Unravelling higher order chromatin organisation through statistical analysis

Benjamin L. Moore

Doctor of Philosophy
The University of Edinburgh
2015

DECLARATION

This thesis presents my own work, wherever the contributions of others were involved this is clearly indicated. It has not been submitted for any other degree or professional qualification.

—Benjamin L. Moore (2015)

ACKNOWLEDGEMENTS

Firstly, I must thank my supervisor Colin Semple for all the valued discussions, ideas and mentorship throughout my PhD. Thanks also to my secondary supervisor Stuart Aitken, who proved to be an inexhaustible source of statistical expertise.

On a personal note, it would be remiss of me not to thank my mother for her support during the past three years (not to mention the prior twenty-two). A special thank you to Nika for the pep talks and good times along the way.

Finally I'd like to acknowledge the open source community at large, members of which have used their time and knowledge to help build the countless tools used throughout this thesis and far beyond. From the Linux kernel to the R programming language, Emacs to Inkscape, and LaTeX in which this document is written — everything in this thesis was made possible by the oft-unacknowledged contributors to open source projects.

ABSTRACT

Recent technological advances underpinned by high throughput sequencing have given new insights into the three-dimensional structure of mammalian genomes. Chromatin conformation assays have been the critical development in this area, particularly the Hi-C method which ascertains genome-wide patterns of intra and inter-chromosomal contacts. However many open questions remain concerning the functional relevance of such higher order structure, the extent to which it varies, and how it relates to other features of the genomic and epigenomic landscape.

Current knowledge of nuclear architecture describes a hierarchical organisation ranging from small loops between individual loci, to megabase-sized self-interacting topological domains (TADs), encompassed within large multimegabase chromosome compartments. In parallel with the discovery of these strata, the ENCODE project has generated vast amounts of data through ChIP-seq, RNA-seq and other assays applied to a wide variety of cell types, forming a comprehensive bioinformatics resource.

In this work we combine Hi-C datasets describing physical genomic contacts with a large and diverse array of chromatin features derived at a much finer scale in the same mammalian cell types. These features include levels of bound transcription factors, histone modifications and expression data. These data are then integrated in a statistically rigorous way, through a predictive modelling framework from the machine learning field. These studies were extended, within a collaborative project, to encompass a dataset of matched Hi-C and expression data collected over a murine neural differentiation timecourse.

We compare higher order chromatin organisation across a variety of human cell types and find pervasive conservation of chromatin organisation at multiple scales. We also identify structurally variable regions between cell types, that are rich in active enhancers and contain loci of known cell-type specific function. We show that broad aspects of higher order chromatin organisation, such as nuclear compartment domains, can be accurately predicted in a variety of human cell types, using models based upon underlying chromatin features. We dissect these quantitative models and find them to be generalisable to novel cell types, presumably reflecting fundamental biological rules linking compartments with key activating and repressive signals. These models describe the strong interconnectedness between locus-level patterns of local histone modifications and bound factors, on the order of hundreds or thousands of basepairs, with much broader compartmentalisation of large, multi-megabase chromosomal regions.

Finally, boundary regions are investigated in terms of chromatin features and co-localisation with other known nuclear structures, such as association with the nuclear lamina. We find boundary complexity to vary between cell types and link TAD aggregations to previously described lamina-associated domains, as well as exploring the concept of meta-boundaries that span multiple levels of organisation. Together these analyses lend quantitative evidence to a model of higher order genome organisation that is largely stable between cell types, but can selectively vary locally, based on the activation or repression of key loci.

LAY SUMMARY

Each human cell contains DNA that would extend for two metres if fully straightened. Instead, this same length of DNA is highly compacted into micrometre-sized cell nuclei. Recently experimental methods such as Hi-C have been developed which allow the inspection of this folded state, generating counts of how frequently chromosomal regions are interacting with each other. These counts can be statistically analysed to reveal different levels of structures, including loops between two distant locations, knot-like domains of self-interacting regions, and broad stretches of mostly active or inactive regions.

In this work, we bring together Hi-C datasets from several different publications and combine these with a large number of chromatin datasets that quantify, for example, levels different DNA-binding proteins as well as modifications to DNA packing histone proteins. We used these datasets to build predictive models of active and inactive states across each human chromosome in three different cell types, and achieved high predictive accuracy. We then compare and contrast these models, and use them to identify the key features which define active and inactive states.

We also analyse the boundaries between domains and compare these across cell types. We find the domains themselves are highly conserved between cell types, but observe different chromatin features marking domain boundaries. Further collaborative work involved analysis of boundaries from Hi-C data taken over successive time points, where boundary markings were found to persist as cells differentiate from stem cells.

Overall we find the three-dimensional DNA structures within cells are highly similar even between human embryonic stem cells and cells derived from blood. Where there are differences, these areas tend to highlight biological activity specific to that cell type.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3C	Chromosome conformation capture (derivatives: 4C, 5C, Hi-C)
AUROC	Area under the receiver operating characteristic
CAGE	Cap analysis of gene expression
ChIP-seq	Chromatin immunoprecipitation following by high-throughput sequencing
DI	Directionality index
ENCODE	The encyclopaedia of DNA elements
ESC	Embryonic stem cell
FDR	False discovery rate
FISH	Fluorescent *in-situ* hybridisation
GC	Guanine and cytosine (content of a DNA sequence)
GO	Gene ontology
Hi-C	Genome-wide 3C experiment using high-throughput sequencing
HMM	Hidden Markov Model
ICE	Iterative correction and eigenvector expansion
IF	Interaction frequency
LAD	Lamina associated domain
MAD	Median absolute deviation
MSE	Mean squared error
NAD	Nucleolus associated domain
OOB	Out-of-bag
PCC	Pearson correlation coefficient
PCR	Polymerase chain reaction
PLS	Partial least squares

RF Random Forest

RMSE Root mean-squared error

RVS Regions of variable structure

SHH Sonic Hedgehog

TAD Topologically associating domain

TSS Transcription start site

ZRS Zone of polarising activity regulatory sequence

PUBLISHED MATERIAL

Some materials contained in this thesis have previously been published in:

> Moore BL, Aitken S and Semple CA (2015) Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biology*, **16**:100.
> doi:10.1186/s13059-015-0661-x

Parts of Section **??** are used in a manuscript which is under review at the time of writing:

> Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, Barbieri M, Moore BL, Aitken S, Morris KJ, Itoh M, Kawaji H, Jaeger I, Hayashizaki Y, Carninci P, Forrest ARR, FANTOM Consortium, Semple CA, Dostie J, Pombo A, Nicodemi M (2015) Hierarchical organization of chromosome folding and its re-organization underlies transcriptional changes in cellular differentiation. *Under review*

1 INTRODUCTION

1.1 GENOME ORGANISATION

It is often 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 on the order of micrometers (μm). This is achieved through a complex organisational hierarchy, ranging from how chromosomes are arranged in the nucleus in territories, to chromatin interactions with the nucleolus or nuclear periphery, down to how DNA is wrapped around nucleosomes (for recent reviews, see **? ?**). While the biophysics of the latter level of organisation may be well understood, more broadly there is still much to learn 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.[? ?] 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 of DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[?] was the first sequencing-based method of assaying nuclear architecture. The 3C method uses formaldehyde to cross-link nuclear proteins in place, trapping genomic regions that were physically co-located through bound proteins, then a frequent-cutter restric-

tion enzyme shears the sample into DNA fragments. Next, under dilute conditions, these fragments are ligated together. The dilute conditions favour ligations between fixed fragments, with the aim of generating hybrid fragments from 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 the 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.

Rapid advancements in sequencing technology allowed the original 3C method to be further developed, first through microarrays, 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 stage in the evolution of the 3C method was an all-versus-all assay, capable of assaying pairwise interaction frequencies between all restriction fragments of a genome. Such an assay was published by Lieberman Aiden *et al.*[?] and named Hi-C (Fig. **??**). The Hi-C method added biotin tagging to pull-down only ligated fragments for sequencing. At the time of the assay's publication, resolution of Hi-C data for analysis was limited by sequencing depth—of particular concern given the enormous pairwise interaction space between all restriction fragments produced by a 6-cutter enzyme—but the falling costs of sequencing and proven utility of the assay meant subsequent Hi-C papers incrementally upped their sequencing depth, culminating at the time of writing to the point where analyses are starting to 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 contain several inherent biases. Restriction fragment properties, such as their length, GC content and mappability, were confounding interaction frequency estimates and therefore required normalisation 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.

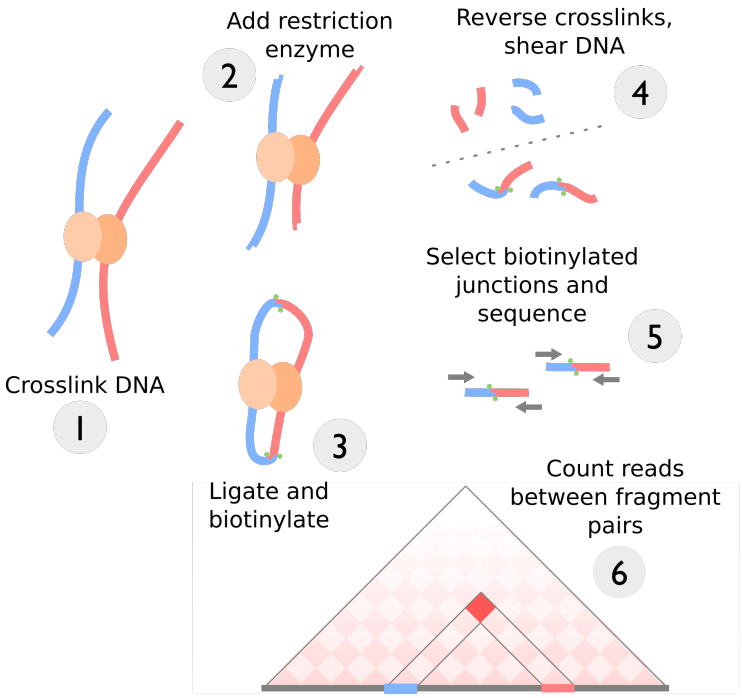

Figure 1: Steps in the Hi-C assay. Schematic of the Hi-C experimental procedure as described in ?]

In situ Hi-C is another, more recent refinement of the Hi-C method from some of those who developed the original Hi-C method.[?] In contrast to TCC, fixation and ligation now happen in place within intact cell nuclei. The observed improvements with this *in situ* procedure, however, are similar: interactions are assayed with greatly reduced noise and again many fewer *trans* contacts are reported.[?]

Hi-C and the variants introduced thus far are population-level assays, reporting summed interaction counts over a cell population. 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 has 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 of this Hi-C variant is that a single restriction fragment can ligate to at most one other fragment per experiment, meaning even if 100% yield were to be achieved, any $n \times n$ restriction fragment interaction matrix could have at most $\frac{n}{2}$ nonzero entries; 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, and 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.[?] When combined with the observations of TCC and *in situ* Hi-C, 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 an altogether different Hi-C variant which attempts to address the resolution problems associated with the standard assay by enriching for functional interactions using *a priori* selection for loci of interest.[?] Indeed, a suggestion in the original Hi-C paper was that resolution could be improved by either increased sequencing or using hybrid capture.[?] Since the release of Capture-C, more Hi-C variants with a target enrichment step have been developed, including Capture Hi-C (CHi-C)[?] and HiCap.[?] These methods have been applied to genome-wide target sets (e.g. CHi-C assayed 22,000 human promoters[?]) and so it could be said that they are to Hi-C as exome-capture is to whole-genome sequencing, in the contexts of chromosome conformation capture and variant discovery respectively.

1.1.3 Chromosome compartments

In the paper describing the Hi-C technique, **?**] described low-resolution structures they name "A" and "B" nuclear compartments. These are genomic 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 often lamina-associated regions, with little transcription and repressive histone modifications such as H3k9me3.[??]

These A and B compartments were identified through a continuous principle component eigenvector profile, derived from normalised Hi-C contact matrices[?] (Fig. **??**). This approach can be intuitively understood as formulated by **?**]:

1. A tartan pattern on normalised Hi-C matrices indicates two preferentially-contacting and exclusive compartments (Fig. **??**).

2. Assume a function (c) that maps a given genomic bin to its compartment, using a positive number for compartment A and a negative for compartment B.

3. The interaction frequency between bins i and j is thus $c(i) \cdot c(j)$. (Note that this rule alone is sufficient to generate a tartan pattern: if i and j are in the same compartment, the product will be positive, and for compartments of opposing type, negative.[?])

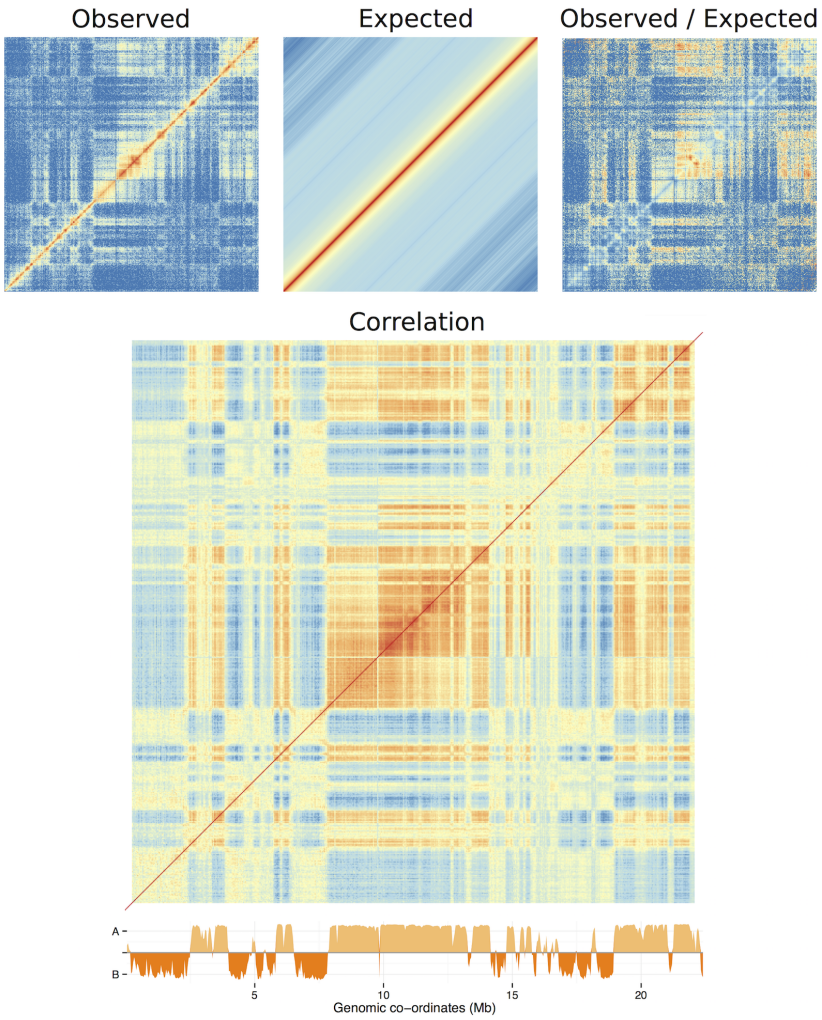

Figure 2: **Derivation of A/B an compartment profile from Hi-C data.** Intrachromosomal observed interaction frequencies (O) are averaged along super-diagonals to give a distance-normalised expected matrix (E). The Pearson correlation of the O/E matrix then can undergo eigenvector expansion; in most cases eigenvector **v** with the largest eigenvalue, λ, then reflects A/B compartmentalisation.[?] Matrices are coloured from blue (lowest values), through yellow, to highest values in red.

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 compartment identity, hence degrees of compaction or activity.[? ?]

1.1.4 Topologically associating domains

The falling cost of high-throughput sequencing has enabled increasingly deep sequencing of Hi-C experiments. Sequencing depth is the main resolution-limiting resource for this assay; in order to increase the analysis resolution while maintaining the same level of coverage requires an exponential increase in the total amount of sequencing.[? ?] Nevertheless such deep sequencing has recently been achieved in a handful of landmark studies.

In experiments totalling around two billion total sequencing reads, ?] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. At the same time, ?] 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 uncover what are now known as "topologically associating domains" (or TADs), observable as off-diagonal blocks in a contact map which exhibit 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 **??**). TADs have since been reported in a variety of metazoan organisms including dog,[?] *Drosophila*[? ?] and *C. elegans*[?] yet comparable structures are not found in higher plants such as *Arabidopsis*[? ?] or in yeast.[? ?]

