with the Illumina cluster amplification, but also inserts a molecular identifier in the form of the pseudo-random digestion site introduced by the transposon. Next, instead of the standard adapters, a custom adapter containing a T7 motif is introduced to each fragment. The sample is then linearly amplified by *in vitro* RNA transcription using the T7 polymerase. This key step produces large amounts of template while maintaining the original input ratios between the different templates in the sample. Hence, this provides the basis for the resampling aspect of the technique and enables the detection of chromatin fibre interactions in small cell populations. The template is then selected using specially designed probes[4], which assay regions of interest. The probes anneal to their target fragments and are then extended by reverse transcription.

[4] To assist in the design of the primers, a Nodewalk primer helper tool was added to the ChipSeqNav tool (http://www.chipseqtools.org/chipseqnav.html). The tool screens the vicinity of a restriction site for uniquely mapped primers with an annealing temperature matching a user defined range and produces a list of primer candidates that can be then manually inspected to select the optimal design. The tool also overlays the restriction site above a 800+ chipseq display which provides the user input over the possible binding sites and regulatory elements covering a given restriction site. The context display of the restriction site along with the automatic "uniqueness" screening significantly reduces the time required to design Nodewalk probes.

Lastly, the cDNA template is exponentially amplified using standard primers to produce vast libraries of interactor sequences.

The resulting protocol provides novel functionality that differentiate it from existing technologies. The RNA step and PCR amplification step provide a capture enrichment of 500,000x, which makes it possible to probe multiple loci in a single MiSeq run. Even the latest and most efficient Cap3C technique, HiCap, requires a high throughput HiSeq run [96], thereby severely limiting the number of groups able to afford this technique and run it on a daily basis. The RNA step also provides a resampling ability that is not available in any existing Cap3C techniques. The large amounts of linearly amplified RNA template can be produced in such quantities as to allow multiple re-queries of the same source material. Importantly, this enables the Nodewalk to query a range of loci for unknown interacting sequences and then iteratively query the neighbors of the neighbors on the same source template, making this technique unique among other existing "C" techniques. This feature provides ways to validate interactions and to explore the entire interactome of a functional element without examining the entire set of possible interactions. Furthermore, the efficiency of the RNA step makes this technique suitable for the analysis of ultra low input, such as patient derived FACS-sorted cells. Finally, the low input enables the ligation step to be performed at ultra-low concentrations which both minimize spurious ligation events and are even able to quantitate per cell interaction frequencies.

Using this technique we investigated the interacting partners of the *c-MYC* promoter locus in a human colon cancer cell line (HCT116). The choice of this particular locus was driven by the previous studies that have reported valid and interesting interactions between the locus and a disease associated point mutation in colorectal cancer [134]. These known functional interactions were well documented and therefore provided positive controls, as recent studies have also shown [94]. By applying the technique we were not only able to recapitulate known interactions, but by iteratively querying the newly found interactors we also rediscovered the original bait providing a 2 fold validation of the method. Furthermore, the resulting network after a single iteration clearly showed an enrichment of enhancers both in *cis* and in *trans*. The interacting regions included active, poised and super-enhancers suggesting a large network of enhancer "factories" involving multiple chromosomes, reminiscent of the previously described active chromatin hub concept [135].

Given *c-MYC's* well documented role in colon cancer, we then assayed human primary colonic epithelial cells (HCEC) to study the difference between the 3D interactome of cancer cells and primary cells. While the *c-MYC* locus in HCEC cells displayed a similar enrichment of enhancer interactors both in *cis* and in *trans*, the variety of enhancers was more restricted than that in HCT116 cells. Using K-Core analysis to dissect the structure of the chromatin hubs, we found that hubs with high k-core value are highly enriched in enhancer chromatin marks in the cancer state, but not in the primary cells. Furthermore, these chromatin hubs are usually found in the proximity (on the linear chromosome) of genes with higher expression, than the rest of the network.

Finally, we demonstrate that even ultra-low input sources (i.e. less, than 200 cells) can be assayed with the technique. The results of these experiments showed not only high concordance between the different samples, but also revealed the transient nature of certain interactions that are present only in some of the low-input samples.

In summary, the Nodewalk represents a middle ground between the WG3C/Cap3C experiments and the 4C one-to-many approach by introducing an iterative low-cost high sensitivity strategy. Moreover, it also includes a critical built in Drosophila negative control for spurious ligation events which have been largely ignored by published work. The unique characteristics of the Nodewalk technique make it the ideal choice for research focused on few specific regulatory elements or is restricted to limited input sources.

3.4 PAPER IV: NODEWALK ANALYSIS PIPELINE SOFTWARE SUITE

The Nodewalk technique described in Paper III provides a powerful and practical insight into focused interactomes. To accelerate discovery of these powerful datasets we designed the Nodewalk Analysis Pipeline (NAP). In the emerging field of 4C and WG3C large number of tools have recently become available including preprocessing pipelines [100,104,107,136], bias correction [98,99,102] and visualization [82,105,137–139] tailored to the specific characteristics of the underlying data. The NAP suite aims to provide the functionality available in some of these tools but tailored for Nodewalk datasets through a visual and user friendly interface.

