1 INTRODUCTION

1.1 GENOME ORGANISATION

It's oft-stated that the DNA within each human cell would extend for two metres fully extended. Instead that same length of DNA packs into a cell nucleus with a diameter in the order of micrometers (μm). This is achieved through a complex organisation hierarchy, ranging from how chromosomes are arranged in the nucleus in territories, to nuclear interactions with the nucleolus or periphery, down to how DNA is wrapped around nucleosomes (for recent reviews, see 1,2). While the biophysics of the latter level of organisation may be well understood, more broadly little is known about the guiding principles and functional importance of higher order chromatin organisation.

This introductory section will describe the current state-of-the-art in chromosome conformation capture experimental methods, as well as criticisms and considerations when interpreting these data, and discuss what is currently understood or theorised about the structure and function of higher order genome organisation. We compare competing models which attempt to recapitulate mechanisms of chromatin folding, and also explore some of the best understood organisational strata in mammalian higher order genome organisation.

1.1.1 C-methods and Hi-C

Classical studies of chromosome conformation relied on microscopy techniques to visualise nuclear architecture, most commonly fluorescence *in situ* hybridisation (FISH). These techniques led to the discovery of "chromosome territories", regions of the nucleus wherein distinct chromosomes were thought to occupy, and more broadly identified the non-random arrangement of loci in three-dimensional space.[3,4] Finer details of chromatin organisation, such as the proposed 30 nm fibre, were also introduced through microscopy-based techniques. Techniques such as FISH are powerful for precise inspection of single genes, but are low-throughput and offer limited resolution.[3]

With the advent DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[5] was the first sequencing-based method of assaying nuclear architecture. The method uses formaldehyde to cross-link nuclear proteins in place, trapping genomic regions that were physically co-located through bound proteins, then to apply a frequent restriction enzyme to shear the sample into fragments. Next, under dilute conditions, DNA

fragments are ligated together. The dilute conditions favour ligations between fixed fragments, with the aim of generating hybrid fragments form two genomic regions which were close together in the original preparation. Cross-linking can then be reversed and, in the case of the original 3C method, measured by quantitative PCR using pre-designed primers for your fragments of interest. The end result is a relative measure of interaction frequency between any two regions of interest, in theory directly proportional to their distance in three-dimensional space.

The rapid advancement of sequencing, allowed the original 3C method to be further developed, first through microarray technology, then using high-throughput sequencing. Two protocols were proposed for a 3C-inspired one-to-many assay[6,7] (both named 4C), whereby interactions were measured for a specific "viewpoint" fragment against all other restriction fragments genome-wide. The same year a many-to-many assay (5C) allowed measurements for all restriction fragments within a specified region.[8]

The final step was an all-versus-all assay, capable of assaying pairwise interaction frequencies between all restriction fragments of a genome. This assay was published by Lieberman Aiden *et al.*[9] and named Hi-C (Fig. 1). The Hi-C method added biotin tagging to pull-down only ligated fragments for sequencing. At the time of publication, resolution of Hi-C data for analysis was limited by sequencing depth, given the huge number of restriction fragments produced by a 6-cutter enzyme (HindIII and NcoI were used in[9]) but the falling costs of sequencing and proven utility of the assay meant subsequent Hi-C papers incrementally increased their sequencing depth, to a point where analysis could be performed at the level of individual restriction fragments, genome-wide.[10–13]

1.1.2 Hi-C variants

The interaction maps produced by Hi-C were found to exhibit several inherent biases. Fragment properties, such as their length, GC content and mappability, were confounding interaction frequency estimates and therefore needed to be normalised-away before subsequent analysis.[14,15] A range of statistical techniques were developed to correct for these latent variables,[16–19] while experimentalists instead looked to improve on the experimental procedure itself.

Tethered chromosome capture (TCC)[20] was the first attempt to increase the signal to noise ratio of Hi-C contacts. In this method, ligations take place on a fixed surface, with the aim of preventing spurious ligations between fragments in solution which were not cross-linked. Kalhor *et al.*[20] reported a large decrease in observed interchromosomal contacts in their tethered library, suggesting many of those originally observed were caused by spurious ligation of non-crosslinked fragments.

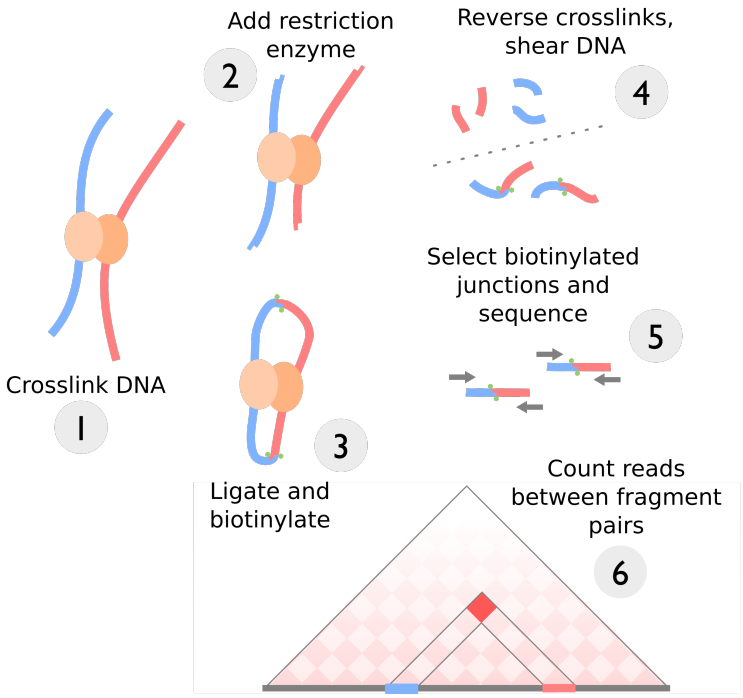

Figure 1: Steps in the Hi-C assay. Schematic of the Hi-C experimental procedure as described in Lieberman Aiden *et al.*[9]