?] 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 bias, reasoning that domain boundaries are present when a strongly upstream biased region is adjacent to a region of opposite bias (Fig. **??**). These boundaries themselves were investigated and found to show suggestive functional enrichments for DNA binding proteins including CTCF, long thought to act as an insulator of chromatin state (Section **??**). Deletion of a CTCF site has been found to disrupt the corresponding TAD border, while removal of some other enriched factors had little effect.[? ? ?] ?] also performed some comparative analysis, reporting large and significant overlap of domain boundary positions both within species and between human and mouse.

Since then, several studies have investigated the functional implications of TADs. A proposed biological explanation is that TADs delimit functional contacts, such as those between enhancers and promoters, and so could inhibit spurious contacts with

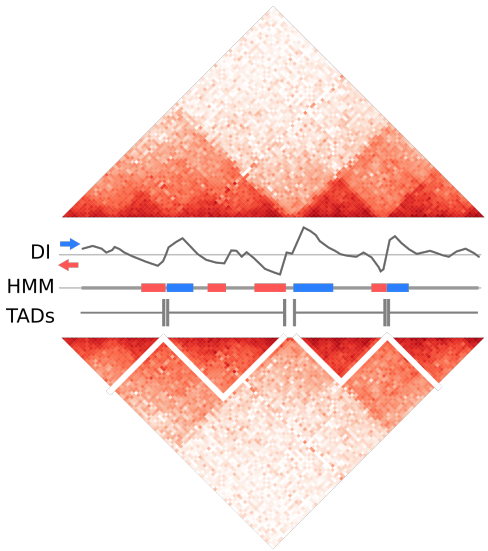

Figure 3: Dixon *et al.* pipeline for calling topologically associating domains. First a directionality index (DI) is calculated for each bin based on the ratio of upstream:downstream contacts. Secondly a hidden Markov model (HMM) is used to infer the most likely state sequence that emitted the DI variable. 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 these are labelled border regions up to a size threshold of 400 kb, whereafter those regions are unclassified.[?] Additional details are given in Methods **??**.

other nearby genetic elements.[? ?] Moreover, hormonal treatment of human breast cancer cells reported coordinated expression responses within TADs, suggesting they also function as domains of transcriptional co-regulation or "regulons".[?] 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.[?] At the time of writing, the functional significance of TADs is the subject of ongoing debate within the chromatin biology field.

1.1.5 Other proposed structures

Since the discovery of TADs, multiple publications have proposed either complementary or altogether different classes of chromatin domains. For example, **?**] developed a tuneable algorithm which identifies "alternative topological domains". The authors use dynamic programming to search for an optimal 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.[?] These multiscale domains were found to be smaller, on average, than those previously reported by **?**], even when applied to the same Hi-C experimental data (with a mean size of 200 kb as opposed to $\approx$ 1 Mb). However the domains of **?**] 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 cell chromosomes found a similarly hierarchical organisation of physical domains, and also was able to relate these to "epigenomic domains" which exhibited specific sets of enrichment signatures representing active, null, polycomb-associated and telomeric regions.[?] This study provided evidence for the first time that contact domains are linked with average epigenomic enrichments.

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 (Section **??**).[?] This domain size is close to those of **?**] (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. Additionally, these sub-compartment types exhibit average epigenomic enrichments along the lines of those reported in *Drosophila* by **?**] and so potentially provide an overarching concept that connects the above-mentioned TADs, alternative domains and epigenomic 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 thus potentially describe the polymer dynamics of chromatin folding.

1.2.1 Fractal globule

?] tested a number of theoretical models of genome folding to see which best explained the observed power-law scaling between distance and interaction frequency ($\mathrm{IF} = \frac{1}{dist^{-\alpha}}$ where $\alpha \approx 1.08$). The authors sought to distinguish between two models of genome organisation: the previously-described "fractal globule"[? ?] and a less-structured "equilibrium globule" (Fig. **??**). The study found that a theoretical fractal globule,

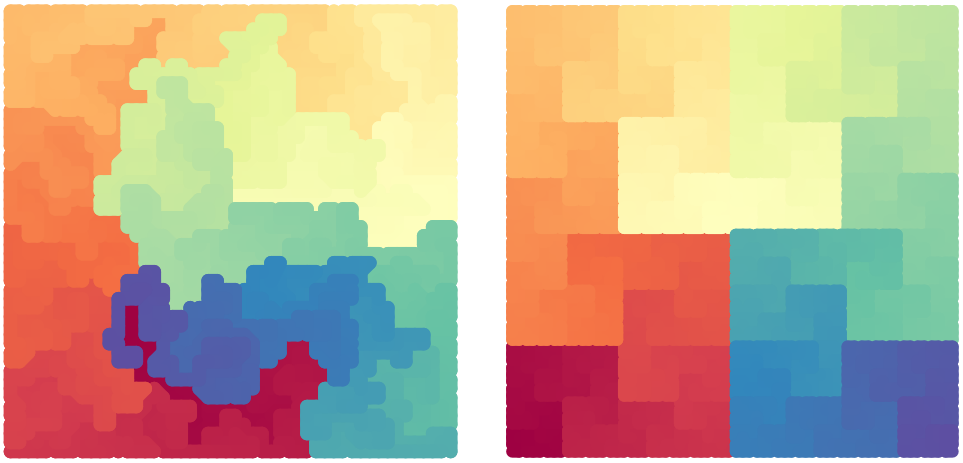

Figure 4: Comparison of theoretical models of chromatin folding. Two theoretical models of chromatin folding are shown simulated along a 2D polymer, coloured from start to finish as blue–green–red. An equilibrium globule is represented by a Hamiltonian path through a grid network (*left*) and is compared to a fractal globule model, here represented by a Hilbert curve (*right*).

embodying scale-independent and self-similar aggregate folding, showed a better fit to 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 idea formed the basis for **?**] to develop the "strings and binders switch" (SBS) model, where the authors simulated polymer folding in the presence of DNA binding factors, such as the known genome architectural protein CTCF (Section **??**). The SBS model was developed in an attempt to consolidate global Hi-C measures of contact scaling with those values from C-based experiments on smaller regions and FISH studies, which report a range of scaling parameters. The authors also explore the different observed values of α (the coefficient describing the power law relationship between interaction frequency and genomic distance, introduced in the previous section) across cell lines and even chromosomes, and find

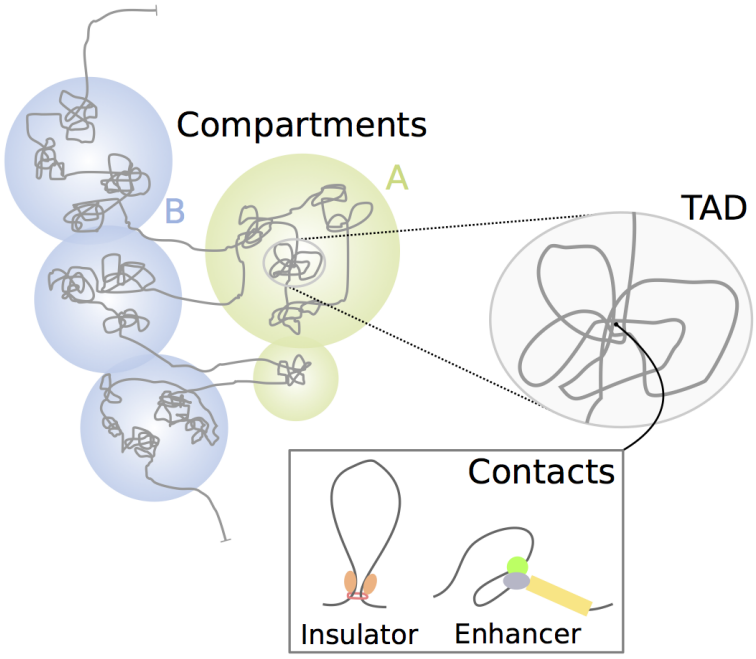

Figure 5: Levels of higher order chromatin organisation. Cartoon showing how functional contacts, such as loops between bound CTCF insulators (Section **??**), occur within TADs (Section **??**) which in turn are found within A or B compartments (Section **??**).

that their mechanistic model can explain each case using variable concentrations of binders which cause phase-switching between accessible and compacted chromatin, with a fractal globule organisation existing at the phase transition boundary.[?]

This SBS model achieves broad explanatory power for a range of observed power law coefficients (α), and does so 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.[?]

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[? ?] (Fig. **??**). A model region, the β-globin locus and its locus control region (LCR) located 40-80 kb away,[?] has been studied since the early 1980s,[? ? ? ?] and is an interesting example of a well-characterised looping event. Current knowledge suggests the β-globin locus forms loops with the multiple distal *cis*-enhancer elements that make up the LCR, forming an active chromatin hub (ACH).[?] Within such a hub, regulatory signals could be efficiently integrated to dictate the overall activity of the target locus.[? ?] It is now thought that the majority of active promoters are engaged with multiple, often cell type specific, regulatory looping events.[? ?]

A notable component of many long-range looping events is the CCCTC-binding transcription factor (CTCF),[? ? ?] already mentioned as a component of TAD boundaries (Section **??**) and as a proposed looping factor in the SBS model (Section **??**). CTCF is strongly conserved in higher eukaryotes,[?] ubiquitously expressed and embryonic lethal, but it is not tied to a single biological function — instead CTCF has been described as a "multivalent factor",[?] 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 stabilise the aforementioned loops linking the β-globin locus with its distal LCR.[?] This looping role, potentially undertaken in combination with other architectural proteins such as Mediator and cohesin,[? ?] 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.[?] 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.[?]

A recent Hi-C study by **?**] 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. **?**] also report a 6-fold overall increase in expression when comparing those promoters participating in a looping event with those that do not. Furthermore, 86% of these loops involved CTCF bound regions, with roughly the same overlapping proportion involving cohesin subunits RAD21 and SMC3. The authors thus propose that a CTCF-binding motif can act as an "anchor", which can then be bound by a transitive complex of other architectural proteins.[?] A majority of these loops (65%) also demarcated a topological domain, and at much higher resolution than previously observed (Section **??**). Another striking finding of this research was that loops almost always occur in between bound CTCF motifs with a convergent orientation,[?] though questions remain over why this should be the case, especially when considering the interactions of a flexible polymer in 3-D solution.[?]

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 ChIP-seq peaks were found to occur at TAD boundaries in human and mouse cells[?] and similarly around 25% of TAD borders had no observable CTCF binding.[?] 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.[?]

1.3 CRITICISMS OF C-METHODS

C-methods are a relatively new and rapidly 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 **??**), the Hi-C assay typically takes place using chromatin from a cell population (though proof-of-concept single-cell experiments have been reported[?]). An obvious consideration, then, is that interaction counts reflect the average over a large number of cells, often including unsynchronised populations at different stages of the cell cycle.[?] Given evidence that, while the interphase chromosomes exhibit cell-to-cell variability, the mitotic state is much more static,[? ?] one might expect even a small proportion of dividing cells to add a detectable amount of 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 in a variety of ways, such as using haploid cell lines (e.g. **?**) or through haplotype phasing using detectable sequence differences with deep sequencing, or by focusing on an allosomal region (e.g. **?**). An altogether different and inventive solution is to use the inherent proximity-ligation information produced by C-methods to discriminate haplotypes,[?] an idea since extended to deconvolution problems in metagenomics.[? ?]

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.[?] This results in on the order of 10 million restriction fragments with a total pairwise interaction space of 10^{12}.[?] 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 using a four-cutter restriction enzyme, for example MboI,[?] 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[?]), 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.[?]

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",[?] 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 the assumption that this proxy is accurate, that methods have been developed that use a known FISH distance to then calibrate genome-wide Hi-C distances,[?] however it need not be the case that population-level interaction frequencies capture physical distance.[?] 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.[?]

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 the nuclear lamina or nucleolus, for example, rather than any specific functional relationship between the two loci.[?] In addition, a functional enhancer–promoter interaction will necessarily constrain the contacts of other nearby regions, potentially causing highly-reproducible "bystander interactions"[?] 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,[?] 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 **? ?**). 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 datasets have empowered research groups worldwide to tackle complex biological questions through a variety of machine learning approaches.

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.[? ? ?] 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.[?] 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[?] coupled with statistical quality control[? ? ?] to ensure data is comparable between different data producers and is of consistently-high accuracy. ENCODE spin-off projects have also aimed to build similar genomics resources for mouse[?] and, more recently, *Drosophila* and *C. elegans*.[?] 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.[?]

Data generated by ENCODE consortium members also has a proven utility in modelling techniques based on machine learning. 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.[?] Independently another chromatin state predictor named `ChromHMM` was developed.[? ?] As the name suggests, this approach instead used multivariate hidden Markov models (HMMs) 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,[?] but more recently these two methods were combined to call a consensus set of just 7 chromatin states.[?] Since their publication, a study was able to experimentally validate many of these state predictions.[?] 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,[?] gene regulation,[?] cell cycle-associated genes[?] and enhancer identification[?] to name but a few. One such study in particular, that of **?**], 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.

1.5 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 human cell types?
2. Can we predict higher order chromatin structure from locus-level features?
3. How do the characteristics of boundaries demarcating 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 **??**) and employ machine learning

techniques and other statistical analyses to explore how these locus-level features relate to higher order chromatin structure.

2 METHODS

2.1 HI-C DATA

2.1.1 Mapping

Raw Hi-C reads were downloaded from published datasets (Table **??**) through the Gene Expression Omnibus (GEO)[?] or the Short Read Archive (SRA)[?] with identifiers: GSE35156 (H1 hESC), GSE18199 (K562) and SRX030113 (GM12878). These paired reads were mapped independently to the human reference genome (build `hg19` / `GRCh37`.

Mapping was performed using the `hiclib` python library[?] and `bowtie2`[?] with the `--very-sensitive` flag. An iterative mapping approach was used to maximise the number of aligning fragments.[?] Each fragment end was aligned first using short terminal sub-sequences. Those unmapped or with ambiguous mapping were then taken forward into the next iteration and extended until the entire fragment end had been aligned. Those remaining pairs with one or more unmapped ends were discarded. This approach is designed to maximise uniquely-alignable fragment ends, while avoiding mismappings from reads that cross a restriction fragment junction.[?]

2.1.2 Filtering

After mapping, interactions are first aggregated into restriction fragments then by regular binning at various resolutions (particularly 40 kb, 100 kb and 1 Mb). Several filters were applied at this stage, with the following cases removed:[? ?]

- Reads directly adjacent to a restriction enzyme site (within 5 bp)
- Identical read pairs (presumed PCR duplicates)
- Very large restriction fragments ($>$ 100 kb) which are likely from a repetitive or poorly-assembled region

Table 1: Public Hi-C data used in this work.

Cell line	Total reads	Accession	Citation
Gm12878	31×10^6	SRX030113	**?**
H1 hESC	331×10^6	GSE35156	**?**
K562	36×10^6	GSE18199	**?**

- Extremely over-represented fragments (top .05%) which may throw-off the subsequent derivation of principle component eigenvectors

2.1.3 Correction

Iterative correction and eigenvector expansion (ICE) is an approach to normalisation and processing Hi-C data and is implemented as part of the the `hiclib` library written in `python`.[?] The iterative correction algorithm performs matrix balancing with the aim of generating a doubly stochastic matrix from raw interaction counts.[?] That is, such that symmetric matrix **A** has both row and columns of equal sum. In practice, this effectively enforces "equal visibility" of each fragment, correcting for previously-described biases in interaction recovery such as GC-content and fragment length[?] but without explicitly modelling these latent variables.

This correction procedure thus converts actual interaction counts into normalised interaction frequencies (IF), and to relative rather than absolute quantities. Scaling of IFs permits comparison of Hi-C experiments with very different sequencing depths (as is the case in this work, see Table **??**). Despite differences in the levels of sequencing, otherwise the experimental methods underlying the produced Hi-C data were similar: the HindIII restriction enzyme was used in each case and the Hi-C protocol was largely unchanged (for example, we did not consider data from Hi-C variants such as tethered conformation capture[?] and *in-situ* Hi-C[?]).