The suite provides a comprehensive set of tools which includes preprocessing, analysis and post-processing. The preprocessing modules include the NodewalkMapper interface that allows users to upload fastq files that are then mapped and summarized. The general QC reports produced at this stage are made available through the NodewalkStats interface. The preprocessing stage maps, filters and summarizes the library. The filtering is a multistep process where each read pair is validated as a mis-annealing event or a valid ligation event. This stage provides information about the library efficiency and the false ligation rate as measured by the number of human-Drosophila ligation events. Critically this stage produces both the key measurements and the summarization strategies that feed into the upstream analysis tools.

The preprocessing module produces 4 different summarization strategies: by restriction fragment, fixed binning, averaged window and by regulatory element. Each of these methods provide different strengths. While restriction fragments provide the maximum resolution for 3C technologies[5], the fixed binning approach commonly used in HiC datasets is very useful when comparing between different datasets. Alternatively, averaged windows allow combining information from different fragments to enhance the significance. In the NAP we introduce the support for variable region length bed files, which allows users to determine the proximity between functional regions (Enhancers, CTCF binding sites, Promotors, etc, etc).

The summarization strategy is performed on 3 different key parameter indicators. The first parameter, ctTot, is the total number of valid reads overlapping an interacting region. Although the ctTot performs well as a ligation event proxy, it is known that the cDNA amplification round performed at the end of the protocol introduces several biases that can affect its efficacy in quantifying interaction frequency. To compensate for the amplification bias, the number of distinct restriction sites is reported as the ctPos. As mentioned earlier, as the 3C template is digested with the transposon system, the near - random digestion of the transposon in the Nextera system provides a molecular identifier, because each fragment is

[5] New protocols involving DNase rather than restriction enzymes could theoretically provide point of contact resolution[140]

theoretically digested at a different site. As shown in the Nodewalk manuscript, the ctPos provides an even better approximation of the original ligation event count, but it is also limited. Firstly, we observed that the digestion by the Nextera system had preferences to digest fragments at specific sites and furthermore the number of distinct digestion sites was limited by the length of the fragment. Therefore, smaller fragments have lower chance of producing high ctPos values. To address this, we introduced a degenerate barcode sequence instead of the barcode in the adapters introduced at the very first stages of the protocol, as described previously [141,142]. This sequence acts as a unique molecular identifier (UMI). The reported ctUMI is reported as the number of distinct sequences for all reads mapping to the same restriction fragment thereby effectively reducing PCR amplification bias.

Next the analysis tools provide 3 different applications aimed at analyzing the reproducibility / variation between samples (3CRepro), annotating and exporting Nodewalk datasets to different formats (3CAnnotate) and visualizing the coverage of each interaction (3CCov). Finally the post processing tool (3CEnrich) enables users to run 4 different commonly used enrichment assays using a Nodewalk-specific background model. When deployed in a high-end server, this interface provides a responsive enrichment analysis tool, which users can query in real-time enabling them to fine tune search parameters to optimize enrichment values.

In summary the NAP suite of tools provide user friendly set of tools to perform the most common operations involved in the analysis of Nodewalk data: mapping, filtering, summarization, visualization and enrichment. The tool enables non-technical users to execute a complex pipeline in high-end servers and to use the computational power of these servers to interactively fine tune enrichment parameters, something that is not possible in web-based public tools. The NAP suite provides cradle to publication analysis for Nodewalk datasets.

4 SUMMARY / OUTLOOK

In the first paper we describe the division of labor between the CTCF binding sites within a regulatory element, contributing to the concept that "role-casting" every CTCF as insulator is incorrect. Thus, we proposed that the functional outcome of CTCF-defined chromatin conformation is as varied as the conformations themselves. In the second paper we describe a new microscopy based tool designed to assess nuclear conformation at a single cell level. The third and fourth papers describe a novel sequencing based technique designed to further query the 3D neighborhood of functional elements in small cell populations. The work summarized here unveils the importance of a chromatin conformation regulator and presents new techniques to expand the analysis of this phenomena in light of the limitations of existing techniques. The analysis of the functional and topological features of the nuclear architecture are slowly beginning to unfold, emerging techniques provide further opportunities to combine them to further explore this largely unknown and interesting field.

4.1 NODEWALK V2 / 4C, UNLIMITED TEMPLATE THROUGH BIOTINYLATED PRIMERS

In paper III we describe the Nodewalk protocol, and it its ability to resample the same source template in order to iteratively uncover the network of interactors by querying the neighbors of the neighbors (walking the nodes). This feature relies on the large amount of RNA template produced by the T7 RNA Polymerase. Despite the large amount of template that is generated, it is a fixed amount which must be carefully stored due to RNA's instability. A further improvement of the technique would be made possible by using biotinylated primers at the initial Nextera amplification step. This would essentially create chimeric DNA templates (ligation products) that can be easily purified and covalently attached to beads or a surface. This would allow not only to skip the Dnase treatment step, but as the DNA template would be physically attached to a surface it would make it trivial to generate RNA template, extract the newly created RNA and store the DNA template for future use.