Hi-C is a population-level assay, as the retrieved interaction counts are from a huge number of different cells. As well as building population-averaged models of genome structure, it is also of interest to probe cell-to-cell variability through single-cell approaches. For instance, it's been estimated that long-range contacts identified with C-methods may occur in as few as 10% of cells at any one time.[4]

In the first single-cell Hi-C study, Nagano *et al.*[21] aimed to explore this cell-to-cell variability by performing the Hi-C assay on single, hand-selected nuclei. An obvious limitation this Hi-C variant is that a single restriction fragment can ligate to at most one other fragment, meaning even if 100% yield were to be achieved, any $n \times n$ restriction fragment interaction matrix could at most populate $\frac{n}{2}$ cells; in practice, the realised yield of this first single cell Hi-C experiment was just 2.5%.[21] Nevertheless, single-cell Hi-C was able to reproduce findings from population-based (or "ensemble") Hi-C, such as preferential interactions between active domains, but also was able to dissect *trans* interactions, suggesting high cell-to-cell variability leads to their relatively uniform appearance in normal Hi-C interaction maps.[21] Combined with observations from TCC which gave evidence that interchromosomal contacts were disproportionately the result of spurious ligation,[20] the functional significance of these *trans* interactions seems at best unclear in the general case.

Capture-C is a C-method variant which attempts to address resolution problems associated with the Hi-C genome-wide pairwise assay by enriching for functional

interactions using *a priori* selection of target loci.[22] Indeed, a suggestion in the original Hi-C paper was that resolution could be improved by either increased sequencing or using hybrid capture.[9] Since then, Hi-C variants with a target enrichment step have been developed, including Capture Hi-C (CHi-C)[23] and HiCap.[24] These methods have been applied to genome-wide target sets (e.g. CHi-C assayed 22,000 human promoters[25]) and so it could be said that they are to Hi-C as exome-capture is to a whole-genome sequencing, in the contexts of conformation capture and variant discovery respectively.

In-situ Hi-C was a recent refinement of the Hi-C method, from the publisher of the original method.[13] The principle difference is that fixation and ligation now happen in place, within intact cell nuclei.

1.1.3 Chromosome compartments

In the paper describing the Hi-C technique,[9] Lieberman-Aiden *et al.* described low-resolution structures they name "A" and "B" nuclear compartments. These are regions with a median size of around 5 megabases which showed properties typical of euchromatin and heterochromatin, respectively. A compartments were observed through 3D-FISH to be centrally-positioned in the nucleus and ChIP-seq data showed several hallmarks of transcriptional activity. B compartments, conversely, were heterochromatic and lamina-associated regions, with little transcription and repressive histone modifications such as H3k9me3.[3,9] As expected from positioning data, the co-location of compartment types is also visible in their contact maps.

These compartments were identified through a continuous eigenvector profile, derived from a normalised Hi-C contact matrix[9] (Fig. 2). This approach can be intuitively understood as formulated by Lajoie *et al.*[26]:

1. A tartan pattern on normalised Hi-C matrices indicates two preferentially-contacting compartments (Fig. 2).
2. Assume a function (c) that maps a given genomic bin to its compartment, using a positive number for compartment A and *vice versa*.
3. The interaction frequency between bins i and j is thus $c(i) \cdot c(j)$. (This is enough to generate a tartan pattern: if i and j are in the same compartment, the product will be positive.[26])
4. Our symmetric Hi-C matrix thus contains $c(i)c(j)$ and in this formulation, principle components analysis is finding the basis that minimises the mean-squared error between $c(i)c(j)$ and $c(i)$.

Importantly, this measure holds more information than a simple two-state classification, rather the continuous values can be interpreted as relative levels of compaction or activity.[16,17]

