Unravelling higher order genome organisation [working title]

Introduction

Benjamin L. Moore

June 11, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

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 positioned in territories down to how DNA is wrapped around nucleosomes.

1.1.1 C-methods and Hi-C

Classical studies of chromosome conformation relied on microscopy techniques to visualise nuclear architecture. These techniques led to the discovery of "chromosome territories", regions of the nucleus wherein distinct chromosomes were thought to occupy. Finer details of chromatin organisation, such as the proposed 30 nm fibre, were also introduced through microscopy-based techniques.

With the advent DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[1] 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[2,3] (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.[4]

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.*[5] and named Hi-C. 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[5]) 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.[6–9]

1.1.2 Hi-C variants

The interaction maps produced by Hi-C were noticed 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.[10] A range of statistical techniques were developed to correct for these latent variables[11–14]

Tethered chromosome capture (TCC)[15] 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.

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. The first single-cell Hi-C study[16] ... An obvious limitation of single-cell Hi-C assays is that a single restriction fragment can ligate to at most a single other fragment, meaning even if 100% yield were to be achieved, any $n \times n$ interaction matrix could at most populate $\frac{n}{2}$ cells.

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.[17] It could be said that Capture-C is to Hi-C as exome-capture sequencing is to a whole-genome approach.

In-site Hi-C was a recent refinement of the Hi-C method, from the published of the original method.[9] 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,[5] 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.[5] As expected from positioning data, the co-location of compartment types is also visible in their contact maps.

Figure: correlation matrix of contacts with eigenvector profile

These compartments were identified through a continuous eigenvector profile, derived from a normalised Hi-C contact matrix.[5] 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.[11,12]

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.[5]

In experiments totalling around two billion total sequencing reads, Dixon *et al.*[6] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. The authors noticed smaller domains they designated "topologi-

cal 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. 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.

Dixon *et al.*[6] 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.*[18] developed a tuneable algorithm which identifies "alternative topological domains".

Recent high-resolution studies have been able to resolve ever-smaller levels of sub-structure. Rao *et al.*[9] 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.[9] 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 ORGANISATION

Theoretical mechanistic models of chromatin folding such as the "strings and binders switch" model[19] and the "fractal globule" model[20**?** ,21] 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

1.2.2 Strings, binders and switches

1.3 CRITICISMS OF C-METHODS

Compare / contrast C-methods and FISH, Bickmore lab papers.

1.4 MACHINE LEARNING IN GENOMICS

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[22] algorithm which predicts states such as active promoters and enhancers, using a range of histone marks and other underlying features.[23] Similarly a Random Forest-based algorithm was developed to predict enhancers from histone modification data.[24] 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 [25] combined with Hi-C genome-wide contact maps in a number of human cell types [6,15?] 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