Unravelling higher order genome organisation [working title]

Introduction

Benjamin L. Moore

July 20, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

LIST OF ACRONYMS

3C	Chromsome conformation capture (derivatives: 4C, 5C, Hi-C)
AUROC	Area under the receiver operating characteristic
ChIP-Seq	Chromatin immunoprecipitation following by high-throughput sequencing
ESC	Embryonic stem cell
FISH	Flurorescent *in-situ* hybridisation
GO	Gene ontology
Hi-C	High-throughput version of a 3C experiment
HMM	Hidden Markov Model
ICE	Iterative correction and eigenvector expansion
IF	Interaction frequency
MAD	Median absolute deviation
MSE	Mean squared error
OOB	Out-of-bag
PCC	Pearson correlation coefficient
PLS	Partial least squares
RF	Random Forest
RMSE	Root mean-squared error
TAD	Topologically-associating domains

ABSTRACT

Recent technological advances have given insights into how genomes are folded in three-dimensions, however many open questions remain about the functional importance of this structure, its variability and its relationships with other features of the genomic and epigenomic landscape. In this work we combine Hi-C datasets describing physical genomic contacts with a large and diverse array of chromatin data derived at a much finer scale, including for example levels of bound transcription factors, histone modifications and expression data. These data are then brought together in a quantitative and rigorous way, through a predictive modelling framework and applied statistical analyses.

First we compare higher order chromatin organisation across a variety of human cell types and find pervasive conservation of chromatin organisation at multiple scales. Beyond this we identify structurally variable regions that are enhancer-rich and contain loci of known cell-type specific function. We find broad aspects of higher order chromatin organisation, such as chromosome compartments, to be highly predictable in a variety of human cell types. We dissect these models and find them to be generalisable to novel cell types, due to fundamental biological rules linking compartments with key activating and repressive signals. These models describe the strong interconnectedness between locus-level enrichments and depletions of local marks and bound factors with much broader compartmentalisation of large chromosomal regions.

Finally, boundary regions are investigated in terms of chromatin marks and co-localisation with other known nuclear structures. We find boundary complexity to vary between cell types and link TAD aggregations to previously described lamin-associated domains, as well as exploring the concept of super-boundaries that span multiple levels of organisation. Together these analyses lend evidence to the idea of higher order genome organisation that is largely fixed between cell types, yet one that can selectively vary locally, based on the activation or repression of key loci.

1 INTRODUCTION

1.1 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.[? ?] 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.[?]

With the advent DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[?] was the first sequencing-based method of measuring chromosome conformation. 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[? ?] (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.[?]

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.*[?] 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[?]) 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.[? ? ? ?]