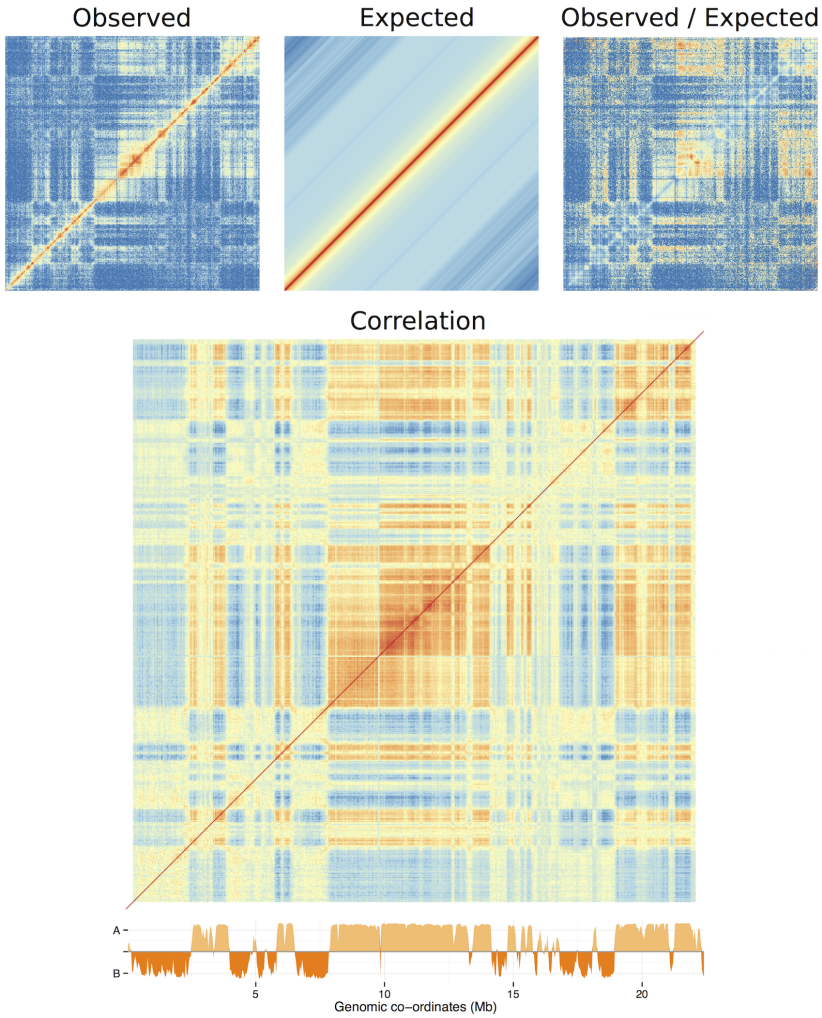

Figure 2: Derivation of A/B compartment profile from Hi-C data. Observed interaction frequencies (O) are averaged along super-diagonals to give a distance-normalised expected matrix (E). The Pearson correlation of the O/E matrix then can undergo eigenvector expansion; in most cases eigenvector $\mathbf{v}$ with the largest eigenvalue, λ, then reflects A/B compartmentalisation.[9]

1.1.4 Topological domains

The falling cost of high-throughput sequencing enabled increasingly deep sequencing of Hi-C experiments. Sequencing is the main resolution-limiting resource for this assay, as to increase the analysis resolution and maintain the level of coverage requires an exponential increase in the total amount of sequencing required.[9,27]

In experiments totalling around two billion total sequencing reads, Dixon *et al.*[10] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. At the same time, Nora *et al.*[28] published an even higher-resolution 5C dataset covering a 4.5 Mb region of the mouse X chromosome. In both of these studies, the authors note "topological associative domains" (or TADs) which were observable as self-interacting, off-diagonal blocks of higher-than-expected self-interaction frequency. With a mean size of around 1 Mb, TADs were recognised as a novel layer of higher order chromatin organisation at a level below the larger A/B compartments (Section 1.1.3). TADs have since been reported in a variety of metazoan organisms including dog,[29] *Drosophila*[30,31] and *C. elegans*[32] yet comparable structures are not found in higher plants such as *Arabidopsis*[33,34] or in yeast.[35,36]

Dixon *et al.*[10] defined a TAD calling algorithm based on the directional bias of a genomic region's contacts, and used a Hidden Markov Model to infer blocks of strongly up- or downstream-biased, reasoning that domain boundaries are present when a strongly upstream biased region is adjacent to a region of opposite bias (Fig. 3). These boundaries themselves were investigated and were found to display suggestive functional enrichments for DNA binding proteins including CTCF, long thought to act as an insulator of chromatin state (Section 1.2.3). Deletion of a CTCF site has been found to disrupt the corresponding TAD border, while removal of some other enriched factors had little effect.[28,37,38] The authors also performed some comparative analysis, reporting large and significant overlap of domain boundary positions both within species and between human and mouse cell lines.[10]

Since then, several studies have investigated the functional implications of TADs. A simple biological explanation is that TADs—by definition—delimit functional contacts, such as those between enhancers and promoters, and so could inhibit spurious contacts with other nearby genetic elements.[2,39] Hormonal treatment of human breast cancer cells reported coordinated expression responses within TADs, suggesting they function as domains of transcriptional co-regulation called "regulons".[40] However the size of TADs means they often span multiple genes, commonly with unrelated functions, so it seems unlikely they can function as regulons in the general case.[1]

1.1.5 Other proposed structures

Filippova *et al.*[41] developed a tuneable algorithm which identifies "alternative topological domains". The authors use dynamic programming to search for an optimal

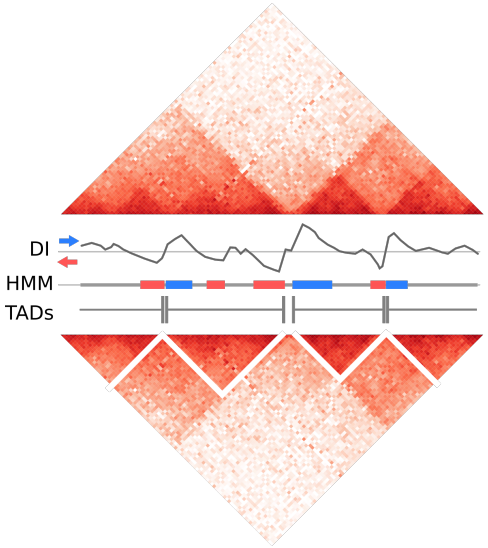

Figure 3: Dixon *et al.* pipeline for calling topological associating domains (TADs). First a directionality index (DI) is calculated for each bin based on the ratio of upstream:downstream contacts. Secondly a Hidden Markov Model (HMM) is used to infer the most likely state sequence that emitted the DI variable. Finally a simple rule is applied whereby a run of high-confidence upstream-biased state calls marks the end of a domain. New domains begin with any subsequent downstream-biased state. Gaps between TAD calls can be observed, and as labelled border regions up to a size threshold of 400 kb, whereafter those regions are unclassified.[10]

set of non-overlapping boundary pairs that maximise intra-domain contacts. The algorithm includes a length scaling factor (γ) which is used to penalise domain size; by varying γ the authors define a subset of "multiscale domains" of heightened persistence across resolutions.[41] These multiscale domains were found to be smaller, on average, than those previously reported by Dixon *et al.*[10], despite be applied to the same Hi-C experimental data (mean size: 200 kb as opposed to $\approx$ 1 Mb). However the domains of Filippova *et al.*[41] show increased intra-domain contacts and stronger boundary enrichments relative to previously-described TADs, indicating this algorithm may generate a more accurate representation of topological domains in mammalian genome organisation. Intriguingly, this study also reports quantitative evidence for hierarchical genome organisation, finding that those domains called at large γ will then combine into larger meta-domains as the γ penalty decreases.