2.1.4 Eigenvector calculation

Additional functionality provided by ICE is the eigenvector expansion of normalised contact maps. Eigenvectors from observed/expected matrices were chosen for consistency with Lieberman Aiden *et al.*,[?] as opposed to the related eigenvectors calculated in Imakaev *et al.*[?] from the corrected maps alone. Briefly, observed contacts (O) are divided by an expected matrix (E) which is generated by averaging the super- and sub-diagonals of the O matrix. That is, the E matrix gives the expected value of interactions at a given distance, hence the O/E matrix is a normalised contact map without the distance decay seen in raw Hi-C contact matrices. Examples of these maps are shown in Figure **??** (Section **??**).

Importantly, for the eigenvector expansion step the first two principle components (PCs) were calculated, and that with the highest absolute Spearman correlation with GC content is taken to reflect A/B compartmentalisation. PC eigenvectors were then orientated to positively correlate with GC, ensuring positive values reflected A compartments and negative values B compartments. Another subtlety is the calculation of eigenvectors per chromosome arm as opposed to per chromosome, this prevents issues with some meta- and submetacentric chromosomes where the first

principle component indicated chromosome arms.[? ?] Eigenvector expansion was performed on both 1 Mb and 100 kb matrices, below these resolutions results became less stable, and besides it has been shown that eigenvectors at higher resolution — when they do indeed capture A/B compartmentalisation — add little, if any, additional information.[?]

2.1.5 TAD calling

TADs were called using the software provided in **?**] and their recommended parameters. This method is introduced in Section **??** (see also Fig. **??**) but will be described here in greater detail.

The TAD calling algorithm is a multi-stage process. Firstly, a statistic called the "directionality index" (DI) is calculated for each bin.[?] The equation for calculating the DI of a given bin is shown (Eqn. **??**), where U represents the sum of reads mapped up to 2 Mb upstream of a given 40 kb bin, and D likewise for downstream contacts. Here E is the expected number of downstream or upstream contacts (equal under the null hypothesis), hence is $E = \frac{U+D}{2}$.

$$\mathrm{DI} = \left(\frac{U-D}{|U-D|}\right)\left(\frac{(D-E)^2}{E} + \frac{(U-E)^2}{E}\right) \tag{1}$$

Equation **??** can be intuitively understood as first determining the direction of the bias (the sign is given by $\frac{U-D}{|U-D|}$) and then calculating the extent of the bias (with $\frac{(D-E)^2}{E} + \frac{(U-E)^2}{E}$ being akin to a χ^2-type statistic).[?]

This DI metric could be used as-is to call domains, as peaks of downstream contacts culminating in a peak of upstream contacts delineate self-interacting domains. However, **?**] instead use a hidden Markov model (HMM; Section **??**) in a manner similar to the strategy we later employed to call compartments (Section **??**).

Here, the DI metric is considered a noisy observation emitted by an unobserved underlying three-state sequence of upstream, downstream or no- directional contact bias.[?] The HMM was fitted to each chromosome with between 1 and 20 Gaussian mixtures allowed per state, however in some cases the expectation-maximisation (EM) algorithm used to parameterise these hidden states failed to converge; such cases were ignored. The Akaike information criterion (AIC) was used to select the optimal number of mixtures (in practice, we found 5–10 were selected).

Finally, given a fully-specified HMM we can calculate the posterior probability of a given state in a specific bin, using the forward-backward algorithm and given its observed data and preceding state sequence. **?**] enforce the heuristic that regions are only classified as downstream- or upstream-biased if the state is called for two consecutive bins, or if a single bin has an especially high posterior probability ($\geq .99$). Domains are called from this state sequence and run from an initial downstream-

biased bin through to the last in a run of ≥ 2 of upstream biased states. This procedure was implemented by **?**[] in `Matlab`.

2.2 ENCODE FEATURES

Genome-wide ChIP-seq datasets for: 22 DNA binding proteins and 10 histone marks were made available by the ENCODE consortium[? ?] along with DNase I hypersensitivity and H2A.z occupancy, for each of the Tier 1 ENCODE cell lines used in this work: H1 hESC, K562 and GM12878. These data were pre-processed using MACSv2[?] to produce signal fold-change relative to input chromatin. In most cases a paired input control was generated by the same laboratory for each ChIP-seq experiment.[?] GC content was also calculated over the same genomic intervals and used in the featureset to give 35 total inputs (Table **??**).

Table 2: ChIP-seq and other public datasets used in this work.

Histone modifications	DNA binding proteins	Other
H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1	ATF3, CEBPB, CHD1, CHD2, CMYC, CTCF, EGR1, EZH2, GABP, JUND, MAX, MXI1, NRSF, POL2, P300, RAD21, SIX5, SP1, TAF1, TBP, YY1, ZNF143	DNase, GC content, H2A.Z

In the analysis of boundaries (Chapter **??**), we also use a measure of sequence conservation in the form of Genomic Evolutionary Rate Profiling (GERP) scores. This measure uses rejected substitutions to assign conservation scores to each genomic site based on a multiple alignment of 35 mammalian genomes.[? ?] A rule of thumb is that a GERP score $\gtrapprox 2$ indicates an evolutionarily-constrained site.[?]

2.2.1 Clustering input features

To quantify collinearity of input features, correlation matrices of genome-wide vectors of input feature measures were hierarchically clustered. The "significance" of observed clustering was assessed using sub- and super-sampled bootstrapping, with stable clusters across sample sizes deemed significant, as implemented in the `pvclust` R package.[?]

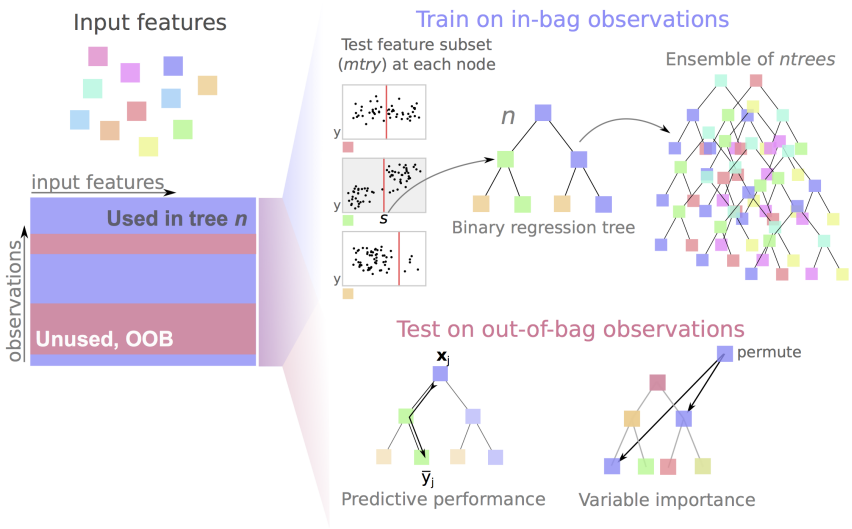

Figure 6: Random Forests overview. Random Forests are an ensemble of bagged, decorrelated classification or regression trees first described by Breiman.[?] These schematics describe how Random Forests are constructed (*upper*) as well as how measures of predictive accuracy and variable importance can be calculated using out-of-bag (OOB) data.

2.3 MODELLING COMPARTMENT EIGENVECTORS

2.3.1 Random Forest

Random Forest (RF) regression,[?] was used as implemented in the R package `randomForest`.[?] The RF algorithm (Fig. **??**) makes use of a collective of regression trees (size *ntrees*), each built from a bootstrapped sample of the training set. In growing each tree, a small number of variables (*mtry*) is tested at each bifurcation node, and that which minimises the variance in child node subsets is selected at an optimal threshold. Having trained a group of trees, these can then be used as predictive tools by inputting a vector of features to each tree and averaging the output leaf node values across the forest. RF regression was used as it is known to be one of the most powerful regression methods developed to date,[? ?] typically providing low bias and low variance predictions without the need for variable selection.[? ?]

Additionally the RF method is an example of "algorithmic modelling"[?] in that it makes no assumptions about the underlying data model. Of the few user-facing parameters, the number of features to test at each node (*mtry*) was set to $\frac{n}{3}$ (where n is the number of input features) and the number of trees in the forest (*ntrees*) was chosen as 200. These parameters are known to be insensitive over a broad range of values,[? ?] as shown in Figure **??**.

Variable importance within Random Forest regression models was measured using mean decrease in accuracy in the out-of-bag (OOB) sample. This represents the average difference (over the forest) between the accuracy of a tree with permuted and

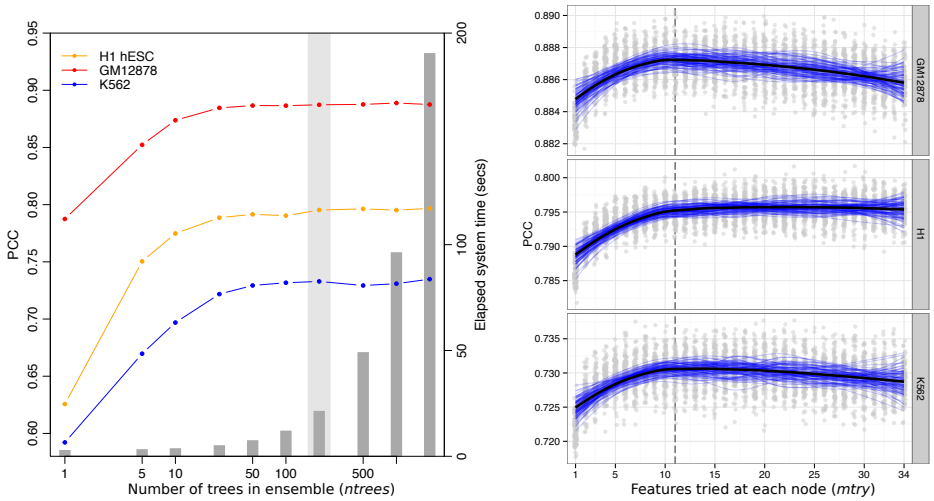

Figure 7: Random Forest parameters are largely insensitive. Two user-facing Random Forest paramters are known to be insensitive over a broad range.[?] Optimisations for *ntrees* (the number of trees in the forests) and *mtry* (the number of features tested at each node) are shown for three different models, with typical default values of 200 trees and $\frac{1}{3}$ of input variables highlighted.

unpermuted versions of a given variable (Fig. **??**), in units of mean squared error (MSE).[? ?]

2.3.2 Model performance

The effectiveness of the RF modelling approach used to predict compartment eigenvectors (Section **??**) was measured by four different metrics. Prediction accuracy was assessed by the Pearson correlation coefficient between the OOB predictions and observed eigenvectors, and the root mean-squared error (RMSE) of the same data. Classification error, where predictions were thresholded into $A \geq 0; B < 0$, was also calculated using accuracy (% correct classifications or True Positives) and area under the receiver operating characteristic (AUROC) curve. Together these give a comprehensive overview of the model performance, both in terms of regression accuracy of the continuous eigenvector, and in how that same model could be used to label discrete chromatin compartments.

For cross-application of cell type specific models, a single Random Forest regression model was learned from all 1 Mb bins for a given cell type. This was then used to predict compartment eigenvectors for all bins from each of the other two cell types (Section **??**).

To test the sensitivity of the models to resolution, we also applied cell-type specific models learnt at 1 Mb resolution to input features binned at 100 kb (Section **??**). This was done by training a Random Forest regression model on all available 1 Mb bins

in a given cell type, then applying that model to the prediction of all compartment eigenvectors derived at 100 kb. Model performance was then assessed as above, with the caveat that here the test set represents a higher-resolution window onto the original training set, therefore we might expect this to inflate the measures of generalisation error.

2.3.3 Hidden Markov models

Hidden Markov models (HMMs) were used both for calling TADs (Methods **??**) and for identifying chromosome compartments (Methods **??**, discussed in Section **??**). Here we briefly introduce the HMM framework in general terms.

HMMs are widely used in computational biology, and have been called "the Legos of computational sequence analysis"[?] due to their wide applicability. HMMs provide a probabilistic modelling framework to explore any system that can be reduced to a 1D state sequence.[? ?] In the discrete case, each state in this sequence is capable of "emitting" one of a number of symbols, each with its own probability. In the continuous case, used in this work, a state instead has an associated emission distribution which can be a simple univariate Gaussian or a more complex mixture model. After an emission, each state has a number of "transition" probabilities, where the sequence can either change to another state or remain in the same state.

A typical representation of an HMM is shown in Equation **??**, where hidden states (θ) emit a sequence of observed states (y).[?]

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & y_{i-1} & \longrightarrow & y_i & \longrightarrow & y_{i+1} & \longrightarrow & \cdots \\ & & \uparrow & & \uparrow & & \uparrow & & \\ \cdots & \longrightarrow & \theta_{i-1} & \longrightarrow & \theta_I & \longrightarrow & \theta_{i+1} & \longrightarrow & \cdots \end{array} \tag{2}$$

HMM state emission and transition parameters can be learned from a sequence of observations via the Baum-Welch algorithm,[?] a special case of the Expectation-Maximisation algorithm applied to HMMs. Following initialisation of transition and emission matrices (this can be random), the Baum-Welch algorithm first performs the E-step, calculating the expected number of transitions and emissions via a Forward-Backward procedure, then the M-step re-estimates transition and emission parameters based on these expected values. These two steps repeat until the HMM parameters converge to within a set tolerance.

Given a fully-specified HMM, the Viterbi algorithm can be applied to find the most probable state sequence given a sequence of observations.[?] Using the notation of Equation **??**, the Viterbi algorithm finds $\hat{\theta} = \text{argmax}_\theta \Pr(y|\theta)$ using dynamic programming.[?]

2.3.4 Stepwise regression

Stepwise regression is a form of model selection applied to multiple linear regression. This simple approach starts with a complete model and serially removes and/or adds variables, then calculates a metric (here we use the Bayesian information criterion, BIC) which weighs the the model likelihood against model complexity. Alternatively, the procedure can be run in the opposite direction and build up a model starting from scratch. In either case, the variable inclusion or exclusion process is iterated until the metric reaches a (local) minimum, thereby generating a parsimonious model which should be less prone to overfitting. Stepwise regression also aids interpretation by selecting representative features from collinear clusters.[?]

It should be noted that despite its continued widespread usage, several statistical issues have been identified with the stepwise procedure for model selection.[? ?]

2.3.5 LASSO

The least absolute shrinkage and selection operator (LASSO) is a form of ℓ_1 regularisation that penalises the sum of absolute values of standardised regression coefficients. By penalising absolute values and sums, rather than squared values as in ℓ_2 regularisation (Ridge regression, for example), coefficients can be shrunk to 0 thereby removing terms from the model. Thus the LASSO combines coefficient shrinkage of techniques like Ridge regression with a type of feature selection by promoting model sparsity.[? ?]

Simply put, the LASSO minimises the sum of squared errors subject to a tuneable constraint on the sum total of absolute model coefficients. In equation form, we are fitting a simple linear model:

$$\begin{aligned} \hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \\ \text{or } \hat{y} = \boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{3}$$

We then wish to find that $\boldsymbol{\beta}$ which minimises $\sum_{j=1}^{n}(\hat{y}_j - y_j)^2$ while at the same time satisfying the inequality:

$$\sum_{i=1}^{p} |\beta_i| \leq c \tag{4}$$

Where here c represents a tuneable parameter inversely proportional to the level of regularisation imposed on the model. It can be seen, for example, that if c is set to the sum of the coefficients fit by ordinary least squares, the LASSO solution will be equivalent. Equation **??** can be contrasted with Ridge regression, where the same in inequality instead constrains $\sum_{i=1}^{p} \beta_j^2$.

Formally, the LASSO problem has been expressed as:[?]

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=i}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{5}$$

This formulation introduces the tuning parameter λ, which translates β coefficients such that larger values of λ place stronger constraints on the coefficient total, and thus encourages greater shrinkage and model sparsity (Eqn. **??**).

In this thesis, we used the `glmnet` R package to fit LASSO models.[? ?] In order to select λ, we use a 10-fold cross-validation approach on a separate held-out training set. We chose that λ which produced a mean cross-validated error within 1 standard error of the minimum, thus favouring a slightly sparser model than the global minimimum.

2.3.6 Other modelling approaches

Linear regression was used as a baseline for comparison with more complicated approaches such as Random Forest (Section **??**). If the modelling accuracy achieved with Random Forest regression could be matched by simple multiple linear regression, the latter could be preferable as a faster and more interpretable modelling framework. For comparison, linear regression models were fitted to matched input feature sets and an intercept term.