In paper I we describe the different interactomes of the CTCF binding sites within the *H19* ICR by using the 4C technique. Upon close inspection, the 4C relies on a high number of PCR cycles in order to essentially dilute out the background. This severely limits the quantitative aspect of the 4C and introduces limitations over which fragments can be amplified by the technique (as large fragments amplify slowly). As in the Nodewalk, a possible improvement of the 4C technique would be to replace the high number of PCR cycles with the incorporation of biotinylated primers that can be used to purify out the target product between the different nested PCR steps. Additionally, another improvement would come from designing 4C primers which keep a 20-50 bp distance from the restriction site in order to differentiate mis-annealing sites from proper interactors.

4.2 BEYOND NODEWALK, LIGATION-LESS 3C

Despite the large number of techniques currently available, all 3C based techniques depend on the ligation of 2 strands to create a chimeric DNA fragment that can be sequenced and mapped. This requirement creates a great constraint in the methods, as not all 3C products ligate in the desired configuration. In many cases the fragments will ligate to themselves creating a self-ligation product, or will not ligate at all. Furthermore the restriction site distribution, chromatin digestibility, GC content, fragment size and nuclear architecture in general produce large biases in the data, which are difficult to quantify and normalize [100]. Perhaps the biggest limitation is in the detection of chromatin hub complexes containing more than 2 fragments/interactors. Except for the 4C, where these hubs can be captured in a circular DNA molecule, these configurations are impossible to detect in all other 3C techniques.

As an alternative, I propose a different strategy which label each end of a complex with a positional barcode. This strategy increases the probability of detecting an interaction by using both ends of each complex and excluding the ligation step all together, thereby allowing the detection of N-fragment complexes. Briefly, cross linking and digestion are done as in any traditional 3C experiment, but instead of ligating each end to each other, the complexes are isolated via different methods. In each of these isolation "chambers" each complex is tagged with the same unique molecular identifier [141,142] (UMI) per complex.

This strategy inherently depends on the ability of isolating a single complex in a constrained volume along with monoclonally amplified UMI containing adapter. This would sound like a formidable challenge, had it not been already resolved in 3 different occasions by the 3 founding NGS technologies. For example, complexes could be covalently bound to surface of a slide which would then be put through the same photolithographic process that created the microarrays [143]. Using photolithographic chamber coupled with a flow cell in which photosensitive capped nucleotides are flooded, it would be possible to "grow" the same positional UMI on any open end fragment of the complex. Alternatively, the emulsion polymerase chain reaction (emPCR) used in the first 454 sequencing [144] technologies could also serve as isolated chambers where the water droplets isolated in the oil emulsion could be seeded with monoclonally amplified UMIs, which would then be mixed at low concentrations with single complexes. Finally, another opportunity arises from the cluster amplification [145] used by the Illumina platform to create the sequencing clusters. Here again it would be possible to create monoclonally amplified clusters of UMI barcoded adapters, which are constrained by a small radius in the flow cell. Similar strategies [146] are currently being applied to mRNA in the emerging field of spatial transcriptomics [147], but this has yet to be applied in the nuclear architecture field.

5 ACKNOWLEDGEMENTS

I would like to thank both the High-throughput Epigenetic Regulatory Organization In Chromatin (HEROIC) and the Knut and Alice Wallenberg Foundation (KAW) for funding my doctoral studies. I would also like to thank the Science and Technology and Telecommunications Ministry of Costa Rica (MICIT) as well as the National Council for Scientific and Technological Research of Costa Rica (CONICIT) for providing a partial scholarship for the master`s program, which allowed me to start my doctoral education.

Next I would like to thank Anita Göndör and Rolf Ohlsson for their support and guidance, and overall for helping a humble computer scientist make the difficult leap into the amazing field of biological information systems. This feat is a testament to their passion and overwhelmingly, their patience in guiding a logical mindset into a fuzzy, stochastic, pleiotropic and dynamic field. I would like to thank also Noriyuki Sumida for letting me into the world of wet lab protocols and introducing me into this field. Sensei Nori, the pipet guru, introduced me to a world where very few bioinformaticians have gone before, a world where kits fail and biology happens. This insight into what can go wrong in the wet lab gave me a perspective without which it is not possible to do bioinformatics. Moreover, he showed me what a perfectly executed ChIP-Seq/RNASeq/GroSeq/Nodewalk should look like, Nori you've set a tall watermark, I don't think I will find datasets as clean as yours. I would also like to thank Anna, Feri, Honglei, Marta, Samer, Xinqi, Chengxi, Maria, Manos, Lluis, Carolina, Li-Sophie, G, Habib for their friendship, support and fruitful discussions through the rough years.