A study of *Drosophila* embryonic chromosomes found a similarly hierarchical organisation of physical domains, and also was able to relate these to "epigenomics domains" showing specific sets of enrichment signatures representing active, null, polycomb-associated and telomeric regions.[30]

Recent high-resolution studies have been able to resolve ever-smaller levels of sub-structure. Rao *et al.*[13] refined the concept of chromosome compartments to "sub-compartments", dividing simple A/B divisions into a total of 5 subtypes. The authors were also able to identify "contact domains" of median size 185 kb, many of

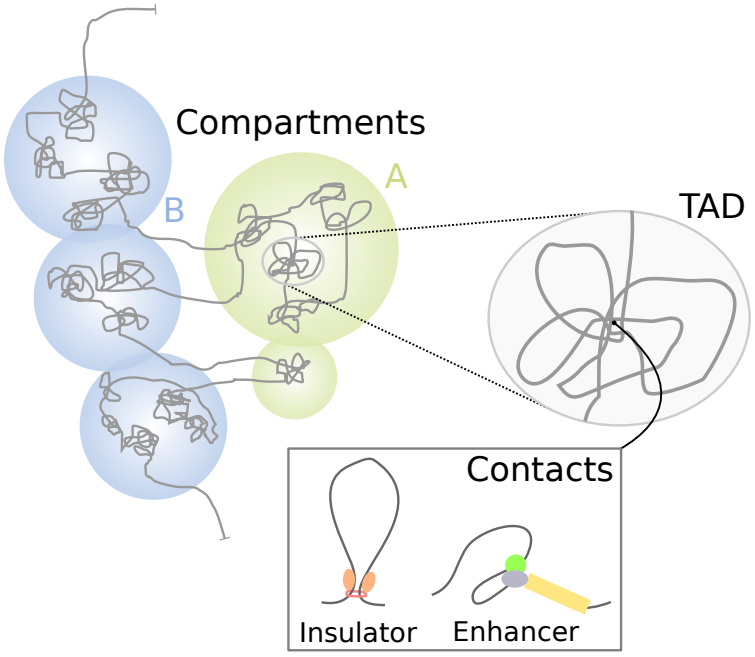

Figure 4: Levels of higher order chromatin organisation. Cartoon showing how functional contacts, such as loops between bound CTCF insulators (Section 1.2.3), occur within TADs (Section 1.1.4) which in turn are found within A or B compartments (Section 1.1.3).

which were associated with identifiable individual looping events (Section 1.2.3).[13] This domain size is close to those of Filippova *et al.*[41] (described above) and the authors here suggest that previously-observed large TADs may be the result of insufficient sequencing; that is, not all boundaries could be detected using 40 kb binned contact maps thus multiple contact domains were unintentionally combined into large domains.

1.2 MODELS OF CHROMATIN FOLDING

Theoretical mechanistic models of chromatin folding such as the "strings and binders switch" model[42] and the "fractal globule" model[9,43,44] have both produced simulated data that reflects empirical C-method observations and potentially describe the polymer dynamics of chromatin folding.

1.2.1 Fractal globule

Lieberman Aiden *et al.*[9] tested a number of theoretical models of genome folding to see which best explained the observed power-law scaling between distance and observed contact frequency (IF $= \frac{1}{dist^{-\alpha}}$ where $\alpha \approx 1.08$). The authors sought to

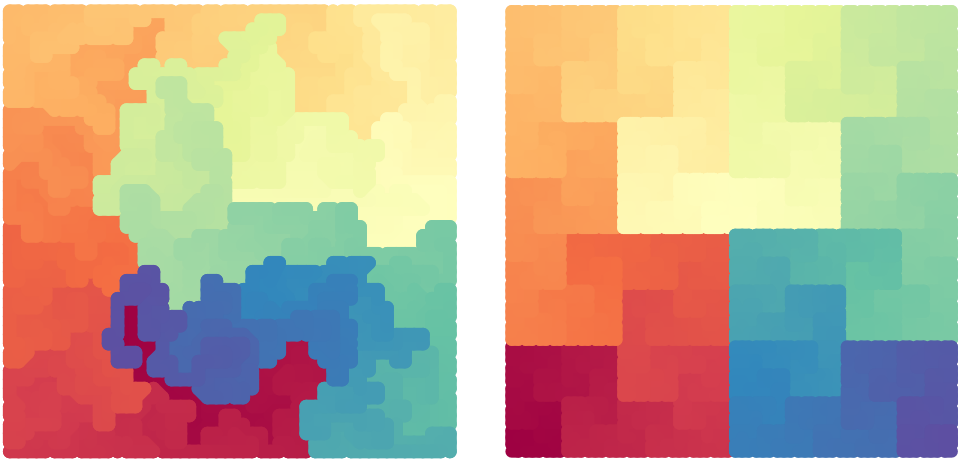

Figure 5: Comparison of theoretical models of chromatin folding. Two theoretical models of chromatin folding are shown simulated along a 2D polymer. An equilibrium globule is represented by a Hamiltonian path through a grid network (*left*) and is compared to a Fractal Globule, here represented by a Hilbert curve (*right*).

distinguish two previously-described models of genome organisation: the "fractal globule" and "equilibrium globule" (Fig. 5). The study found that a theoretical fractal globule, embodying scale-independent self-similar aggregate folding, better fit the observed data than an equilibrium globule null model where simulated polymer folding was allowed to proceed unchecked.