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

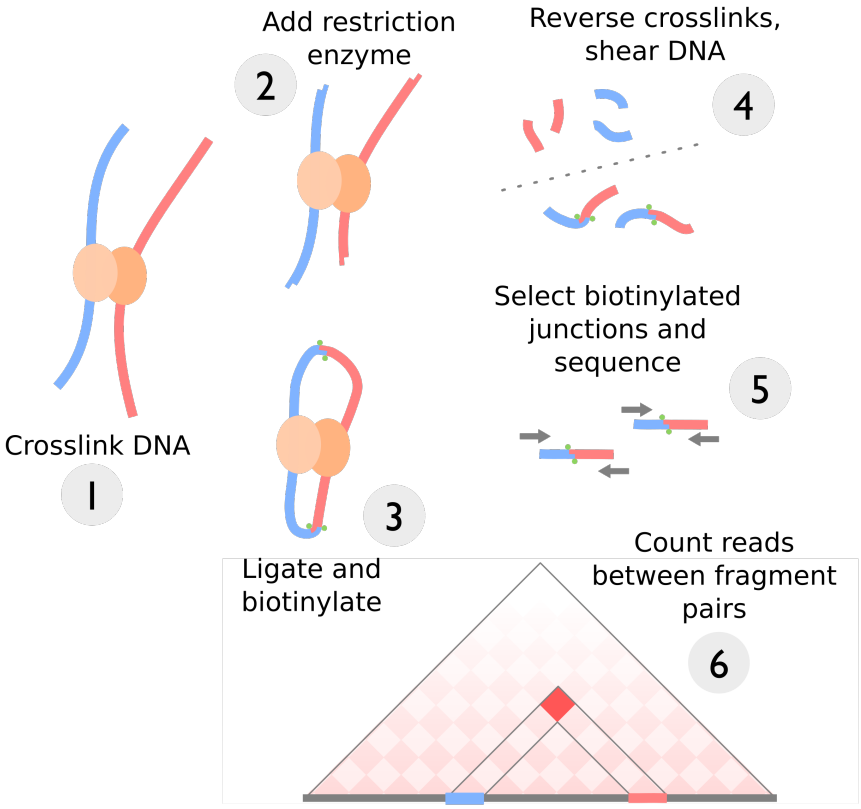

Figure 1: Steps in the Hi-C assay. Schematic of the Hi-C experimental procedure as described in Lieberman Aiden *et al.*[?]

needed to be normalised-away before subsequent analysis.[? ?] A range of statistical techniques were developed to correct for these latent variables,[? ? ? ?] while experimentalists instead looked to improve on the experimental procedure itself.

Tethered chromosome capture (TCC)[?] 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.*[?] 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.

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.[?]

In the first single-cell Hi-C study, Nagano *et al.*[?] 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%.[?] 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.[?] Combined with observations from TCC which gave evidence that

interchromosomal contacts were disproportionately the result of spurious ligation,[?] the functional significance of these *trans* interactions seems at best unclear in the general case.

Capture-C is another recent Hi-C derivative which attempts to address resolution problems associated with the genome-wide pairwise assay by enriching for promoter-enhancer interactions using *a priori* selection.[?] It could be said that Capture-C is to Hi-C as exome-capture sequencing is to a whole-genome approach. Indeed, a suggestion in the original Hi-C paper was that resolution could be improved by either increased sequencing or using hybrid capture.[?]

Use of a cell population also averages away cell-cycle effects, with the vast majority of results coming from cells during interphase (around 97%[?]). Naumova *et al.*[?] looked to assay chromosome conformation specifically over different cell cycle stages, to better understand chromosome compaction during mitosis.

In-site Hi-C was a recent refinement of the Hi-C method, from the published of the original method.[?] 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,[?] 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.[??] 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.[?] 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.[??]

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.[??]

In experiments totalling around two billion total sequencing reads, Dixon *et al.*[?] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. The authors noticed smaller domains they designated "topological associative domains" (or TADs) which were observable as self-interacting, off-diagonal blocks of higher-than-expected self-interaction frequency. They defined a domain 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.

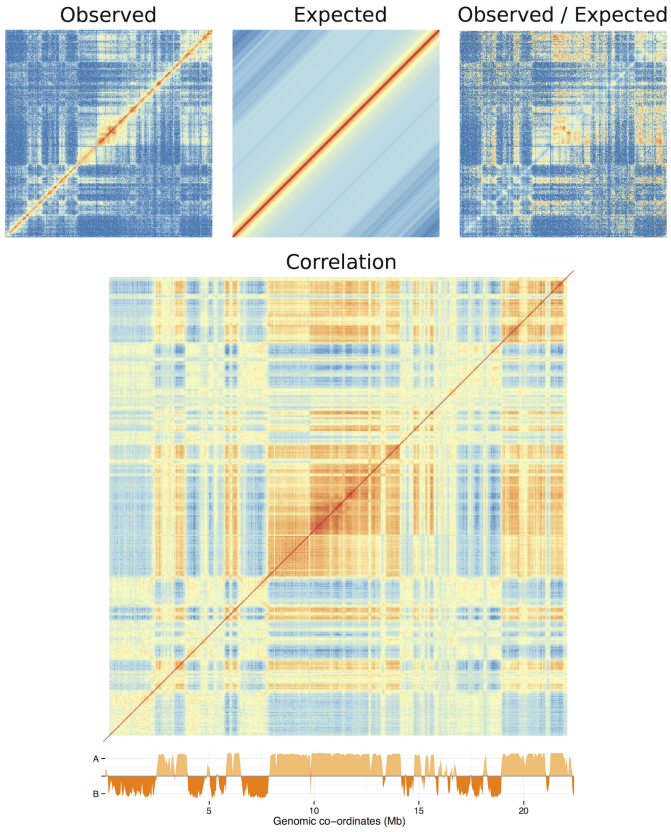

Figure 2: Derivation of A/B compartment profile from Hi-C data. Observed interaction frequencies (O) are averaged along super-diagonals to give a distance-normalised expected matrix (E). The Pearson correlation of the O/E matrix then can undergo eigenvector expansion; in most cases eigenvector $\mathbf{v}$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