Partial least squares (PLS) regression was also used to model compartment profiles (Section **??**). PLS regression is well-suited to highly correlated inputs, employing a dimensionality reduction step to help address this redundancy, yet lacks the interpretability of a multiple linear regression. Similar to RF, PLS regression is aimed at building highly-predictive models rather than understanding singular relationships between a predictor and independent variable.[?] The `plsdepot` R implementation of PLS regression was used in this work.[?]

2.4 VARIABLE REGIONS

2.4.1 Stratification by variability

Median absolute deviation (MAD) was chosen as a robust measure of the variability in a given 1 Mb block between the three primary cell types used in this work (H1 hESC, K562 and GM12878). This simple metric is calculated by taking the median eigenvector value for each genomic bin across the three cell types, then taking the absolute difference between this median value and the two other recorded eigenvectors (as well as itself). Finally, the median of these differences is then calculated to give a MAD value per megabase bin.

Due to only three numbers being considered, the MAD is also equivalent to the minimum absolute difference from the median eigenvector value. That is, a MAD of 0.1 for a given region where cell type K562 has the median eigenvector value, means that both GM12878 and H1 hESC had absolute eigenvectors $\geq$ K562 + 0.1. Larger values report greater dispersion thus more variability between cell types.

Blocks were ranked by this measure and split into thirds that represented "low" variability (the third of blocks with the lowest MAD), "mid" and "high" variability. Each subgroup was then independently modelled using the previously-described Random Forest approach (Section **??**).

In another measure of variability, we also call regions of variable structure (RVS; Section **??**). RVS are those genomic regions whose compartment state differs in one cell type relative to the other two. For example, if a 1 Mb bin was classified as "open" in H1 hESC and "closed" in both K562 and GM12878, this is said to be an RVS, and a "flipped" compartment (to open) in H1 hESC. As can be seen, RVS calls are cell type specific: the same RVS that was called as flipped open in H1 hESC would not be called as variable in K562, as necessarily in this scenario H1 hESC and GM12878 would not be concordant.

2.4.2 Chromatin state enrichment

Chromatin state annotations used in this work were retrieved from `ChromHMM`[?] and `SegWay`[?] combined annotations.[?] These represent the consensus from two independent chromatin state prediction algorithms, and ignore regions of apparent disagreement; hence in theory making more robust and conservative predictions than either algorithm independently. Nevertheless, **?**] caution that in areas of disagreement, each algorithm may highlight differing biological phenomena so ideally should also be considered separately.

The set of state predictions from the combined algorithms are:

1. Predicted transcription start sites (TSS)
2. Promoter flanking regions
3. Transcribed reigons
4. Repressed regions
5. Predicted enhancers
6. Predicted weak enhancer or *cis* regulatory element
7. CTCF-enriched elements

Short, discrete state predictions such as enhancers were considered "shared" if there was an overlapping enhancer annotation in either of the two other cell types,

and labelled as "tissue-specific" otherwise. This was repeated for each of the called chromatin states.

2.4.3 Gene ontology analysis

Regions of variable structure (RVS; Section **??**) were tested for functional enrichments using Gene Ontology (GO) annotations.[?] The DAVID tool[?] was used to compare GO terms for genes located in variable compartments against a background set of genes from all annotated compartment bins.

2.5 BOUNDARIES

2.5.1 TAD boundaries

Having called TADs (Section **??**), we then have a set of boundaries at the start and end of each domain. We generated average boundary enrichment or depletion profiles by averaging input features into 50 kb bins spanning $\pm$450 kb from the central boundary bin.

To test for the enrichment or depletion of a chromatin feature over a given boundary class, a two tailed Mann-Whitney test was used to compare the boundary bin with the ten outermost bins of the window (5 from either side). Therefore we tested whether there was any significant difference in rankings of input feature signal between boundary bins and peripheral bins over all boundary instances per class. The significance level at $\alpha = 0.01$ was then Bonferonni-adjusted for multiple testing correction, and results with p-values exceeding this threshold were deemed significantly enriched or depleted at a given boundary.

2.5.2 Compartments

Eigenvectors were calculated as described previously (Section **??**). A/B compartmentalisation has previously been called simply from the properly-orientated principle component eigenvector, with positive values representing a bin in an A compartment state, and negative values representing a bin in a B compartment state.[?] Using this method, compartment boundaries occur whenever the eigenvector changes sign.

In this thesis compartment boundaries were called by first training a two-state hidden Markov model (HMM) on the compartment eigenvector and then using the Viterbi algorithm to predict the most likely state sequence that produced the observed values (Methods **??**). Justification for this approach is discussed in Section **??** and we also note the similar use of an HMM in TAD calling (Section **??**).

The point at which transitions occurred between compartment states was taken as a compartment boundary which was then extended ±1.5 Mb to give a 3 Mb window in which a boundary was though to occur. Boundary enrichments and alignments were tested in the same manner as TADs (Section **??**).

2.5.3 Boundary comparisons between cell types

To compare boundaries between cells, each TAD and compartment boundary called in K562 and GM12878 was compared with those called in H1 hESC. For each boundary, the minimum absolute difference to the nearest matching boundary in H1 hESC was recorded, and this was then compared with a null model of an equal number of boundaries randomly-placed along available bins (i.e. TAD boundaries were called at 40 kb, hence random boundaries could only be assigned to these same discrete bins). A Kolmogorov-Smirnov test was then used to compare the empirical cumulative distributions of these distances.

2.6 PREDICTING TAD BOUNDARIES

2.6.1 AUC-RF

To predict TAD boundaries we used a classification Random Forest model, built with the AUC-RF algorithm,[?] as implemented in the `AUCRF` R package.[?] This is a form of stepwise model selection which optimises feature subset selection relative to the area under the receiver operating characteristic (AUROC), a metric which captures both the specificity and sensitivity of a classifier hence is better-suited to unbalanced datasets than previously-described RF variable selection methods that used simple classification accuracy.[? ?] The AUC-RF procedure has been successfully applied in other bioinformatics applications, such as in identifying a subset of most relevant variants as part of a genome-wide association study.[?]

More specifically, the AUC-RF algorithm is a backwards elimination procedure, commencing with variable ranking by importance. In these classification RF models, mean decrease in Gini impurity (Methods **??**) was chosen as the measure of variable importance, as it has previously been proven to be more stable than the mean decrease in accuracy.[?] Regardless, impurity and permutation-based importance measures are thought to be largely consistent in many cases.[? ?]

Steps in the algorithm can be summarised as:[?]

1. Train an RF classifier using all available variables and calculate variable importance

2. Remove the least important 10% of variables and fit a new RF model

3. Calculate the AUROC on predictions made with out-of-bag data

4. Iterate steps 2–3 until a single variable model is built

5. Select that model with the highest AUROC as calculated in step 3

Importantly, AUC-RF avoids some of the problems of stepwise regression. Both over-optimistic model performance (due to repeated training and testing on the same pool of data) and errors due to the instability of variable rankings are mediated through an additional cross-validation step.[?]

The input feature set for this model was made up of the same 35 ENCODE features used in models of compartment eigenvectors (Methods **??**), with the addition of counts of Alu repeat elements (as used in Section **??**) and GERP scores as a proxy for the degree of evolutionarily constrained sequence (Methods **??**). TAD boundary bins were called as previously described (Methods **??**) and were resolved to 40 kb. Bins containing a TAD boundary were our true positives (TP), and to generate true negatives (TN) we took matched bins 450 kb upstream of each boundary bin. Of these, a randomly-selected 80% of TP/TN pairs were used as our training and validation set in each cell type, while the remaining 20% of cases were held-out as independent test sets.

2.6.2 Gini importance

Gini importance was the variable importance metric used to rank variables during the AUC-RF procedure (Section **??**) and was also used in the analysis of a two-step model prediction transcriptional output (Section **??**). This measure is calculated as follows.

The Gini impurity, G, of a single node containing some proportions of n classes is calculated as shown (Eqn. **??**), and is observably related to the concept of entropy or information gain.

$$G = \sum_{i=1}^{n} p_i(1 - p_i) \tag{6}$$

In our simple two-class setting (i.e. boundary, b, or non-boundary), this simplifies to Equation **??**. Here it can be seen that a G of 0.5 means the node contains a $50:50$ split of class labels, whereas a node of 95% boundaries has a much lower impurity ($G = 0.095$).

$$G = 2p_b(1 - p_b) \tag{7}$$

To convert the impurity into a measure of importance, we compare G of a parent node with that of its two daughter nodes (G_{d_1} and G_{d_2}), following a partition on a given variable of interest.

This decrease in Gini impurity can then be summed over all splits in which a specific variable has been selected per tree (N_{used}), and the averaged over all trees in the forest ($ntrees$). This approach is described in Equation **??**.

$$I = \frac{1}{ntrees} \sum_{ntrees} \sum_{j=1}^{N_{used}} G_j - (G_{d_1} + G_{d_2}) \tag{8}$$

As can be seen through this derivation, the Gini importance captures information regarding both how frequently a variable is selected at a node, and to what degree, after splitting by said variable, the labelled inputs are now better separated in daughter nodes. Through these concepts it is clear that variables with a larger Gini importance are providing a greater amount of useful, discriminative information to the Random Forest classification model.

2.7 METATAD ANALYSIS

MetaTADs are a conceptual level of genome organisation proposed by collaborators in the Pombo lab (Max Delbrück Center, Berlin). Their method for calling metaTADs involves the constrained hierarchical clustering of those neighbouring TADs with the greatest inter-TAD contacts. This pairing was recursed up to the level of whole chromosomes, thus resulting in a tree of increasing metaTAD aggregation. Since the calculation of metaTADs was performed and designed by collaborators, finer details are omitted here but are discussed fully in the associated manuscript.[?] Our contribution to the analysis of metaTADs is discussed in Section **??**.

2.7.1 Size selection

For boundary analysis of metaTADs, again a similar approach was used to that of TADs (Section **??**) but with metaTADs thresholded to within a given range of domain sizes. Those below 10 Mb were excluded, as to have no lower bound results in $\frac{2}{3}$ of all TAD boundaries likewise considered MetaTAD boundaries, reducing the statistical power to detect any differences. 10 Mb was chosen as a compromise between minimising the overlap between TAD and metaTAD boundaries, while also retaining a large enough sample size (Section **??**). An upper bound of 40 Mb was also chosen, as beyond this threshold inter-TAD contacts were found to be no higher than expected by chance (*personal communication*). In practice, the tree-like structure means any upper-bound has little impact as a filter: in almost all cases, any boundary in a metaTAD of size > 40 Mb will also form metaTADs below this value. Additionally, the hierarchical nature of metaTADs means that some boundaries are present at multiple levels of the

tree. Only one case of each boundary position was tested for feature enrichments, and this was performed as with TAD boundaries (Section **??**).

2.7.2 Collaborator datasets

Our collaborators in the metaTAD project performed ChIP-seq experiments for PolIII (three variants), H3K27me3, CTCF and DNase-I hypersensitivity. Mapped reads from these experiments were processed using MACSv2[?] to give relative signal over background (from an estimated local model), which was then averaged over all boundaries genome wide.

Cap analysis of gene expression (CAGE) data was produced by the FANTOM consortium.[? ?] This method produces sequencing data from the 5′ end of cDNAs, and can be used to quantify expression activity at precise promoter locations.[?] Here, CAGE was performed at multiple points along a neural-differentiation timecourse and tags were clustered to form CAGE TSS (CTSS) in a manner developed for the use within the FANTOM5 project.[?] To count these CTSS over boundary bins, we simply intersect the annotations and count CTSS per bin using `bedtools`.[?]

Gene density over metaTAD boundaries was calculated using UCSC `mm9` gene models.[?] Again simple intersections were taken to count genes over boundaries using `bedtools`[?] and requiring a minimal overlap fraction of at least 0.5% of a bin (250 bp).

2.7.3 LAD coincidence

Lamina associated domains (LADs) are genomic regions which are in contact with lamin proteins A, B and C, found on the inner nuclear membrane (reviewed in **?**). To compare metaTAD boundaries with those of LADs, we made use of previously-published Lamin-B1 DamID microarray probe intensities.[?] For analysis over boundaries, these values were averaged into the same boundary windows as used previously (50 kb bins $\pm$450 kb around boundary, as in Section **??**).

Transitions between high and low lamina association were detected by fitting a linear regression model across each series of consecutive boundary bins (i.e. Lamina assoc. $= \beta \cdot \text{bin} + c$). Linear models which had an absolute coefficient $|\beta| > .05$ were taken as crossing a LAD transition. This threshold is a heuristic which appears to perform well at conservatively selecting clear transitions. As a method of seriation for the y-axis of heatmap figures (e.g. Fig. **??**), boundaries were divided into those that coincided with a lamin transition and those that did not, and members within each group were then sorted by average intensity.

To test the significance of the association between boundaries and lamin transitions, we circularly permuted both TAD and metaTAD boundaries within each chromosome

1000 times, and calculated the proportion of boundaries that crossed LAD boundaries using the same linear regression procedure described above. Empirical p-values were then calculated as the number of permuted results greater than or equal to the observed value.

2.8 GIEMSA BAND COMPARISON

Cytogenic band data and Giemsa stain results were downloaded from the UCSC genome browser (table `cytoBandIdeo`). The genomic co-ordinates are an approximation of cytogenic band data inferred from a large number of FISH experiments.[?]

To compare G-band boundaries with our compartment data, we allowed for a $\pm$500 kb inaccuracy in G-band boundary. For each G-band boundary, the minimum absolute distance to any compartment or TAD boundary was calculated for each cell type. To generate a null model, we calculated 20 circularly-permuted sets of G-bands per chromosome, and recalculated their distance from our compartment boundaries. Differences were then compared as empirical cumulative distributions using a a two-sided Kolmogorov-Smirnov test.

2.9 NUCLEAR POSITIONING

Previously published data on chromosome positioning preference within the nucleus was used to label each chromosome as "inner", "middle" or "outer".[?] Chromosomes whose DAPI ($4'$,6-diamidino-2-phenylindole) hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by Boyle *et al.*[?], made up the "inner" group and included chromosomes 1 and 16. Similarly the "outer" group had enriched signals ($p \leq 5 \times 10^{-3}$) in the outer shell relative to the inner nuclear shell and included chromosomes 2, 3, 11-13 and 18. Remaining chromosomes were assigned to the "middle" group and showed no significant to either inner or outer nuclear shells ($p \geq 0.1$).[?]

The significance of the difference in distribution of eigenvectors in the inner versus outer shell was determined by a two-sided Kolmogorov-Smirnov test, with the alternative hypothesis that the empirical cumulative density function of the inner chromosome eigenvectors F_{inner} was not equal to F_{outer}. This chromosomal positioning data was measured in lymphoblastoid cells, though this level of nuclear organisation is thought to be largely conserved between cell types[? ?] and even among higher primates,[?] so should be comparable across cell types for this purpose.

2.10 MODELLING TRANSCRIPTIONAL OUTPUT

2.10.1 Reproducing a published study

In Section **??** we reproduce and extend a previously published study by **?**[] In doing so, we reuse much of the code and materials made available by the authors and more widely by the ENCODE consortium,[?] of which this paper was a part. Some scripts were extracted from the ENCODE virtual machine,[?] designed to provide an environment in which to reproduce their main findings.[?]

Input features for models of transcription were derived from the January 2011 ENCODE data freeze.[?] Normalised ChIP-seq signals were generated by ENCODE using `wiggler` and retrieved for this study as `bigWig` files. These were averaged into 40×100 bp bins across each GENCODE v7 TSS, to give ± 2 kb windows around each start site. These bins were then used to find the 'bestbin', that which correlates best with transcriptional output on a training subset of TSS.[?] A bin representing the average intensity over the whole gene (TSS to TES) was also considered. That which best correlated on a training set was then used as the representative region for that feature in subsequent modelling steps.[?] The justification for this approach is discussed in Section **??**.

2.10.2 Predicting FANTOM5 expression levels

We transferred this transcriptional modelling approach to what was at the time novel, unpublished CAGE data produced by the FANTOM consortium. This data has since been released in the FANTOM5 series of publications.[?]

Specifically, we used H1 hESC t_0 CAGE data from a differentiation timecourse study. The consortium pre-processed raw CAGE tags into clusters using decomposition-based peak identification.[?] To filter for gene-associated CAGE clusters, we discarded those tag clusters centered on a point > 50 bp from an Ensembl (`v69`) annotated TSS, thereby removing transcribed enhancers and other non-genic regions with detectable transcription. When multiple clusters were linked to the same TSS, that with the highest peak maximum was kept. Expression was matched with ENCODE ChIP-seq data for the H1 hESC cell type (processed as described in Section **??**) and an additional measure of replication timing retrieved from **?**[] (Section **??**).