The fractal globule model was noted for its appealing functional properties. Under this model, for example, the polymer folds are knot-free hence could facilitate local dynamics of repression and activation without wider disruption. Despite this appeal, the authors were careful to state that while their simulations show good agreement with observed data, this does not preclude other organisational models from having similar or greater explanatory power.[9]

1.2.2 Strings and binders switch

Subsequent modelling techniques integrated known biological phenomena as well as polymer models. This formed the basis of Barbieri *et al.*'s[42] "strings and binders switch" (SBS) model, where the authors simulated polymer folding in the presence of DNA binding factors, such as the known genome organiser CTCF (Section 1.2.3). This organisational model was developed in an attempt to consolidate global Hi-C measures of contact scaling with C-based experiments on smaller regions and FISH studies, which found a range of scaling parameters. The authors also explore the different values of α between cell lines and even chromosomes, and find that their mechanistic model can explain each case using variable concentrations of binders

which causes phase-switching between open and compacted chromatin, with fractal globule existing at the phase transition boundary.

This model offers broad explanatory power for a range of observed power law coefficients (α) and from simple underpinnings, but critics point out that simulations were performed on a polymer composed of just 512 monomers so may not be broadly applicable.[17]

1.2.3 Looping and CTCF

Examples have long been known of specific enhancer elements that are brought into close proximity with the promoter(s) they are regulating; under this model, these contacts form a "loop" structure between two potentially distal loci[45,46] (Fig. 4). A model region, the β-globin locus and its locus control region (LCR) located 40-80 kb away,[17] has been studied since the early 1980s,[47–50] and is an interesting example of a well-characterised looping. Current knowledge suggests the β-globin forms loops with the multiple distal *cis*-enhancer elements which make up the LCR, together forming an active chromatin hub (ACH).[51] Within such a hub, regulatory signals could be efficiently integrated to dictate the overall activity of the target locus.[1,3] It is now thought that the majority of active promoters are engaged with multiple, often cell type specific, regulatory looping events.[12,52]

A notable component of many long-range looping events is the CCCTC-binding transcription factor (CTCF),[53,54] already mentioned as a component of TAD boundaries (Section 1.1.4) and as a proposed looping factor in the SBS model (Section 1.2.2). CTCF is strongly conserved in higher eukaryotes,[55] ubiquitously expressed and embryonic lethal, but it is not tied to a single biological function — instead CTCF has been described as a "multivalent factor",[54] capable of regulating transcription, imprinting, dosage-compensation and acting as an insulator.

In the context of genome organisation, CTCF is of interest for its role of anchoring interactions between loci, forming loops. Experimental evidence has shown that interactions between CTCF sites stabilises the aforementioned loops linking the β-globin locus with its distal LCR.[56] This looping role, potentially undertaken in combination with other architectural proteins such as Mediator and cohesion,[46,57] can explain its previously-identified insulator behaviour: CTCF can block the spread of heterochromatin and contacts between enhancers and promoter through topological constraints by forming loops.[54] It must be said, however, that the functional significance of CTCF-mediated loops, and indeed the role of CTCF in even well-studied systems, remains only partially understood.[58]

A recent Hi-C paper, that of Rao *et al.*[13], again brought CTCF and looping to the fore of chromatin conformation research. This study identified around 10,000 individual looping events in the human genome, almost all linking loci over distances

of less than 2 Mb, and around 30% connecting predicted enhancer and promoter chromatin states. Rao *et al.*[13] also found a 6-fold overall increase in expression when comparing those promoters participating in a looping event with those not. Furthermore, 86% of these loops involved CTCF bound regions, with roughly the same proportion involving cohesion subunits RAD21 and SMC3. The authors thus proposes that a CTCF-binding motif formed the "anchor" for this transitive complex of architectural proteins.[13] A majority of these loops (65%) also demarcated a topological domain, and at much higher resolution than previously observed (Section 1.1.4). Another striking finding of this research was that CTCF loops almost always occur in between bound motifs with a convergent orientation,[13] though questions remain over why this should be the case, especially when considering the interactions of a flexible polymer in 3-D solution.[59]

While the evidence linking CTCF and genome architecture is substantial, it should be noted that from a global perspective as few as 15% of all CTCF sites were found to occur at TAD boundaries in human and mouse cells[10] and similarly around 25% if TAD borders had no observable CTCF binding.[39] These facts indicate that CTCF alone is neither necessary nor sufficient for the formation of higher order chromatin structures such as TADs. Indeed, the degree of insulation at a given genomic site was recently reported to correlate with the degree of co-binding of a range of architectural proteins including not only CTCF but cohesin, condensin and the transcription complex TFIIIC, among others.[60]

1.3 CRITICISMS OF C-METHODS

C-methods are a relatively new and developing set of assays, especially compared to long-standing microscopy techniques which have for decades been used to visualise chromosome conformation. In this section, we discuss some of the limitations and issues with applying or interpreting the results of C-methods.

1.3.1 Cell populations

As previously mentioned (Section 1.1.2), the Hi-C assay typically takes place in a cell population (though proof-of-concept single-cell experiments have been reported[21]). An obvious consideration, then, is that all interaction counts reflect the average over a large number of cells, often including unsynchronised populations at different stages of the cell cycle.[2] Given evidence that, while the interphase chromosomes exhibit cell-to-cell variability, the mitotic state is much more static,[61,62] one might expect even a small proportion of dividing cells to add a detectable bias to averaged genome-wide contact maps.