To Erik Aurell, Erik Fredlund, Gema Sanz and David Gomes I would like to thank for providing guidance in statistics and bioinformatics. Working in complete isolation as the only bioinformatician at MTC for many years, your input and collaboration made it possible for me to advance in the field in parallel to my development in Biology and BioTechnology. To Margareta Wilhelm and Galina Selivanova I would like to thank you for introducing me to new fields in cancer research, expanding my vision beyond my specialization in nuclear chromatin architecture.

I would like to thank my wife, Bita, who stood at my side through innumerable failures. To my mother, Gabriela Woodbridge, who taught me the meaning of the word resilience and the unquestionable priority of moral principles in every aspect of one's life. There is no pride nor value in accomplishing a goal, personal or scientific, if one's principles were compromised on the road to achieve it. To my stepfather Manrique and my brother Federico, who gave me the confidence to leave a cosy well paid industry job in warm Costa Rica to come face a huge challenge in icy Sweden as a student. To the rest of my family for standing by and cheering me on through times of doubt and weakness. I would like to thank the country of Sweden, for giving me an experience far more significant and far greater than any impact factor or any number of publications. Finally, I would like to thank those that I lost to cancer, my father Ricardo Fernández, my father in law Ebrahim Daemi, my friends John Grieves and Arturo Rosabal because it was their memory and everlasting presence that drove me here.

6 REFERENCES

1. Wang D, Gribskov M. Examining the architecture of cellular computing through a comparative study with a computer. J R Soc Interface. 2005;2: 187–195. doi:10.1098/rsif.2005.0038

2. Darwin C. On the Origin of the Species by Natural Selection [Internet]. Murray; 1859. Available: http://www.citeulike.org/group/1788/article/1262916

3. Soyfer VN. The consequences of political dictatorship for Russian science. Nat Rev Genet. 2001;2: 723–729. doi:10.1038/35088598

4. Watson JD, Crick FHC, Others. Molecular structure of nucleic acids. Nature. 1953;171: 737–738. Available: http://www.nature.com/physics/looking-back/crick/

5. Franklin RE, Gosling RG. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature. 1953;172: 156–157. Available: http://www.ncbi.nlm.nih.gov/pubmed/13072614

6. Felsenfeld G. A brief history of epigenetics. Cold Spring Harb Perspect Biol. 2014;6. doi:10.1101/cshperspect.a018200

7. Holliday R. Epigenetics: a historical overview. Epigenetics. 2006;1: 76–80. Available: http://www.ncbi.nlm.nih.gov/pubmed/17998809

8. Imhof A, Bonaldi T. “Chromatomics” the analysis of the chromatome. Mol Biosyst. 2005;1: 112–116. doi:10.1039/b502845k

9. Deans C, Maggert KA. What do you mean, “epigenetic”? Genetics. 2015;199: 887–896. doi:10.1534/genetics.114.173492

10. Laurent M, Charvin G, Guespin-Michel J. Bistability and hysteresis in epigenetic regulation of the lactose operon. Since Delbrück, a long series of ignored models. Cell Mol Biol . 2005;51: 583–594. Available: http://www.ncbi.nlm.nih.gov/pubmed/16359608

11. Beisson J, Sonneborn TM. CYTOPLASMIC INHERITANCE OF THE ORGANIZATION OF THE CELL CORTEX IN PARAMECIUM AURELIA. Proc Natl Acad Sci U S A. 1965;53: 275–282. Available: http://www.ncbi.nlm.nih.gov/pubmed/14294056

12. Kunkel TA. DNA replication fidelity. J Biol Chem. 2004;279: 16895–16898. doi:10.1074/jbc.R400006200

13. Gardner RL, Lyon MF. Biological Sciences: X Chromosome Inactivation studied by Injection of a Single Cell into the Mouse Blastocyst. Nature Publishing Group; 1971;231: 385–386. doi:10.1038/231385a0

14. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet. 1975;14: 9–25. Available: http://www.ncbi.nlm.nih.gov/pubmed/1093816

15. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. Science. 1975;187: 226–232. Available: http://www.ncbi.nlm.nih.gov/pubmed/1111098

16. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38: 23–38. doi:10.1038/npp.2012.112

17. Olins AL, Olins DE. Spheroid chromatin units (v bodies). Science. sciencemag.org; 1974;183: 330–332. Available: http://www.ncbi.nlm.nih.gov/pubmed/4128918

18. Finch JT, Klug A. Solenoidal model for superstructure in chromatin. Proc Natl Acad Sci U S A. National Acad Sciences; 1976;73: 1897–1901. Available: http://www.ncbi.nlm.nih.gov/pubmed/1064861

19. Histone Modification Table | CST Cell Signaling Technology [Internet]. [cited 22 Aug 2015]. Available: http://www.cellsignal.com/common/content/content.jsp?id=science-tables-histone

20. An encyclopedia of DNA elements [Internet]. Available: https://www.encodeproject.org/

21. Roadmap Epigenomics Project [Internet]. Available: http://www.roadmapepigenomics.org/

22. Xu H, Papatsenko D, Ma'ayan A, Lemischka I. Chapter 22 - Biological and Quantitative Models for Stem Cell Self-Renewal and Differentiation. In: Dekker AJMWV, editor. Handbook of Systems Biology. San Diego: Academic Press; 2013. pp. 427–441. doi:10.1016/B978-0-12-385944-0.00022-8