Input data for models of FANTOM5 CAGE are shown in Table **??**.

Table 3: ENCODE datasets generated in the H1 hESC cell line and used in models of transcriptional output.

Histone modifications	Other
H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1	HDAC6, DNase I, H2A.Z, Input

2.11 4C DATA ANALYSIS

For computational analysis of 3C-seq data (also known as 4C), the experimental protocol used by our collaborators recommends the `r3Cseq` R package,[? ?] part of the BioConductor repository[? ?] for the R programming environment.[?]

This package produces normalised interaction frequencies which are comparable between experiments and then assigns statistical significance to any identified contacts, thereby reporting regions that co-localise to a greater degree than expected by their genomic proximity alone.

2.11.1 Normalisation

The normalisation procedure for 4C data is adapted from a previous method for normalising deepCAGE data between samples.[?] In short, the reverse-cumulative distribution of read counts per restriction fragment is fitted to a power-law model; this effectively encodes the *a priori* expectation of exponential decay of the number of contacts as distance increases from the viewpoint. Transformed read counts per million (RPM) can then be retrieved from a standardised reverse cumulative distribution, parametrised with an empirical coefficient for this power-law relationship ($\alpha = -1.35$).[?]

This normalisation procedure has the effect of making the output RPM value independent of the original experiment's sequencing depth and, more importantly, acts to reduce the impact of artefacts and errors by enforcing the expected power-law relationship of restriction fragment read counts.

2.11.2 Significance estimation

The `r3Cseq` package[?] also attempts to assign a measure of statistical significance to observed contact frequencies. This is done through a simple method of background estimation based on observed values. The justification for this non-independent estimate of background signal is that a relatively small proportion of observed contacts are ex-

pected to be significantly enriched, thus will not unduly perturb an average signal.[?] An improved method that avoids this assumption has since been developed where a background model was iteratively fitted, with outlier removal at each revision.[?]

Here a non-parametric cubic smooth spline is fitted to normalised read count data using a heuristic smoothing parameter. This model then provides an expected level of interaction at a given distance from the viewpoint in *cis*. From this, it is simple to calculate a Z–score as:

$$Z = \frac{O - E}{\sigma} \tag{9}$$

Where σ is the standard deviation of residuals from the observed (O), expected (E) difference. This Z-score can then be converted to a p-value which in turn is corrected for multiple testing using bootstrapped estimates of false-discovery rate (FDR) q-values[?] (as implemented in the `qvalue` R package[?]). This Z–test approach assumes a normally-distributed test statistic, an assumption that typically does not hold on 4C data where interactions distal to the viewpoint are increasingly sparse, however this approach and variants thereof have been applied in a variety 4C and 5C analyses (e.g. **? ? ? ? ? ?**). Some publications (e.g. **?**) use a more appropriate distribution to assign p-values to a Z-type statistic, such as the Weibull (extreme value) distribution.

While we are mostly concerned with these *cis* interactions, `r3Cseq` also offers significance testing for *trans* interactions between the viewpoint and restriction fragments on different chromosomes. Here instead of distance scaling, the expected (E) term in Equation **??** is just the genome-wide background average, excluding regions $\pm$ 100 kb around the viewpoint.[?] This means the absolute values of normalised RPMs reported for *trans* interactions are in practice upscaled, being equivalent to experimental RPMs less the most deeply-sequenced regions, i.e. the viewpoint and immediately adjacent regions.

2.12 SCRIPTS AND OTHER ANALYSES

Much of this work has been performed by writing custom scripts in the R programming language.[?] Code for the majority of analyses described in this thesis are available through a public git repository hosted on `github` at github.com/blmoore/3dgenome (instructions on how to reproduce analyses and figures are included therein). A special mention goes to the packages of Hadley Wickham which are used throughout, especially `ggplot2`[?] and `dplyr`[?].

The programming language python[?] was also employed to a lesser-extent, as were command–line tools such as `bedtools`[?] and `SAMtools`[?]. Additionally command-line `BigWig*` tools[?] were used, as well as the UCSC genome browser and associated data tracks.[? ? ?]

3 REANALYSIS OF HI-C DATASETS

3.1 INTRODUCTION

Since the initial publication of the Hi-C technique in 2009,[?] there has been rapid advancement of both the technique itself and the resolution at which interaction frequencies have been analysed. From proof-of-concept analyses at 1 megabase (Mb) and 100 kilobase (kb) resolution,[?] subsequent experiments achieved resolutions first of 40 kb[?], then 10 kb[?] and most recently 1 kb,[?] enabling bona fide genome-wide fragment-level analysis for the first time.

Such rapid progression in the field has resulted in a wide variety of public Hi-C datasets being available, albeit with differing qualities. With proper correction and at a suitable resolution, these interaction frequencies can be compared and contrasted both within and between species.

In this chapter we uniformly reprocessed publicly-available human Hi-C datasets in order to address fundamental questions about the stability of higher order genome organisation between cell populations from the same species. Previously Hi-C studies have compared two samples, such as K562 against GM06990[?] or IMR90 against GM12878.[?] Here we make use of three Hi-C datasets corresponding to extensively-studied human cell lines: K562, GM12878 and H1 hESC. Together these make up the "Tier 1" cell lines studied by the ENCODE consortium,[?] and hence have huge amounts of matched locus level features, such as ChIP-seq and histone modification data available.

By combinatorial reanalysis of these cell-matched datasets, we can comprehensively investigate the quantitative relationships between higher order chromatin structure and locus level chromatin features.

3.2 HI-C REPROCESSING

Each Hi-C dataset used in this work was reprocessed from raw sequencing reads using the same pipeline (Methods **??**). Briefly, raw sequencing reads were sourced from three different publications: **?**], **?**] and **?**]. These reads were mapped to human genome build `hg19` using an iterative mapping procedure that maximised the number of uniquely mappable reads from each sample (Methods **??**).

Next a filtering step was applied, which removed those fragment pairs that were likely artifactual or erroneous (Methods **??**). A correction step was then applied,

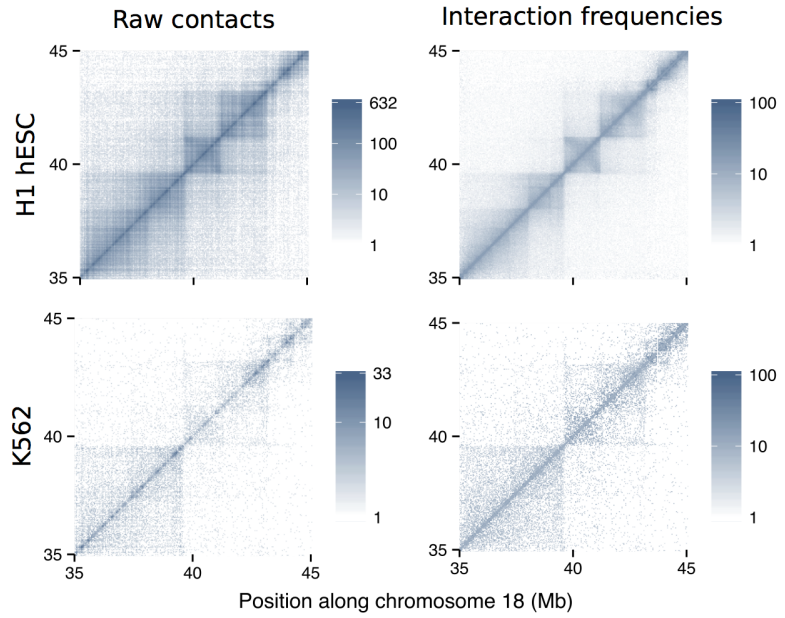

Figure 8: Iterative correction converts raw counts to normalised interaction frequencies. The sample with highest sequencing depth (H1 hESC) is shown alongside a sample with much lower sequencing depth (K562) both before and after iterative correction and normalisation procedures were applied (Methods **??**) at 40 kb resolution for a 10 Mb section of human chromosome 18. Fill gradients are on a $\log_{10}$ scale.

whereby biases such as mappability and GC content were removed to give each fragment equal visibility (Methods **??**). Overall these steps produced comparable maps of interaction frequency in different cell types, despite their differing origins (Fig. **??**).

Figure **??** shows a 10 Mb region of chromosome 18 before and after filtering and normalisation in two different cell types. Self-interacting domains visible in the deeply-sequenced H1 hESC cell type also become more visible in the K562 cell type after normalisation. In addition many of the long-range and intra-domain contacts visible in each raw contact map are down-weighted during the normalisation procedure, indicating their prominence was enhanced by biases or other sources of noise in the experimental procedure (Fig. **??**).

The relative sparsity of the K562 Hi-C contact map compared to that of the H1 hESC cell line should also be noted (Fig. **??**). At the time much of this study was performed, deeply-sequenced Hi-C datasets for cell lines K562 and GM12878 were not available, thus the majority of analyses were performed at a lower resolution of 1 Mb to further reduce the impact of variable sequencing depth between cell lines (Methods **??**).

3.3 COMPARTMENT PROFILES

After uniformly reprocessing each Hi-C dataset and calling compartment eigenvector profiles (Methods **??**), we can compare these between three human cell lines. We find compartment profiles have a striking concordance (Fig. **??**), despite the variable sources of both sample material and experimental data. This strong correlation of higher order chromatin structure (at the level of compartments) between three very different human cell types shows that the vast majority of genomic regions appear to be constitutively present in either the A or B compartments regardless of cell lineage.

This close correspondence also validates our approach of combining these different datasets, and suggests our uniform pipeline is successfully accounting for differences in sequencing depth and other batch effects. The pairwise Pearson correlation coefficients between these independent measures are all in the interval $[.75, .8]$ (Fig. **??**). Also of note is that each cell type independently shows a similar bimodal distribution of compartment eigenvector, indicative of the two distinct underlying A/B compartment states (Fig. **??**).

3.4 DOMAIN CALLS

3.4.1 Compartments

The continuous compartment eigenvector is most commonly used as-is to classify A/B compartments by thresholding based on sign: typically the eigenvector is orientated such that positive values reflect A compartments and negative values B compartments.[? ?] However, given that compartments are understood to be generally broad and alternating domains along a chromosome, often aligning with other large domains such as LADs, an improved classification method might penalise the calls of very short compartment calls, which may be the result of noise. For this reason, instead of using raw eigenvector values we consider observed values as emissions from unobserved underlying states (Fig. **??**). We built a hidden Markov model (HMM; Methods **??**) to represent these states through a well-described probabilistic framework.[?]

Firstly we designed the HMM to have two unobserved states with univariate Gaussian distributed emissions (representing our A and B compartments; Fig. **??**). To parameterise these states we used the iterative Baum-Welch algorithm, an Expectation-Maximisation procedure designed for HMMs. Having parameterised the HMM for each cell type, we then use the Viterbi algorithm to infer the most likely state sequence to have generated our observed data. This two-state sequence is then used to assign compartment identities to genomic bins. A schematic of this procedure is shown in Figure **??**.

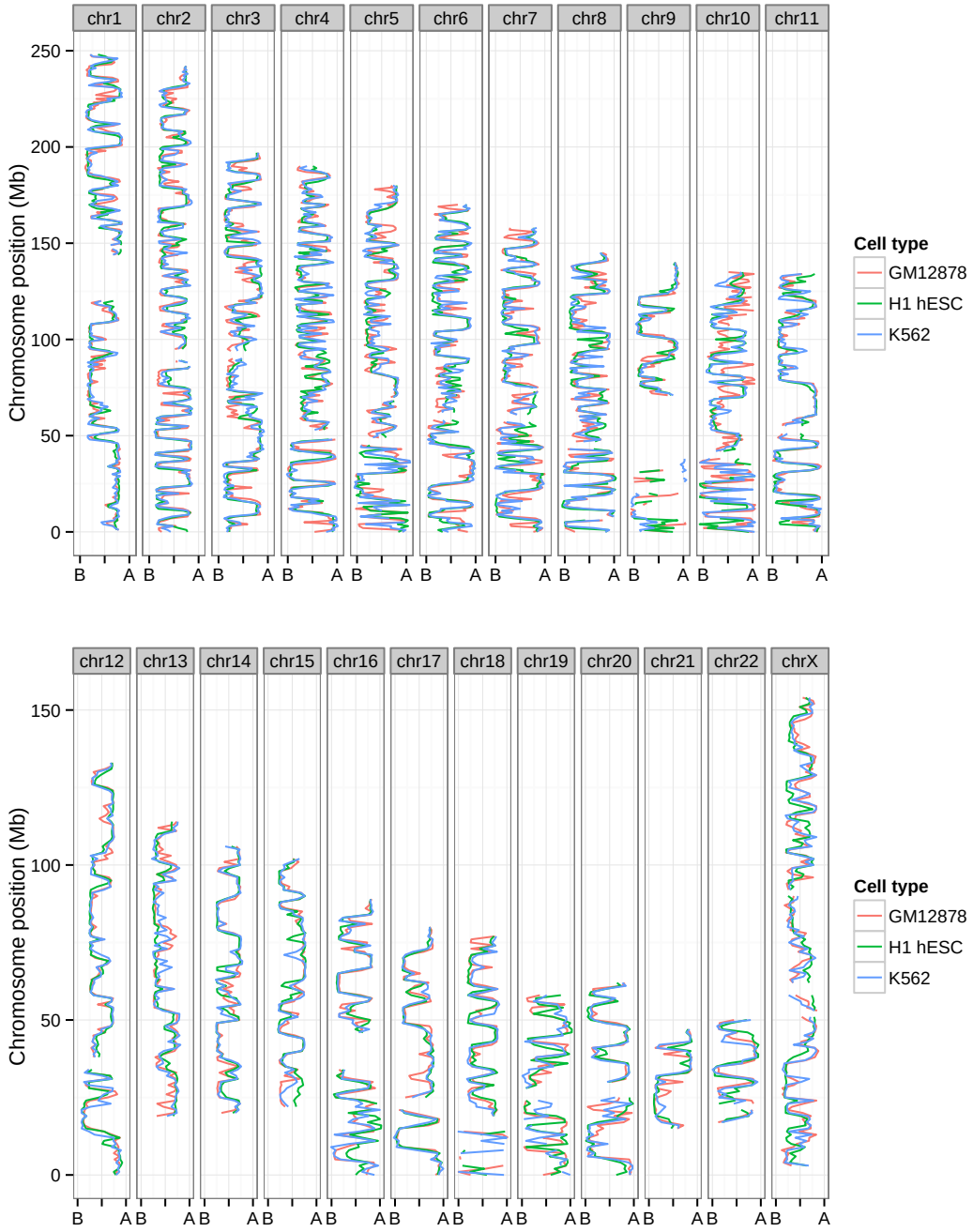

Figure 9: Compartment profiles are observably well-correlated between human cell types and across all chromosomes. Compartment eigenvectors are plotted along the lengths of each human chromosome (chrY and chrM are omitted). In each case the overlaid profiles show strong concordance between the three different human cell types under study.

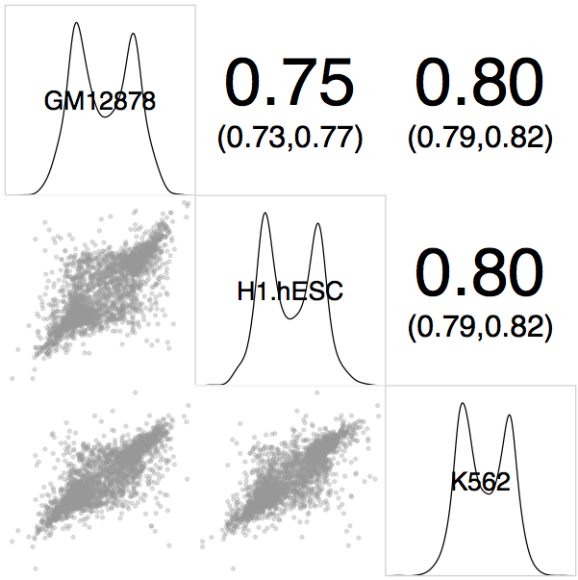

Figure 10: Compartment eigenvectors are highly correlated between human cell types. Megabase resolution compartment eigenvector values are shown in a plot matrix. *Upper triangle*: Pearson correlation coefficients between pairs, with 95% confidence intervals; *diagonal*: kernel density estimates of eigenvector values per cell type; *lower triangle*: pairwise *x*-*y* scatterplots of compartment eigenvector values.