1.3.2 Ploidy

A more esoteric consideration with C-methods data is that organisms under study are typically diploid, while maps of chromosome organisation are commonly collapsed onto a haploid pseudo-genome. Haplotype conformation can be delineated from C-methods data a variety of ways, such as using haploid cell lines (e.g. 13) or via detectable sequence differences using either deep sequencing or a targeted area (e.g. 63). An altogether different and inventive solution is to use the inherent proximity-ligation information produced by C-methods to discriminate haplotypes,[11] an idea since extended to deconvolution problems in metagenomics.[64,65]

1.3.3 Resolution

The resolution of a Hi-C experiment has a hard-limit imposed by the choice of restriction enzyme. For example, the commonly-used HindIII enzyme is a six-cutter that recognises the motif AAGCTT and cuts approximately every 4 kb, on average.[3] This results in on the order of 10 million restriction fragments with a total pairwise interaction space of 10^{12}.[26] The depth of sequencing required to cover this interaction space is cost-prohibitive, so in practice analysis takes place with data aggregated into bins of either fixed length or fixed number of restriction fragments.

More recent studies have switched to a four-cutter restriction enzyme, for example MboI,[13] which increases this upper-bound on resolution to the order of hundreds of basepairs (e.g. theoretical mean fragment size of 422 bp in mouse[24]), but again ultra deep-sequencing is required to realise such resolutions during analysis. A downside of using more frequent restriction enzymes is the potential side-effect of promoting more non-specific ligations by increasing the concentration of fragments in solution.[13]

Realistically and in most instances, an experimental design may either target high-resolution interactions through targeted 4C or 5C, or low-resolution genome-wide interactions — but not both.

1.3.4 Biological interpretation

A key consideration with C-methods is that, when accurately stated, the assays are measuring "the frequency at which sequences are ligated together by formaldehyde cross-linking",[66] which is then assumed to be a proxy for physical distance within the nucleus. This is a marked difference from aforementioned FISH methods, where the physical distance is observed directly, albeit through the addition of non-native probes. So strong is this assumption, that methods have been developed that use a known FISH distance to then calibrate genome-wide Hi-C distances,[67] however it need not be the case that population-level interaction frequencies capture physical

distance.[26] Consider, for example, a tight enhancer–promoter interaction occurring in 50% of cells, but not at all in the other half. In this scenario, the two loci would have an intermediate interaction frequency overall, which is then converted to a distance measure that reflects the realities of neither cell sub-population. For similar reasons, the transience of an interaction cannot be directly inferred from its interaction frequencies: a weak interaction frequency may be the result of either the same fleeting contact in many cells, or stable contacts in only a subset of cells.[26]

When interpreting C-methods data it should also be kept in mind that even verifiable contacts are by no-means functional. To elaborate, C-methods may find two regions to be strongly co-localised, but an understanding of the region may explain their co-localisation to be caused by mutual interaction with a nuclear lamina or nucleolus, for example, rather than any specific functional relationship between the two loci.[17] In addition, a functional enhancer–promoter interaction will necessarily constrain the contacts of other nearby regions, potentially causing highly-reproducible "bystander interactions"[17] that are nevertheless uninteresting from a functional perspective.

1.3.5 Other considerations

An additional and separate issue identified with C-methods, specifically 3C in this instance, emerges from reports that the observed ligation frequency is as low as 1% of expected values in a model system,[68] potentially magnifying the relative influence of noise and artefacts.

1.4 MACHINE LEARNING IN GENOMICS

Machine learning offers a powerful framework for understanding complex datasets, such as those produced in large-scale genomics studies. Problems in the field such as gene prediction and inferring regulatory networks can be approached by employing a learning algorithm, either in a supervised way based on a known truth set, or through unsupervised methods aimed at pattern detection or clustering (for reviews see 69,70). If a successful predictive model can be built, it can then be dissected to explore statistical rules which may impart novel biological insight. As a toy example, learning a highly-accurate model of enhancer prediction could itself identify novel epigenetic marks indicative of enhancers, generating testable hypotheses about how enhancers are activated.

In this section, we introduce recent and high-profile machine learning applications in the context of the ENCODE consortium, and give examples of how their vast datasets have empowered research groups worldwide to tackle complex biological

questions through a variety of machine learning approaches. We then discuss research broadly aligned with the aims of this thesis, those attempting to advance an understanding higher order chromatin structure through machine learning and related techniques.

1.4.1 ENCODE

The Encyclopaedia of DNA Elements (ENCODE) is a consortium project started over a decade ago with the ambitious aim of comprehensively cataloguing all functional elements in the human genome.[71–73] This project involves huge amounts of data production from a diverse array of experimental methods, such as: ChIP-seq, DNase-seq, RNA-seq, CAGE, DNase-seq and ChiA-PET.[74] Importantly these methods were applied to a range of human cell types, including many well-studied immortalised cell lines as well as primary cells and tissues, and according to standardised experimental methods[75] coupled with statistical quality control[73,76,77] to ensure data is comparable between different data produces and of consistently-high accuracy. Despite ENCODE's human-focus, there also exists spin-off projects aimed at building similar genomics resources for mouse[78] and, more recently, *Drosophila* and *C. elegans*.[79] Together these data sources offer an unparalleled resource for comparative and within-species genomics research, and as such have been used in at least 1200 publications to date.[80]

Data generated by ENCODE consortium members has a proven utility in genomics research. Notably two ENCODE-associated groups have released models which classify the human genome into discrete "chromatin states", such as actively transcribed regions or gene promoters. The first, named `SegWay`, trained a dynamic Bayesian network on 31 ENCODE-generated input variables and called an unsupervised 25-state genome segmentation in the ENCODE pilot region.[81] Independently another chromatin state predictor named `ChromHMM` was developed.[82,83] As the name suggests, this approach instead used multivariate Hidden Markov Models (HHMs) and has the capability to learn a single generative model over multiple cell types. Original runs of the model called 51 chromatin states using over 40 input features,[84] but more recently these two methods were combined to call a consensus set of just 7 chromatin states.[85] Since their publication, a study was able to experimentally validate many of these state predictions.[86] This discretisation of the chromatin landscape greatly helps interpretability, at the cost of simplifying the complex underlying data series, and is used for this reason later in this work (Section **??**).