23. Bygren LO, Tinghög P, Carstensen J, Edvinsson S, Kaati G, Pembrey ME, et al. Change in paternal grandmothers' early food supply influenced cardiovascular mortality of the female grandchildren. BMC Genet. 2014;15: 12. doi:10.1186/1471-2156-15-12

24. Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. Cell. 2014;157: 95–109. doi:10.1016/j.cell.2014.02.045

25. Whitelaw E. Disputing Lamarckian epigenetic inheritance in mammals. Genome Biol. 2015;16: 60. doi:10.1186/s13059-015-0626-0

26. Kaati G, Bygren LO, Edvinsson S. Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. Eur J Hum Genet. 2002;10: 682–688. doi:10.1038/sj.ejhg.5200859

27. Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol Cell. 2011;42: 451–464. doi:10.1016/j.molcel.2011.04.005

28. Holmgren C, Kanduri C, Dell G, Ward A, Mukhopadhya R, Kanduri M, et al. CpG methylation regulates the Igf2/H19 insulator. Curr Biol. 2001;11: 1128–1130. Available: http://www.ncbi.nlm.nih.gov/pubmed/11509237

29. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol. 1987;196: 261–282. Available: http://www.ncbi.nlm.nih.gov/pubmed/3656447

30. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517: 321–326. doi:10.1038/nature14192

31. Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zeliger SR, et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature. 2014;513: 115–119. doi:10.1038/nature13458

32. Kearns M, Preis J, McDonald M, Morris C, Whitelaw E. Complex patterns of inheritance of an imprinted murine transgene suggest incomplete germline erasure. Nucleic Acids Res. 2000;28: 3301–3309. Available: http://www.ncbi.nlm.nih.gov/pubmed/10954598

33. Kierszenbaum AL. Genomic imprinting and epigenetic reprogramming: unearthing the garden of forking paths. Mol Reprod Dev. 2002;63: 269–272. doi:10.1002/mrd.90011

34. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006;38: 1341–1347. doi:10.1038/ng1891

35. Tilghman SM, Bartolomei MS, Webber AL, Brunkow ME, Saam J, Leighton PA, et al. Parental imprinting of the H19 and Igf2 genes in the mouse. Cold Spring Harb Symp Quant Biol. 1993;58: 287–295. Available: http://www.ncbi.nlm.nih.gov/pubmed/7956041

36. Haig D. Genetic conflicts in human pregnancy. Q Rev Biol. 1993;68: 495–532. Available: http://www.ncbi.nlm.nih.gov/pubmed/8115596

37. Bergman D, Halje M, Nordin M, Engström W. Insulin-like growth factor 2 in development and disease: a mini-review. Gerontology. 2013;59: 240–249. doi:10.1159/000343995

38. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Chromosomal DNA and Its Packaging in the Chromatin Fiber [Internet]. Garland Science; 2002. Available: http://www.ncbi.nlm.nih.gov/books/NBK26834/

39. Harshman SW, Young NL, Parthun MR, Freitas MA. H1 histones: current perspectives and challenges. Nucleic Acids Res. 2013;41: 9593–9609. doi:10.1093/nar/gkt700

40. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21: 381–395. doi:10.1038/cr.2011.22

41. Falkenberg KJ, Johnstone RW. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. Nat Rev Drug Discov. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;13: 673–691. doi:10.1038/nrd4360

42. Audergon PNCB, Catania S, Kagansky A, Tong P, Shukla M, Pidoux AL, et al. Restricted epigenetic inheritance of H3K9 methylation. Science. 2015;348: 132–135. doi:10.1126/science.1260638

43. Campos EI, Stafford JM, Reinberg D. Epigenetic inheritance: histone bookmarks across generations. Trends Cell Biol. 2014;24: 664–674. doi:10.1016/j.tcb.2014.08.004

44. Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic information. Nat Rev Genet. Nature Publishing Group; 2010;11: 285–296. doi:10.1038/nrg2752

45. Kanduri M, Kanduri C, Mariano P, Vostrov AA, Quitschke W, Lobanenkov V, et al. Multiple nucleosome positioning sites regulate the CTCF-mediated insulator function of the H19 imprinting control region. Mol Cell Biol. 2002;22: 3339–3344. Available: http://www.ncbi.nlm.nih.gov/pubmed/11971967

46. Lee C-H, Wu J, Li B. Chromatin remodelers fine-tune H3K36me-directed deacetylation of neighbor nucleosomes by Rpd3S. Mol Cell. 2013;52: 255–263. doi:10.1016/j.molcel.2013.08.024

47. Pina C, Fugazza C, Tipping AJ, Brown J, Soneji S, Teles J, et al. Inferring rules of lineage commitment in haematopoiesis. Nat Cell Biol. 2012;14: 287–294. doi:10.1038/ncb2442

48. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res. 2012;22: 1735–1747. doi:10.1101/gr.136366.111