In practice, this approach acts to de-noise our compartment calls. Whereas single sign-changes along the series would (under a simple thresholding procedure) have resulted in a single-block compartment, these may now be modelled as noisy emissions from a single unobserved state. An exemplar region is shown in Figure **??**. This shows an approximately 50 Mb region from chromosome 8 with eigenvector data from the H1 hESC cell line. A simple thresholding method in this region calls a total of 12 regions, whereas our HMM method finds only 6 larger regions in the same window. The disparity is caused by very short and single-bin compartments being disfavoured by the HMM-based method (e.g. Fig. **??**).

Having called compartments we can compare their properties across cell types. In each case, a majority of chromosome compartment sizes are in the range 0–10 Mb, with a handful of compartments reaching up to 40 Mb in size (Fig. **??**). Median sizes for compartments called in this work match those reported previously, with a median size of around 5 Mb.[?] Our slightly larger mean compartment sizes (up to 7.6 Mb in H1 hESC) may be due to our altered domain calling procedure (Fig. **??**) and is clearly influenced by some large outliers (Fig. **??**).

Next we compare compartments between the three cell types under study. To do this, we calculate the minimum absolute distance from each boundary in one cell type against those in a designated comparison cell type (we used H1 hESC). The cumulative distribution of these boundary differences is shown (Fig. **??**) and compared to a null distribution of random compartment boundaries (Methods **??**). We find boundaries

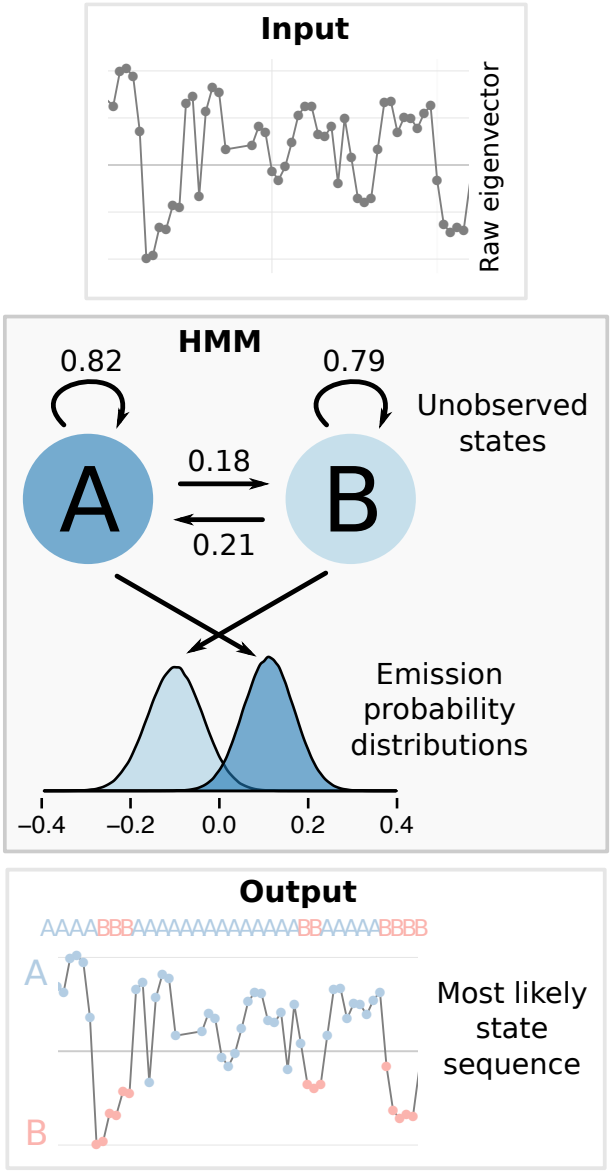

Figure 11: Overview of HMM method for compartment calls. Schematic of HMM method for compartment calls, showing transition probabilities and emissions distributions learned in the GM12878 cell type. A description of HMMs is given in Methods **??**. Emission distributions for each cell type are shown in Figure **??**.

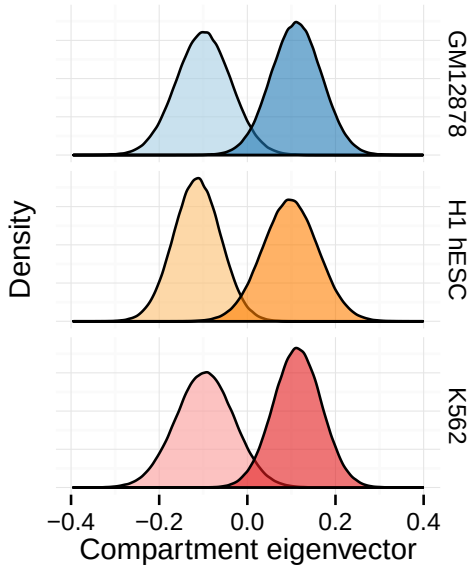

Figure 12: Univariate Gaussian emission distributions for A and B unobserved HMM states. Probability distributions for compartment eigenvector values per HMM state are shown for each cell type. Distributions centred below zero (lighter colours) show the distribution for the B compartment state, distributions centred above zero (darker) represent the A compartment state.

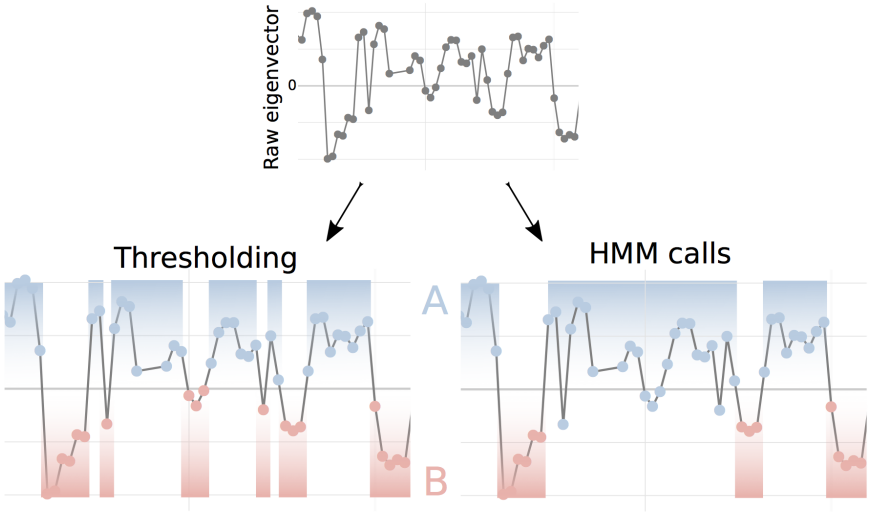

Figure 13: Compartment calls by simple thresholding method or context-aware HMMs. Chromosome compartments have previously been called through simple thresholding at 0,[?] in this work we use a novel HMM-based method to call unobserved states that have emitted our noisy observed values (*right*).

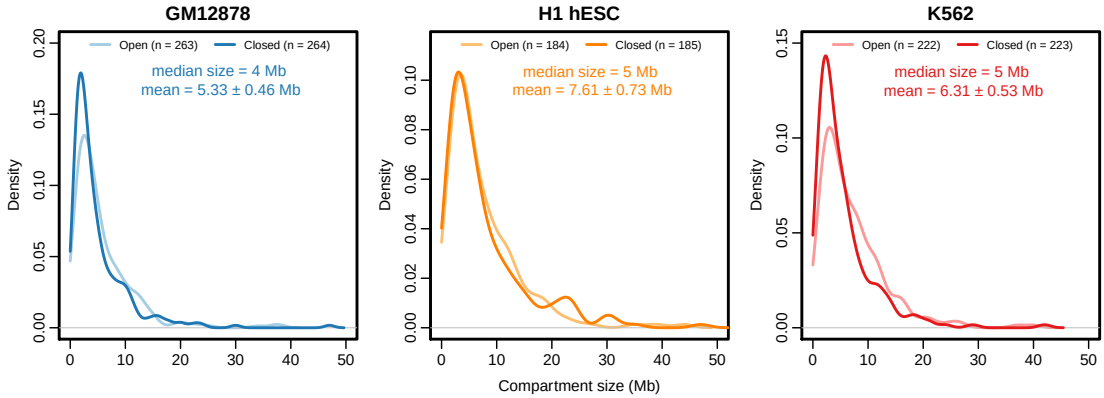

Figure 14: Size distributions of compartments called in three human cell types. Kernel density estimates of compartment sizes (Open: A; Closed: B) are shown per cell type with summary statistics (*inset*), including mean compartment size with 95% confidence intervals.

are significantly more closely aligned across cell types than is expected by chance. For example, genome-wide approximately 37% of compartment boundaries in H1 hESC have a corresponding boundary within 100 kb in GM12878 (and 35% in K562, but only 5% in random boundaries). Comparisons of the cumulative boundary distance distributions yield statistically significant differences relative to the null model (K-S test, K562: $D = 0.47$; $p \approx 0$; GM12878: $D = 0.49$; $p \approx 0$; Fig. **??**).

3.4.2 TADs

Topological associating domains (TADs) are self-interacting blocks of the genome first described by **?**] We applied the original TAD calling method without modification, which uses a measure of the directional contact bias of a fragment (Section **??** and Fig. **??**).

The **?**] method of calling TADs relies on the detection of boundaries,[**?**] thus it is affected by sequencing depth: experiments with sparser contact matrices may not contain enough for a sufficiently high degree of bias to allow a boundary call. This is evident in our datasets even after normalisation, with the deeply-sequenced H1 hESC cell type having approximately 50% more TADs called than in the GM12878 cell type (Fig. **??**). This effect could have been mitigated by down-sampling reads in the H1 cell type, but at a cost of reducing the quality of the best dataset under study. Instead this disparity should just be noted as a potential cofounder in downstream TAD analysis; at lower-resolution such as that used to calculate compartment eigenvectors (1 Mb) this sensitivity to sequencing depth is not evident (Figs. **??**, **??**).

Despite differing numbers, there is still detectable levels of conservation of TADs between cell types (Fig. **??**). Genome-wide, 45% of all H1 TAD boundaries have a

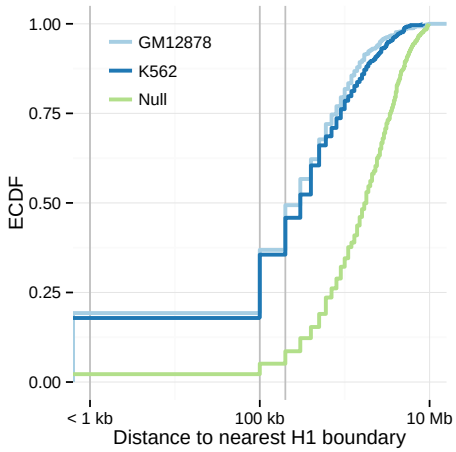

Figure 15: Compartment boundaries are shared between cell types. The empirical cumulative distribution functions (ECDF) of distances between H1 compartments and those called in GM12878 and K562 are shown. Vertical lines mark distances of 0, 1 and 2 bins. Also plotted is the ECDF of a null model, where distances were calculated to shuffled boundaries at a matched resolution (Methods **??**).

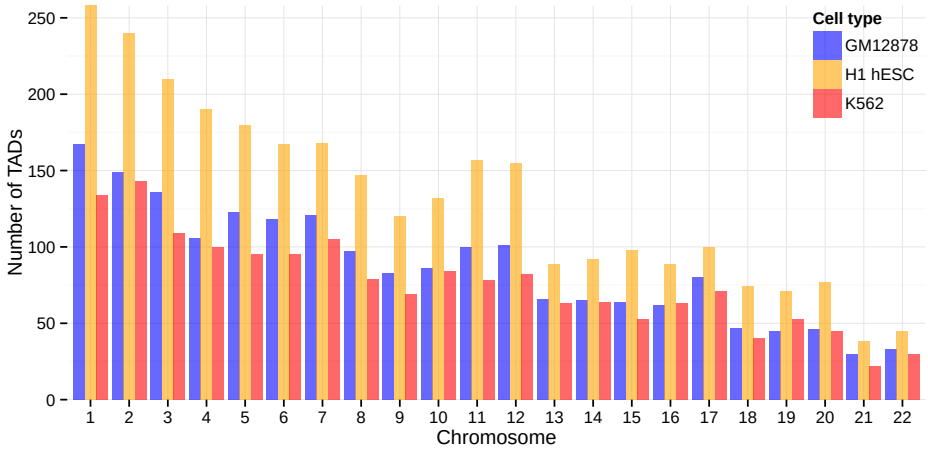

Figure 16: The number of TADs called per chromosome in each cell type under study. A greater number of TADs were called in H1 hESC (2,897 total) than in GM12878 (1,925) or K562 (1,677), due to the difference in sequencing depths in each experiment when matrices were binned at 40 kb resolution.

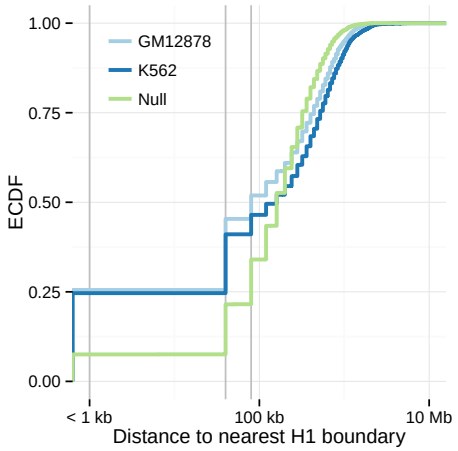

Figure 17: TAD boundaries are shared between cell types. The empirical cumulative density functions (ECDF) of distances between H1 TADs and those called in GM12878 and K562 are shown. Vertical lines mark distances of 0, 1 and 2 bins. Also plotted is the ECDF of a null model, where distances were calculated to shuffled boundaries at a matched resolution (Methods **??**).

matching boundary in GM12878 in the same or an adjacent 40 kb bin (K562: 41%, null model: 22%; K–S test: $D = 0.2$, $p \approx 0$). To illustrate this conservation with a real example, a 20 Mb region of chromosome 2 is pictured (Fig. **??**), highlighting the conservation between both TADs and compartment calls across the three cell types and at multiple scales: from chromosome-wide 1 Mb compartment eigenvectors, to TADs with individual boundaries resolved to 40 kb.

3.5 DOMAIN EPIGENOMICS

The use of well-studied human cell types allows intersection with publicly-available epigenomics datasets, such as those produce by the ENCODE consortium.[?] In total, 35 cell-matched ChIP-seq datasets were available for all three of the tier 1 ENCODE cell lines: GM12878, H1 hESC and K562 (see Methods **??**). In this section we integrate reprocessed Hi-C data and derived higher order domains with these various high-resolution datasets.

3.5.1 A/B compartments

The overwhelming majority of intersected chromatin features are significantly enriched in active A compartments relative to B compartments (Fig. **??**). This is expected since A compartments represent the actively-transcribed and accessible portions of the

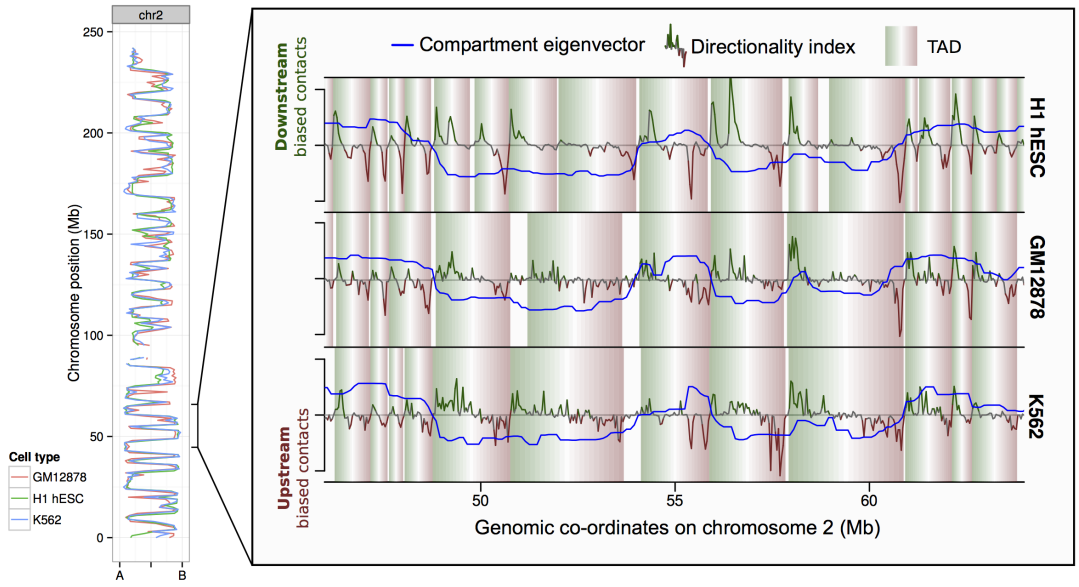

Figure 18: Concordance of chromatin structure at multiple scales over three human cell types. The eigenvector compartment profile is shown for chromosome 2 for three human cell types (*left*). At higher resolution, the zoomed region illustrates conservation of topologically associating domains (TADs) over 20 Mb of the same chromosome.

genome, and have previously been shown to positively correlate with many of the features shown.[? ?]