More broadly, ENCODE data has been used by external researchers to generate input variables for machine learning-based predictive models which describe transcriptional output,[87] gene regulation,[88] cell cycle-associated genes[89] and enhancer identification[90] to name but a few. One such study in particular, that of Dong *et al.*[91], is reproduced and further analysed in this work (Section **??**) and is used as a

template for our own machine learning framework applied in the context of higher order chromatin structure (Chapter **??**). We also make use of ENCODE data in other chapters (e.g. Chapter **??**) due to its comprehensive coverage of model human cell types and stringent data production guidelines referenced above.

1.4.2 Related work

In this thesis we will be applying machine learning and other forms of statistical analysis the gain a greater understanding of the biological underpinnings of higher order chromatin conformation (introduced in Section 1.1). We shall now consider existing and overlapping works, some of which were published after or during the time that the research presented herein was performed.

1.5 AIMS

In the broadest terms, the aims of this work are to investigate the relationship between structure and function of the genome. In particular, we aim to answer the following questions:

1. How does higher order chromatin structure compare across cell types?
2. Can we predict higher order chromatin structure from locus-level features?
3. How do the characteristics of boundaries remarking higher order domains vary between cell types and domain classes?

In an attempt to address these questions, we will bring together the huge volumes of data generated by the ENCODE consortium (Section 1.4.1) and employ machine learning techniques and other statistical analysis to explore how these locus-level features relate to higher order chromatin structure.

REFERENCES

[1] Pombo A, Dillon N (2015) Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*, **16**(4): 245–257.

[2] Fraser J, Williamson I, Bickmore Wa, Dostie J (2015) An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, **79**(3): 347–372.

[3] de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. *Genes & development*, **26**(1): 11–24.

[4] van Steensel B, Dekker J (2010) Genomics tools for unraveling chromosome architecture. *Nature biotechnology*, **28**(10): 1089–1095.

[5] Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science (New York, N.Y.)*, **295**(February): 1306–1311.

[6] Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, **38**(11): 1341–1347.

[7] Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, **38**(11): 1348–1354.

[8] Dostie J, Richmond Ta, Arnaout Ra, Selzer RR, Lee WL, Honan Ta, Rubio ED, Krumm A, Lamb J, *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**(10): 1299–1309.

[9] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950): 289–93.

[10] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398): 376–80.

[11] Selvaraj S, R Dixon J, Bansal V, Ren B (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology*, **31**(12): 1111–8.

[12] Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen Ca, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**(7475): 290–4.

[13] Rao S, Huntley M, Durand N, Stamenova E, Bochkov I, Robinson J, Sanborn A, Machol I, Omer A, *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**(7): 1665–1680.

[14] Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**(11): 1059–65.

[15] Hu M, Deng K, Qin Z, Liu JS (2013) Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. *Quantitative Biology*, **1**(2): 156–174.

[16] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, **9**(10): 999–1003.

[17] Dekker J, Marti-Renom Ma, Mirny La (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, **14**(6): 390–403.

[18] Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics (Oxford, England)*, **28**(23): 3131–3.

[19] Li W, Gong K, Li Q, Alber F, Zhou XJ (2014) Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics (Oxford, England)*, (November): 1–3.

[20] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, **30**(1): 90–8.

[21] Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**(7469): 59–64.

[22] Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R, Higgs DR (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, **46**(2): 205–12.

[23] Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, *et al.* (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome research*, pp. 1854–1868.

[24] Sahlén P, Abdullayev I, Ramsköld D, Matskova L, Rilakovic N, Lötstedt B, Albert TJ, Lundeberg J, Sandberg R (2015) Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, **16**(1): 156.

[25] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, (April): 1–12.

[26] Lajoie BR, Dekker J, Kaplan N (2014) The Hitchhikers Guide to Hi-C Analysis: Practical guidelines. *Methods*, (November).

[27] Tanay A, Cavalli G (2013) Chromosomal domains: epigenetic contexts and functional implications of genomic compartmentalization. *Current opinion in genetics & development*, **23**(2): 197–203.

[28] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398): 381–5.

[29] Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S (2015) Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, **10**(8): 1297–1309.

[30] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**(3): 458–72.

[31] Hou C, Li L, Qin ZS, Corces VG (2012) Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, **48**(3): 471–484.

[32] Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*.

[33] Feng S, Cokus S, Schubert V, Zhai J, Pellegrini M, Jacobsen S (2014) Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis. *Molecular Cell*, **55**(5): 694–707.

[34] Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, Lanz C, Weigel D (2015) Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Research*, **25**(2): 246–256.

[35] Duan Z, Andronescu M, Schutz K, McIlwain S, Mcllwain S, Kim YJ, Lee C, Shendure J, Fields S, *et al.* (2010) A Three-Dimensional Model of the Yeast Genome. *Nature*, **465**(7296): 363–367.

[36] Gong K, Tjong H, Zhou XJ, Alber F (2015) Comparative 3D Genome Structure Analysis of the Fission and the Budding Yeast. *Plos One*, **10**(3): e0119672.

[37] Zuin J, Dixon JR, van der Reijden MIJa, Ye Z, Kolovos P, Brouwer RWW, van de Corput MPC, van de Werken HJG, Knoch Ta, *et al.* (2013) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, pp. 1–6.

[38] Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, Reinberg D (2015) CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, **347**(6225): 1017–1021.

[39] Sexton T, Cavalli G (2015) Review The Role of Chromosome Domains in Shaping the Functional Genome. *Cell*, **160**(6): 1049–1059.