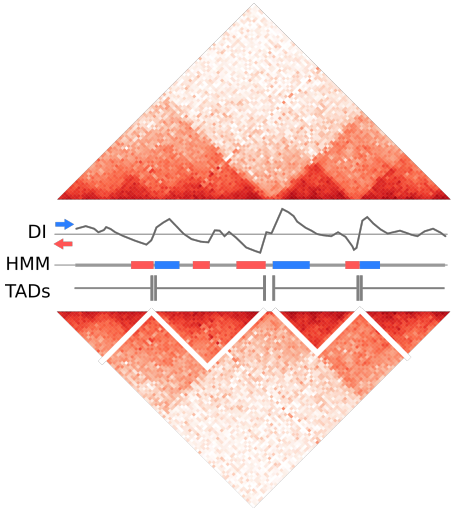

Figure 3: Dixon *et al.*[?] pipeline for calling topological associating domains (TADs). First a directionality index (DI) is calculated for each bin based on the ratio of upstream:downstream contacts. Secondly a Hidden Markov Model (HMM) is used to infer the most likely state sequence that emitted the 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.[?]

Dixon *et al.*[?] also performed some comparative analysis, reporting large and significant overlap of domain boundary positions both within species and between human and mouse cell lines.

1.1.5 Other proposed structures

Filippova *et al.*[?] developed a tuneable algorithm which identifies "alternative topological domains".

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.[?]

Recent high-resolution studies have been able to resolve ever-smaller levels of sub-structure. Rao *et al.*[?] 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 which were associated with identifiable individual looping events.[?] The authors also 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[?] and the "fractal globule" model[???] 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.*[?] 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 ($IF = 1/dist^{-\alpha}$ where $\alpha \approx 1.08$). The authors sought to distinguish two previously-described models of genome organisation: the "fractal globule" and "equilibrium globule". The authors 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.[?]

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[?] "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 CCCTC-binding factor (CTCF).[?] 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 500 monomers.

1.2.3 Looping

1.2.4 Cell cycle changes

Chromosome structure has been assayed both through mitosis[?] and Studies have also focused on the edge-case of chromatin structures on X-chromosomes.

1.3 CRITICISMS OF C-METHODS

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.[?] More recent studies have switched to a four-cutter restriction enzyme, for example MboI,[?] which increases this upper-bound on resolution to the order of hundreds of basepairs (i.e. naively, $4^4 = 256$bp fragments, on average). 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.[?]

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",[?] 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,[?] yet it remains unclear to what extent these two methods are compatible.

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,[?] 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. 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.

The link between epigenomic features and local chromatin state has been analysed computationally in a number of publications, notably in developing the Hidden Markov Model-based ChromHMM[?] algorithm which predicts

states such as active promoters and enhancers, using a range of histone marks and other underlying features.[?] Similarly a Random Forest-based algorithm was developed to predict enhancers from histone modification data.[?] However few studies have spanned all of these levels of chromatin structure and nuclear organisation, and it is not yet known how locus-level chromatin features may be related to higher order genome organisation.

1.4.1 ENCODE

The recent comprehensive ChIP-seq datasets produced by the ENCODE consortium[?] combined with Hi-C genome-wide contact maps in a number of human cell types[? ? ?] present a remarkable opportunity to investigate the relationships between local chromatin features and higher order structure. In this work, a machine-learning approach was employed to model the compartmental characteristics of large genomic regions based on their aggregate levels of various histone marks and DNA binding proteins. Dissection of the resulting models was then used as a means of gleaning biological insights into the basis of higher order structure and of highlighting important differences between cell types.

1.5 AIMS

In the broadest terms, the aims of this work are to investigate the relationship between structure and function of the genome.