Exceptions to this rule are few. However the repressive histone modification H3k9me3 is found more often in B compartments in two cell types, as is the P300 transcription factor (Fig. **??**). Also of note is the histone variant H2A.Z which is significantly enriched in A compartments in GM12878 and K562, but this relationship is reversed in the embryonic stem cell line (Fig. **??**). Recent evidence suggests specialised roles for H2A.Z in regulating both repression and activation during embryonic stem cell differentiation, acting as a "general facilitator".[?] Additionally H2A.Z has been reported to mark histone octamers for depletion, thereby permitting gene activation during differentiation.[?] Potentially, then, the H2A.Z enrichment in B compartments could be driven by regions soon to be de-repressed as the stem cell differentiates.

3.5.2 TAD classes

Unlike compartments, initially TADs were not thought to be correlated with blocks of chromatin features.[?] Later studies have linked TADs with such enrichments, first in *Drosophila*[?] and later in human cells, where it was argued TADs are merely a low-resolution window to smaller "sub-compartments", bearing similar active and inactive

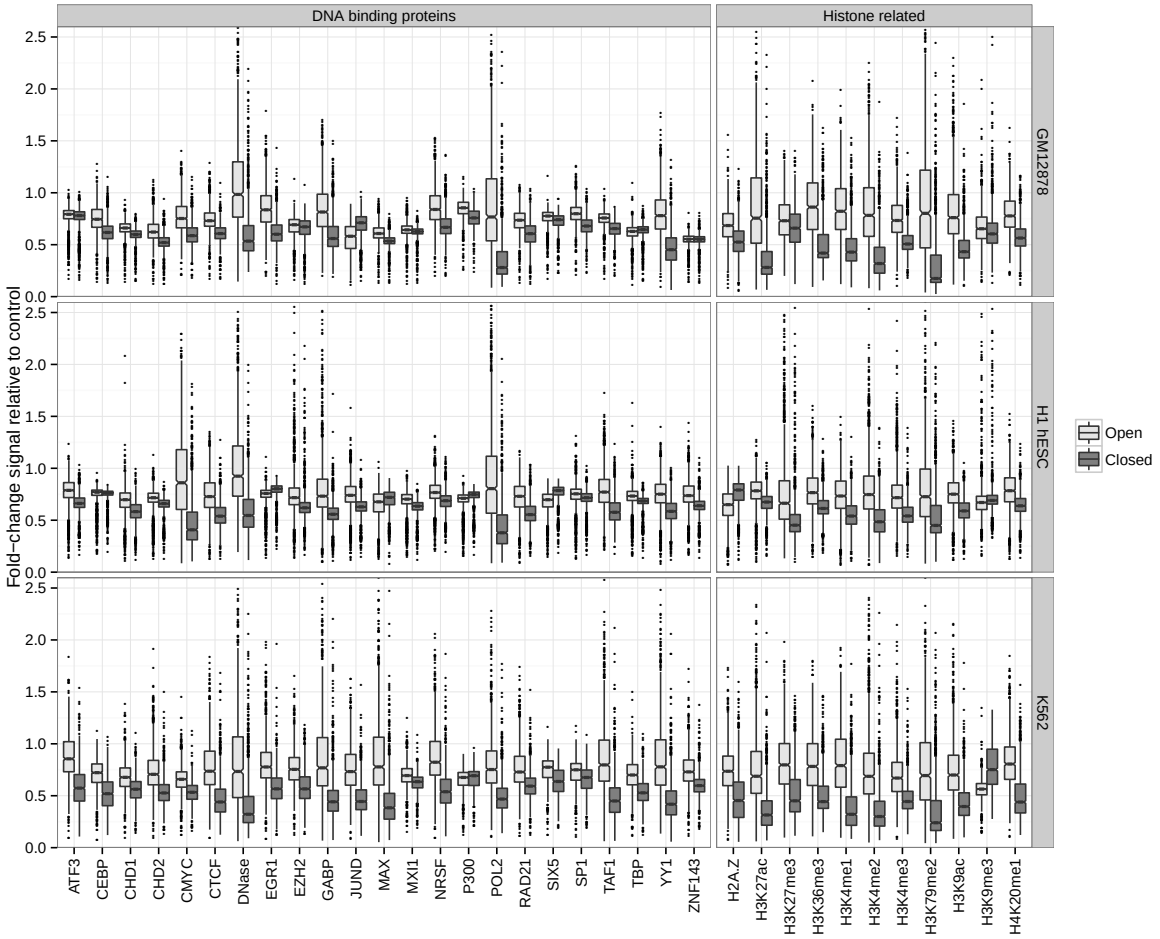

Figure 19: The chromatin signatures of A/B compartments. Notched boxplots summarise the distribution of each feature over 1 Mb bins in open (A) and closed (B) compartments genome-wide.

marks to their much-larger namesakes.[?] Here we look for evidence that TADs called in our human cell types correspond to the "epigenomic domains" identified in *Drosophila* by **?**] Epigenomic domains were identified through supervised clustering of "physical domains" (TAD analogues called in *Drosophila*) by their average enrichment for selected epigenomic features of known function, for example enrichment of H3K27me3 mark was used to call Polycomb (PcG) associated domains.

We found that TADs called across the three cell types used in this work could similarly be clustered into transcriptionally active (active), repressed heterochromatin (null) and polycomb-associated (PcG) domains, based on the patterns of DNase hypersensitivity, H3k9me3 and H3k27me3, respectively (Fig. **??**). *Drosophila* physical domains were clustered into four categories, with three of those matching our annotations. The fourth *Drosophila* domain type was enriched for the HP1 protein (therefore likely centromeric) for which we did not have ENCODE ChIP-seq data in all human cell types under study.

This analysis reveals that active compartments typically cover both active and PcG-associated TADs, while B compartments appear more homogeneous and are composed mostly of H3k9me3-enriched heterochromatin even when considering fine-grained TAD structures rather than megabase-sized genomic blocks (Fig. **??**).

These findings also link with recent work that suggested TADs are windows into "sub-compartments"[?] which more closely reflect the functional enrichments of compartments. However, in our data we did not find statistical support for the suggested 5 classes of sub-compartment; instead, an ensemble of algorithms for optimising the number of cluster centroids voted for two or three clusters of TADs (*data not shown*). This is not wholly surprising as **?**] report sub-compartments only on extremely deep-sequenced samples, and at a scale of organisation below that of TADs.

3.6 VARIABLE REGIONS

Despite the vast majority of the genome being in matched chromatin compartments across human cell types (Fig. **??**), there are also regions of disagreement. Reasons for observable differences include technical errors and biases, but also more interesting functional explanations, where cell-type specific activation or repression is reflected in changes in higher order structure.

To conservatively call regions of variable structure (RVS), we used HMM-called compartment states and selected those which were either: i) open in one cell type and closed in both others or ii) closed in one cell type and open in both others. This left sets of RVS which could be considered as "flipped open" or "flipped closed" in a given cell type.

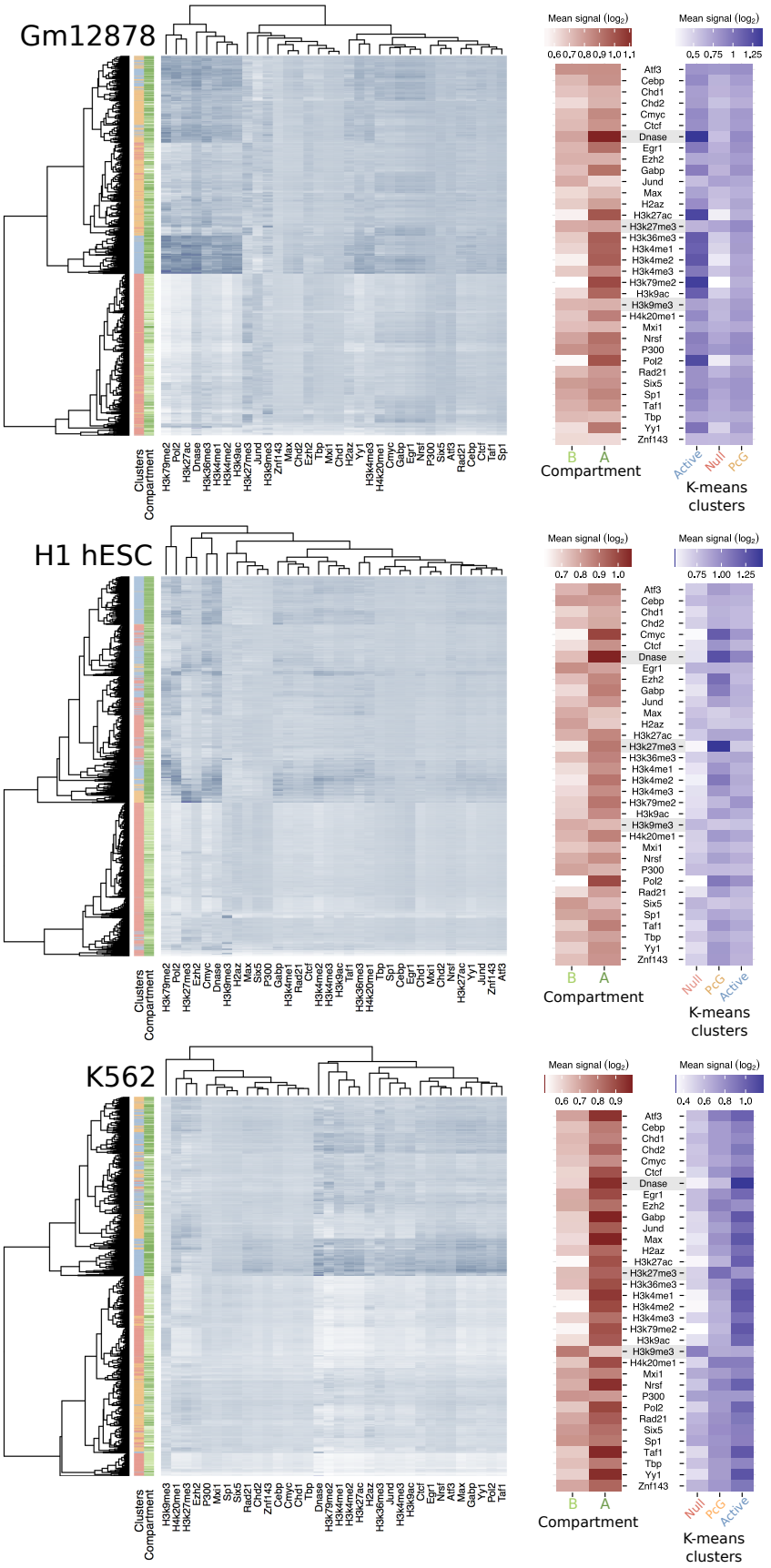

Figure 20: TADs reflect epigenomic domains. Following the *Drosophila* results of ?], clustering of TAD domains by mean $\log_2$ signal of 34 ENCODE features distinguishes null, active and polycomb-associated (PcG) domains, as well as reflecting the encompassing A/B compartments.

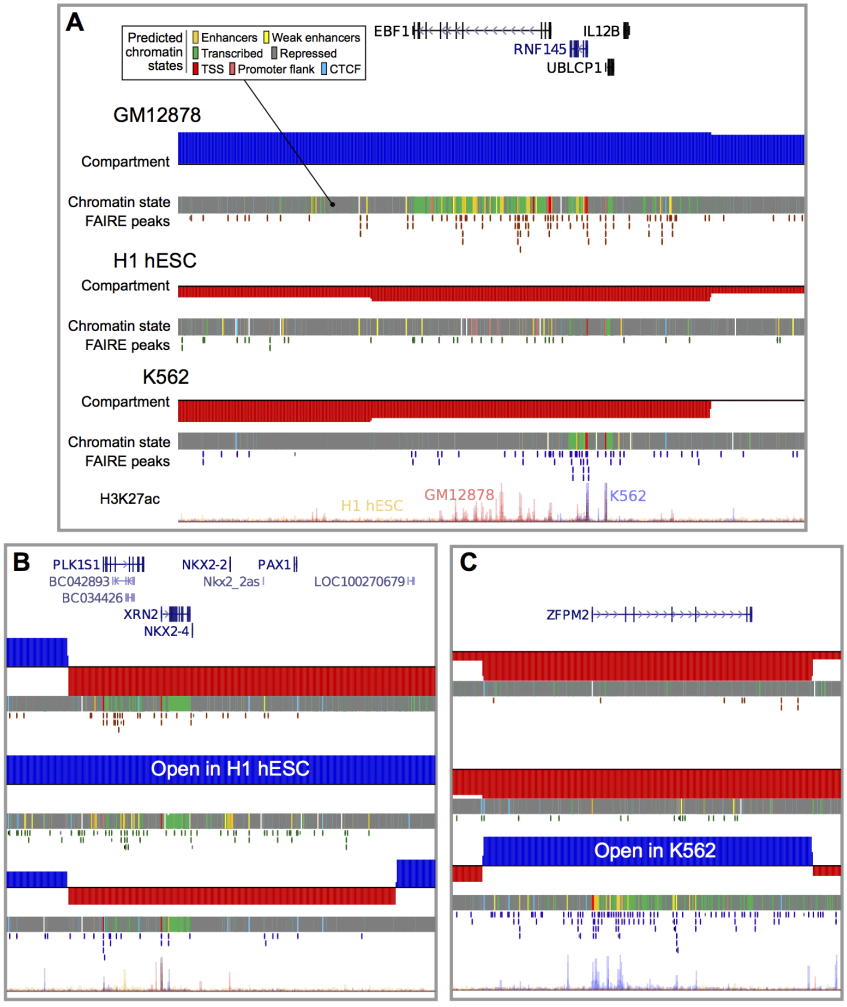

Figure 21: Structurally variable regions indicate cell type specific biology. Regions occupying the active A nuclear compartment in one cell type, but the repressive B compartment in the other two, were selected and ranked by the number of predicted active enhancers. Examples of particular interest from the top 5 regions per cell type are shown: (A) the chr5:158-159 Mb region which occupies the open A compartment in GM12878 cells, (B) the chr20:21-22 Mb region which is open in H1 hESC, (C) the chr8: 106-107 Mb region which is open in K562. Displayed tracks are: known (UCSC) genes, compartment eigenvectors, chromHMM/Segway combined chromatin state predictions, open chromatin FAIRE peaks, and H3K27ac signal.

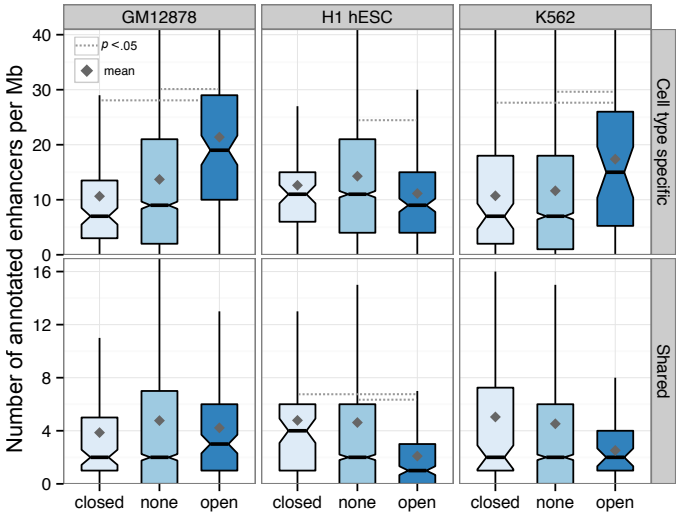

Figure 22: Regions of variable structure are enriched for cell type specific enhancers. Numbers of predicted enhancer states (cell type specific or shared between two or more cell types) are shown for regions with altered (open or closed) and non-altered (none) compartments in each cell type.