[40] Le Dily F, Bau D, Pohl a, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RHG, Ballare C, *et al.* (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, **28**(19): 2151–2162.

[41] Filippova D, Patro R, Duggal G, Kingsford C (2014) Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, **9**: 14.

[42] Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, Pombo A, Nicodemi M (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(40): 16173–8.

[43] Mirny La (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, **19**(1): 37–51.

[44] Grosberg AY, Nechaev S, Shakhnovich E (1988) The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique*, **49**(12): 2095–2100.

[45] Kadauke S, Blobel Ga (2009) Chromatin loops in gene regulation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, **1789**(1): 17–25.

[46] Sexton T, Bantignies F, Cavalli G (2009) Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. *Seminars in Cell and Developmental Biology*, **20**(7): 849–855.

[47] Banerji J, Rusconi S, Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**(2 Pt 1): 299–308.

[48] Engel JD, Tanimoto K (2000) Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. *Cell*, **100**(5): 499–502.

[49] Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. *Science (New York, N.Y.)*, **281**(5373): 60–63.

[50] Tolhuis B, Palstra RJ, Splinter E, Grosveld F, De Laat W (2002) Looping and interaction between hypersensitive sites in the active ??-globin locus. *Molecular Cell*, **10**(6): 1453–1465.

[51] van de Corput MPC, de Boer E, Knoch Ta, van Cappellen Wa, Quintanilla a, Ferrand L, Grosveld FG (2012) Super-resolution imaging reveals 3D folding dynamics of the -globin locus upon gene activation. *Journal of Cell Science*, pp. 4630–4639.

[52] Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**(7414): 109–13.

[53] Ong CT, Corces VG (2014) CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics*, **15**(4): 234–46.

[54] Phillips JE, Corces VG (2009) CTCF: Master Weaver of the Genome. *Cell*, **137**(7): 1194–1211.

[55] Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology*, **16**(6): 2802–2813.

[56] Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, De Laat W (2006) CTCF mediates long-range chromatin looping and local histone modification in the ??-globin locus. *Genes and Development*, **20**(17): 2349–2354.

[57] Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong CT, Hookway Ta, Guo C, *et al.* (2013) Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, **153**(6): 1281–95.

[58] Gómez-Díaz E, Corces VG (2014) Architectural proteins: regulators of 3D genome organization in cell fate. *Trends in Cell Biology*, **24**(11): 703–711.

[59] Nichols M, Corces V (2015) A CTCF Code for 3D Genome Architecture. *Cell*, **162**(4): 703–705.

[60] Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, Corces VG (2014) Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome biology*, **15**(5): R82.

[61] Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny La, Dekker J (2013) Organization of the mitotic chromosome. *Science (New York, N.Y.)*, **342**(6161): 948–53.

[62] Dekker J (2014) Two ways to fold the genome during the cell cycle : insights obtained with chromosome conformation capture. **7**(1): 1–12.

[63] Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJG, Zhu Y, Kaaij LJT, van Ijcken W, Gribnau J, *et al.* (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development*, **25**(13): 1371–83.

[64] Burton JN, Liachko I, Dunham MJ, Shendure J (2014) Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps. *G3 (Bethesda, Md.)*, **4**(July): 1339–1346.

[65] Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen Ja, Darling AE (2014) Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, **2**: e415.

[66] Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, Bickmore WA (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. pp. 2778–2791.

[67] Shavit Y, Hamey FK, Lio' P (2014) FisHiCal: an R package for iterative FISH-based calibration of Hi-C data. *Bioinformatics (Oxford, England)*, pp. btu491–.

[68] Gavrilov Aa, Golov AK, Razin SV (2013) Actual ligation frequencies in the chromosome conformation capture procedure. *PloS one*, **8**(3): e60403.

[69] Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science*, **349**(6245): 255–260.

[70] Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, **16**(6): 321–332.

[71] Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, Collins F, Gingeras T, Kampa D, *et al.* (2004) The ENCODE (ENCyclopedia of DNA elements) Project.

[72] Qu H, Fang X (2013) A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. *Genomics, Proteomics and Bioinformatics*, **11**(3): 135–141.

[73] ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414): 57–74.

[74] Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, *et al.* (2011) A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**(4).

[75] Landt S, Marinov G (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome . . .*, (Park 2009): 1813–1831.

[76] Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**(7515): 453–456.

[77] Marinov GK, Kundaje A, Park PJ, Wold BJ (2013) Large-Scale Quality Analysis of Published ChIP-seq Data. *G3 (Bethesda, Md.)*, **4**(February): 209–223.

[78] Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**(7527): 355–364.

[79] Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, *et al.* (2014) Comparative analysis of metazoan chromatin organization. *Nature*, **512**(7515): 449–452.

[80] ENCODE DCC (2015) Encode news, may 13 2015. https://www.encodeproject.org/news/. Accessed: 2015-07-29.

[81] Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes Ja, Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, **9**(5): 473–476.

[82] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345): 43–9.

[83] Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3): 215–6.

[84] Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, **28**(8): 817–825.

[85] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes Ja, *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2): 827–41.

[86] Kwasnieski JC, Fiore C, Chaudhari HG, Cohen Ba (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*.

[87] Cheng C, Yan Kk, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome biology*, **12**(2): R15.

[88] Althammer S, Pagès A, Eyras E (2012) Predictive models of gene regulation from high-throughput epigenomics data. *Comparative and functional genomics*, **2012**: 1–13.

[89] Cheng C, Ung M, Grant GD, Whitfield ML (2013) Transcription Factor Binding Profiles Reveal Cyclic Expression of Human Protein-coding Genes and Non-coding RNAs. *PLoS Computational Biology*, **9**(7).

[90] Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS computational biology*, **9**(3): e1002968.

[91] Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome biology*, **13**(9): R53.