49. Venkatesh S, Workman JL. Set2 mediated H3 lysine 36 methylation: regulation of transcription elongation and implications in organismal development. Wiley Interdiscip Rev Dev Biol. 2013;2: 685–700. doi:10.1002/wdev.109

50. Göndör A, Ohlsson R. Replication timing and epigenetic reprogramming of gene expression: a two-way relationship? Nat Rev Genet. 2009;10: 269–276. doi:10.1038/nrg2555

51. Wigler MH, Axel R. Nucleosomes in metaphase chromosomes. Nucleic Acids Res. 1976;3: 1463–1471. Available: http://www.ncbi.nlm.nih.gov/pubmed/958895

52. Kireev I, Lakonishok M, Liu W, Joshi VN, Powell R, Belmont AS. In vivo immunogold labeling confirms large-scale chromatin folding motifs. Nat Methods. 2008;5: 311–313. doi:10.1038/nmeth.1196

53. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326: 289–293. doi:10.1126/science.1181369

54. Cremer T, Cremer M. Chromosome territories. Cold Spring Harb Perspect Biol. 2010;2: a003889. doi:10.1101/cshperspect.a003889

55. Branco MR, Pombo A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. 2006;4: e138. doi:10.1371/journal.pbio.0040138

56. Kaufmann R, Müller P, Hildenbrand G, Hausmann M, Cremer C. Analysis of Her2/neu membrane protein clusters in different types of breast cancer cells using localization microscopy. J Microsc. 2011;242: 46–54. doi:10.1111/j.1365-2818.2010.03436.x

57. Passarge E. Emil Heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. Am J Hum Genet. 1979;31: 106–115. Available: http://www.ncbi.nlm.nih.gov/pubmed/377956

58. Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S. Chromosome territories--a functional nuclear landscape. Curr Opin Cell Biol. 2006;18: 307–316. doi:10.1016/j.ceb.2006.04.007

59. Reddy KL, Feinberg AP. Higher order chromatin organization in cancer. Semin Cancer Biol. 2013;23: 109–115. doi:10.1016/j.semcancer.2012.12.001

60. Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B. Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat Genet. 2006;38: 1005–1014. doi:10.1038/ng1852

61. Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, et al. Single-cell dynamics of genome-nuclear lamina interactions. Cell. 2013;153: 178–192. doi:10.1016/j.cell.2013.02.028

62. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453: 948–951. doi:10.1038/nature06947

63. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SWM, Solovei I, Brugman W, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell. 2010;38: 603–613. doi:10.1016/j.molcel.2010.03.016

64. Meuleman W, Peric-Hupkes D, Kind J, Beaudry J-B, Pagie L, Kellis M, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 2013;23: 270–280. doi:10.1101/gr.141028.112

65. Gerace L, Blobel G. The nuclear envelope lamina is reversibly depolymerized during mitosis. Cell. 1980;19: 277–287. Available: http://www.ncbi.nlm.nih.gov/pubmed/7357605

66. Gruenbaum Y, Margalit A, Goldman RD, Shumaker DK, Wilson KL. The nuclear lamina comes of age. Nat Rev Mol Cell Biol. 2005;6: 21–31. doi:10.1038/nrm1550

67. Zuleger N, Boyle S, Kelly DA, Las Heras JI de, Lazou V, Korfali N, et al. Specific nuclear envelope transmembrane proteins can promote the location of chromosomes to and from the nuclear periphery. Genome Biol. BioMed Central Ltd; 2013;14: R14. Available: http://genomebiology.com/2013/14/2/R14/

68. Harr JC, Luperchio TR, Wong X, Cohen E, Wheelan SJ, Reddy KL. Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. J Cell Biol. 2015;208: 33–52. doi:10.1083/jcb.201405110

69. Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet. 2001;10: 211–219. Available: http://www.ncbi.nlm.nih.gov/pubmed/11159939

70. Paddy MR, Belmont AS, Saumweber H, Agard DA, Sedat JW. Interphase nuclear envelope lamins form a discontinuous network that interacts with only a fraction of the chromatin in the nuclear periphery. Cell. 1990;62: 89–106. doi:10.1016/0092-8674(90)90243-8

71. Taniura H, Glass C, Gerace L. A chromatin binding site in the tail domain of nuclear lamins that interacts with core histones. J Cell Biol. 1995;131: 33–44. Available: http://www.ncbi.nlm.nih.gov/pubmed/7559784

72. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15: 272–286. doi:10.1038/nrg3682

73. Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature. 2012;482: 221–225. doi:10.1038/nature10805

74. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010;107: 21931–21936. doi:10.1073/pnas.1016071107

75. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature. 2013;494: 497–501. doi:10.1038/nature11884

76. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell. 2015;162: 900–910. doi:10.1016/j.cell.2015.07.038

77. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012;149: 1233–1244. doi:10.1016/j.cell.2012.03.051

78. Hou C, Zhao H, Tanimoto K, Dean A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc Natl Acad Sci U S A. 2008;105: 20398–20403. doi:10.1073/pnas.0808506106

79. Göndör A, Ohlsson R. Chromatin insulators and cohesins. EMBO Rep. 2008;9: 327–329. doi:10.1038/embor.2008.46

80. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153: 307–319. doi:10.1016/j.cell.2013.03.035

81. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, et al. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Mol Cell. 2015;58: 362–370. doi:10.1016/j.molcel.2015.02.014

82. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159: 1665–1680. doi:10.1016/j.cell.2014.11.021

83. Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 2015;162: 108–119. doi:10.1016/j.cell.2015.05.048

84. Zhang Y, McCord RP, Ho Y-J, Lajoie BR, Hildebrand DG, Simon AC, et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell. 2012;148: 908–921. doi:10.1016/j.cell.2012.02.002

85. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502: 59–64. doi:10.1038/nature12593

86. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012;22: 490–503. doi:10.1038/cr.2012.15

87. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet. 2011;43: 630–638. doi:10.1038/ng.857

88. Bickmore WA. Chromatin Structure and Domains. eLS. John Wiley & Sons, Ltd; 2001. doi:10.1038/npg.els.0005279

89. Maglione M, Sigrist SJ. Seeing the forest tree by tree: super-resolution light microscopy meets the neurosciences. Nat Neurosci. Nature Publishing Group; 2013;16: 790–797. doi:10.1038/nn.3403

90. Miele A, Dekker J. Mapping cis- and trans- chromatin interaction networks using chromosome conformation capture (3C). Methods Mol Biol. 2009;464: 105–121. doi:10.1007/978-1-60327-461-6_7

91. Miele A, Gheldof N, Tabuchi TM, Dostie J, Dekker J. Mapping chromatin interactions by chromosome conformation capture. Curr Protoc Mol Biol. 2006;Chapter 21: Unit 21.11. doi:10.1002/0471142727.mb2111s74

92. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006;38: 1348–1354. doi:10.1038/ng1896

93. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2007;2: 988–1002. doi:10.1038/nprot.2007.116

94. Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun. 2015;6: 6178. doi:10.1038/ncomms7178

95. Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet. 2014;46: 205–212. doi:10.1038/ng.2871

96. Sahlén P, Abdullayev I, Ramsköld D, Matskova L, Rilakovic N, Lötstedt B, et al. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol. 2015;16: 156. doi:10.1186/s13059-015-0727-9

97. Schmid MW, Grob S, Grossniklaus U. HiCdat: a fast and easy-to-use Hi-C data analysis tool. BMC Bioinformatics. 2015;16: 277. doi:10.1186/s12859-015-0678-x

98. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28: 3131–3133. doi:10.1093/bioinformatics/bts570

99. Li W, Gong K, Li Q, Alber F, Zhou XJ. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. Bioinformatics. 2015;31: 960–962. doi:10.1093/bioinformatics/btu747

100. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43: 1059–1065. doi:10.1038/ng.947

101. Hwang Y-C, Lin C-F, Valladares O, Malamon J, Kuksa PP, Zheng Q, et al. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics. 2015;31: 1290–1292. doi:10.1093/bioinformatics/btu801

102. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9: 999–1003. doi:10.1038/nmeth.2148

103. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res. 2013;41: e132. doi:10.1093/nar/gkt373

104. Van de Werken HJG, Landan G, Holwerda SJB, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods. 2012;9: 969–972. doi:10.1038/nmeth.2173

105. Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, Hovig E. HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. Bioinformatics. 2014;30: 1620–1622. doi:10.1093/bioinformatics/btu082

106. Williams RL Jr, Starmer J, Mugford JW, Calabrese JM, Mieczkowski P, Yee D, et al. fourSig: a method for determining chromosomal interactions in 4C-Seq data. Nucleic Acids Res. 2014;42: e68. doi:10.1093/nar/gku156

107. Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EEM, Huber W. FourCSeq: analysis of 4C sequencing data. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv335

108. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

109. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18: 1851–1858. doi:10.1101/gr.078212.108

110. Kent WJ. BLAT −− The BLAST-Like Alignment Tool BLAT — The BLAST-Like Alignment Tool. Genome Res. 2002; 656–664. doi:10.1101/gr.229202

111. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

112. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. 2010;Chapter 11: Unit 11.7. doi:10.1002/0471250953.bi1107s32

113. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

114. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

115. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7: 562–578. doi:10.1038/nprot.2012.016

116. Anders S. HTSeq: Analysing high-throughput sequencing data with Python. URL http://www-huber embl de/users/anders/HTSeq/doc/overview html. 2010;

117. Anders S, Huber W. Differential expression of RNA-Seq data at the gene level--the DESeq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL). 2012; Available: http://watson.nci.nih.gov/bioc_mirror/packages/2.11/bioc/vignettes/DESeq/inst/doc/DESeq.pdf

118. Yan H, Evans J, Kalmbach M, Moore R, Middha S, Luban S, et al. HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. BMC Bioinformatics. 2014;15: 280. doi:10.1186/1471-2105-15-280

119. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012;7: 1728–1740. doi:10.1038/nprot.2012.101

120. Castellano G, Le Dily F, Pulido AH, Beato M, Roma G. Hi-Cpipe: a pipeline for high-throughput chromosome capture.

121. Peng C, Fu L-Y, Dong P-F, Deng Z-L, Li J-X, Wang X-T, et al. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. Nucleic Acids Res. 2013;41: e183. doi:10.1093/nar/gkt745

122. Phanstiel DH, Boyle AP, Heidari N, Snyder MP. Mango: a bias-correcting ChIA-PET analysis pipeline. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv336

123. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137: 1194–1211. doi:10.1016/j.cell.2009.06.001

124. Bergström R, Whitehead J, Kurukuti S, Ohlsson R. CTCF regulates asynchronous replication of the imprinted H19/Igf2 domain. Cell Cycle. 2007;6: 450–454. Available: http://www.ncbi.nlm.nih.gov/pubmed/17329968

125. Pant V, Kurukuti S, Pugacheva E, Shamsuddin S, Mariano P, Renkawitz R, et al. Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. Mol Cell Biol. 2004;24: 3497–3504. Available: http://www.ncbi.nlm.nih.gov/pubmed/15060168

126. Pugacheva EM, Tiwari VK, Abdullaev Z, Vostrov AA, Flanagan PT, Quitschke WW, et al. Familial cases of point mutations in the XIST promoter reveal a correlation between CTCF binding and pre-emptive choices of X chromosome inactivation. Hum Mol Genet. 2005;14: 953–965. doi:10.1093/hmg/ddi089

127. McAdams HH, Adam A. It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 1999;15: 65–69. doi:10.1016/s0168-9525(98)01659-x

128. Söderberg O, Gullberg M, Jarvius M, Ridderstråle K, Leuchowius K-J, Jarvius J, et al. Direct observation of individual endogenous protein complexes in situ by proximity ligation. Nat Methods. 2006;3: 995–1000. doi:10.1038/nmeth947

129. Chen X, Yammine S, Shi C, Tark-Dame M, Göndör A, Ohlsson R. The visualization of large organized chromatin domains enriched in the H3K9me2 mark within a single

chromosome in a single cell. Epigenetics. 2014;9: 1439–1445. doi:10.4161/15592294.2014.971633

130. Therizols P, Illingworth RS, Courilleau C, Boyle S, Wood AJ, Bickmore WA. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science. 2014;346: 1238–1242. doi:10.1126/science.1259587

131. Vallot C, Hérault A, Boyle S, Bickmore WA, Radvanyi F. PRC2-independent chromatin compaction and transcriptional repression in cancer. Oncogene. 2015;34: 741–751. doi:10.1038/onc.2013.604

132. Göndör A, Rougier C, Ohlsson R. High-resolution circular chromosome conformation capture assay. Nat Protoc. 2008;3: 303–313. doi:10.1038/nprot.2007.540

133. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462: 58–64. doi:10.1038/nature08497

134. Wright JB, Brown SJ, Cole MD. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol. 2010;30: 1411–1420. doi:10.1128/MCB.01384-09

135. De Laat W, Grosveld F. Spatial organization of gene expression: the active chromatin hub. Chromosome Res. 2003;11: 447–459. Available: http://www.ncbi.nlm.nih.gov/pubmed/12971721

136. Babraham Bioinformatics - HiCUP Hi-C Analysis Pipeline [Internet]. [cited 9 Sep 2015]. Available: http://www.bioinformatics.babraham.ac.uk/projects/hicup/

137. Shavit Y, Lio' P. CytoHiC: a cytoscape plugin for visual comparison of Hi-C networks. Bioinformatics. 2013;29: 1206–1207. doi:10.1093/bioinformatics/btt120

138. Phanstiel DH, Boyle AP, Araya CL, Snyder MP. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics. 2014;30: 2808–2810. doi:10.1093/bioinformatics/btu379

139. Walter C, Schuetzmann D, Rosenbauer F, Dugas M. Basic4Cseq: an R/Bioconductor package for analyzing 4C-seq data. Bioinformatics. 2014;30: 3268–3269. doi:10.1093/bioinformatics/btu497

140. Ma W, Ay F, Lee C, Gulsoy G, Deng X, Cook S, et al. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Methods. 2015;12: 71–78. doi:10.1038/nmeth.3205

141. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012;9: 72–74. doi:10.1038/nmeth.1778

142. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11: 163–166. doi:10.1038/nmeth.2772

143. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991;251: 767–773. Available: http://www.ncbi.nlm.nih.gov/pubmed/1990438

144. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

145. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309: 1728–1732. doi:10.1126/science.1117389

146. Larsson C, Koch J, Nygren A, Janssen G, Raap AK, Landegren U, et al. In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. Nat Methods. 2004;1: 227–232. doi:10.1038/nmeth723

147. Crosetto N, Bienko M, van Oudenaarden A. Spatially resolved transcriptomics and beyond. Nat Rev Genet. 2015;16: 57–66. doi:10.1038/nrg3832