3.6.1 Chromatin state enrichment

Given our conservative definition of RVS (Section **??**), such notable changes between transcriptionally permissive and repressive environments might be expected to be associated with cell-type-specific biology, such as functional chromatin states. To test this, we used consensus predicted chromatin state annotations, built from two machine learning algorithms, `ChromHMM`[?] and `SegWay`[??], and tested for enrichment or depletion in our set of RVS (Methods **??**).

We found that RVS show a striking enrichment for cell-type specific enhancers in both of our derived cell lines, but not in embryonic stem cells (Fig. **??**). This observation is consistent with the undifferentiated embryonic stem-cell type lacking lineage-specific enhancer contacts active in its pluripotent state. The same pattern was not seen for enhancers shared between two or more of the cell types under study. We observed a similar enrichment for cell-type-specific transcription but not for several other chromatin states including promoter activity (Fig. **??**).

Together these state enrichments suggest the identified RVS often reflect functional changes at regions of cell-type specific biology, with heightened enhancer and transcriptional activity in the relevant cell type (Fig. **??**). Combined with the observed large-scale concordance of higher order chromatin organisation between cell types (Figs. **??**, **??**), these results reinforce a model of organisation whereby chromatin organisation is largely conserved and static across cell types, but also permits local flexibility in order to activate or repress regions of biological importance to a given cell type.

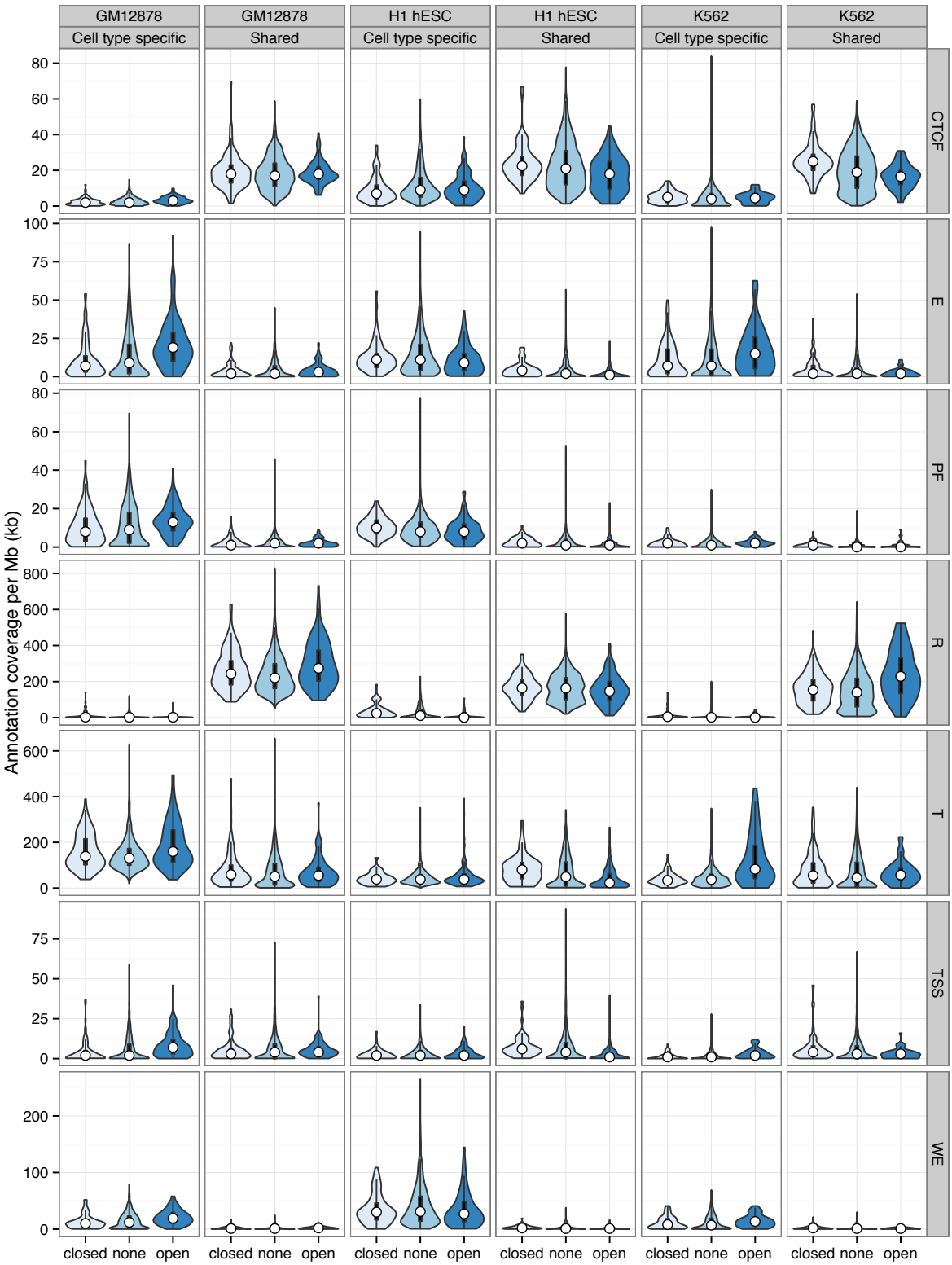

Figure 23: Distributions of features across all predicted chromatin states in regions of variable higher order structure. Distributions of the average coverage of predicted chromatin states in each Mb per cell type are shown as bean plots. Predicted chromatin states are those from **?**] and are labelled as: TSS: promoter and TSS; PF: promoter flanking region; E: enhancer; WE: weak enhancer or cis-regulatory element; CTCF: CTCF enriched element; T: transcribed region; R: repressed or low-activity (Methods **??**).

3.6.2 Gene ontology enrichment

Specific examples of RVS highlight genes of interest (Fig. **??**) but should be coupled with statistical evidence prior to suggestions of a general trend. For this reason we used Gene Ontology (GO) terms to test for functional enrichment within open RVS per cell type.

Functional enrichments of genes found in "flipped open" RVS in each cell type were calculated using DAVID[? ?] and filtered by false discovery rate (FDR $< .05$; Methods **??**). This revealed slight enrichments for keywords "blood", "oxygen carrier" and "β haemoglobin" in the K562 cell type, a multipotent cell type which is known to show properties of an early erythrocyte, among others.[?] However, in the other two cell types we did not find significant enrichments across regions, except for artefacts caused by violations of the independence assumption used in GO term hypergeometric testing. Specifically, our RVS blocks were at least 1 Mb each so generally contain more than one gene, thus enrichments were seen for those gene families known to form paralogous clusters along chromosomes, such as olfactory receptors. The full results of these tests are given in the appendix (Tables **??**, **??**, **??**).

The size of RVS could also explain why we do not capture the relationship hinted at by our cherry-picked examples (Fig. **??**). Given that regions contain multiple (often unrelated) genes, we can imagine a case where a cell type specific locus is activated and moves into a more central position, disturbing adjacent genes which remain in a repressed state. Thus the cell type specific signals contained within the sum of all RVS in a given cell type could be obscured by the noise of adjacent genes captured in these broad compartment transitions.

3.6.3 Contact changes

A defining characteristic of active A compartment regions is a preferential bias in contacting other A compartment regions.[?] However, it is not clear whether cell-type-specific transitions in higher-order structure are solely compartment-level phenomena, or involve other structural strata. We therefore examined the genome-wide contact profiles of each region of variable cell-type-specific chromatin structure in detail.

If our cell-type-specific RVS are mediated by finer-scale structural levels (such as TADs) we might expect to see predominantly short-range contact changes in their underlying contact profile. Instead, we found that variable regions preferentially interact with other A compartment regions in the cell types in which they are active (Fig. **??**), but not in the other cell types in which they are inactive. This supports the idea that these cell-type-specific regions are undergoing compartment-level transitions, disproportionately mediated by the formation of long-range contacts, while also not precluding additional changes at lower levels such as TADs. Furthermore these contact shifts, particularly when coupled with the observed functional enrichments

for transcriptional machinery and enhancer activity (Section **??**), are consistent with active RVS selectively entering into "transcription factories", sites of co-ordinated transcription between potentially distal loci.[?]

3.7 NUCLEAR POSITIONING

Chromosome positioning within the nucleus is known to reflect gene density, with the most gene-dense chromosomes occupying the centre of the nucleus in human cells.[?] **?**] used a Hi-C variant to recreate probability density maps of chromosome positions which again reflected this feature of higher order chromatin organisation, and also reported active regions were more diffuse than inactive. A testable hypothesis with the eigenvector data used in this work is that active A compartments are enriched in the central nucleus of our human cell types, and B compartments are preferentially located in the nuclear periphery.

To test this, published data on chromosome positioning preference within the nucleus was used to label chromosomes as "central" or or "edge".[?] Chromosomes whose DAPI hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by **?**], made up the "central" group and included chromosomes 1, 16, 17, 19 and 22. Similarly the "edge" group had enriched signals ($p \leq 5 \times 10^{-3}$) in the outer shell relative to the inner nuclear shell and included chromosomes 2, 3, 4, 7, 8, 11, 13 and 18. The remaining chromosomes showed no significant preference to either inner or outer nuclear shells at $\alpha = 0.05$.[?]

We found that positive eigenvectors (reflecting A compartments) did show a modest relative enrichment in centrally-positioned chromosomes relative to those located at the nuclear periphery (Fig. **??**). The significance of the difference in distribution of eigenvectors in the central and edge of the nucleus was determined by a two-sided Kolmogorov-Smirnov (K-S) test, with the null hypothesis that there is no difference between the empirical cumulative density functions of the central chromosome eigenvectors ($F_{central}$) and peripheral (F_{edge}). The difference was found to be statistically significant in each cell type ($H_0 : F_{edge} = F_{central}$; GM12878: $D = 0.11$, $p < 6 \times 10^{-4}$; H1 hESC: $D = 0.12$, $p < 8 \times 10^{-8}$; K562: $D = 0.10$, $p < 5 \times 10^{-3}$)

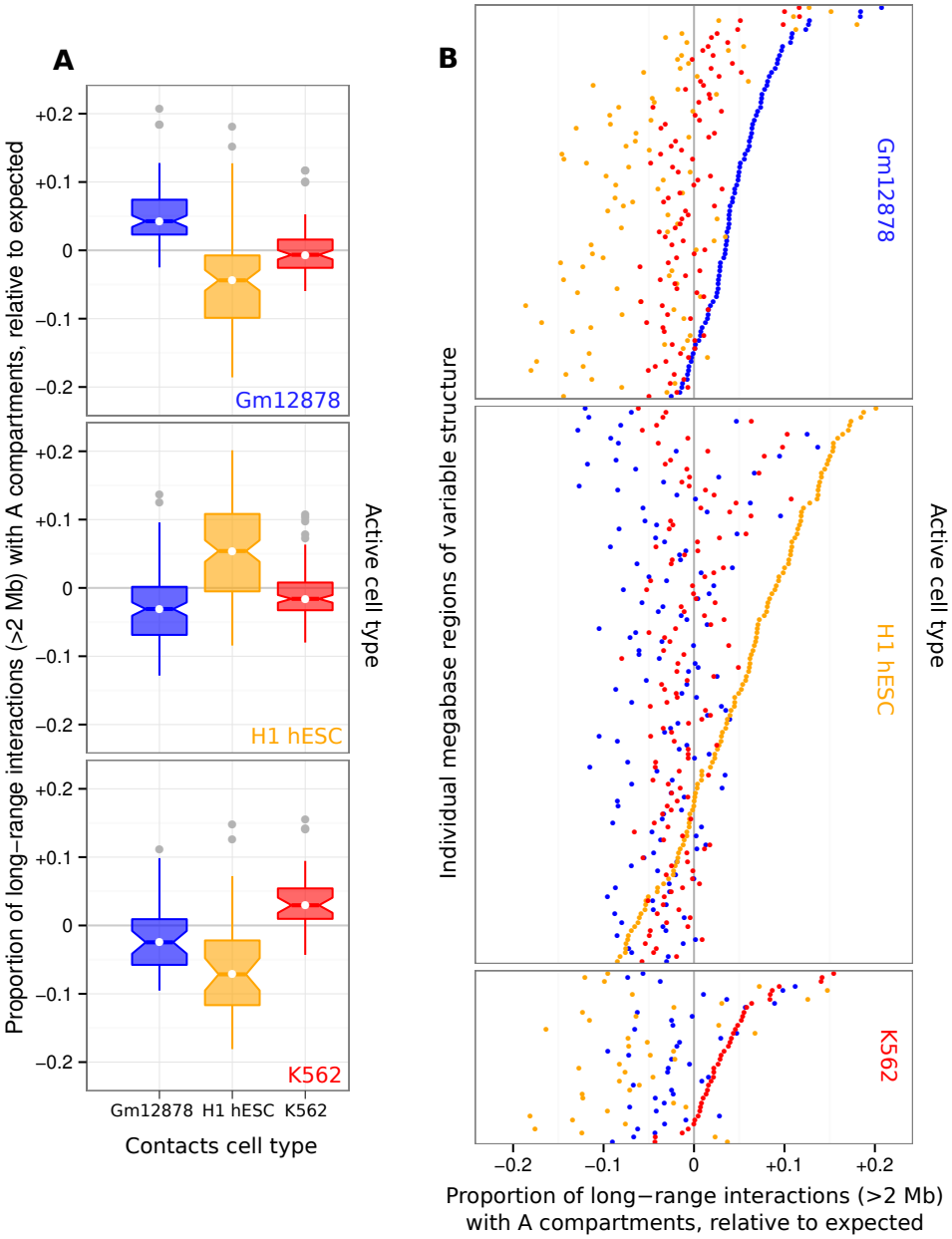

Figure 24: **Regions of variable higher order structure change their genome-wide contact profiles to favour active compartments.** Genome-wide normalised contacts were summed for each "open" region of variable structure and the relative proportion of those that were with active / A compartments is shown across the three cell types used in this study. Proportions were subtracted from the genome-wide average per cell type, such that positive values indicate a greater than expected interaction bias with active compartments. These data are presented both as a summary notched boxplot (A) and with each individual region visualised, sorted by relative proportion of A contacts in active cell type (B).

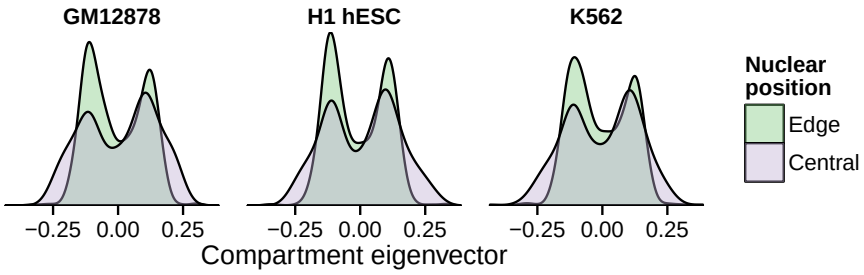

Figure 25: **Chromosomes located at the nuclear periphery hold a greater proportion of inactive B compartments than those in the central nucleus.** Kernel density estimates show the distributions of A (positive eigenvectors) and B compartments (negative eigenvectors) at the edges of the nucleus and at its centre. Positioning data from **?**] (Methods **??**).

4 INTEGRATIVE MODELLING AS A TOOL TO EXPLORE BIOLOGICAL SYSTEMS

4.1 INTRODUCTION

Large-scale chromatin data has recently been produced by multiple consortia, most notably ENCODE[?] (Section **??**) but also the NIH Roadmap Epigenomics project.[?] The breadth and depth of this new data offers unprecedented opportunities to advance our understanding of the complex biology of the chromatin landscape. To this end, studies have already enjoyed success in integrating these data through modelling techniques, with the subsequent dissection of these models revealing novel insights into complex biological phenomena.

Recent studies have shown convincingly that local chromatin state measurements can accurately predict expression levels of genes on a genome-wide basis. **?**] designed a linear model to predict steady-state mRNA levels in mouse embryonic stem cells based on just four predictors: 3 histone modifications (H3K36me3, H3K4me2 and H3K27me3) and Pol-II occupancy. Remarkably, the linear model was found to explain 84.6% of an estimated 91% maximal variance that could be explained (as calculated through a detailed determination of noise).[?] An additional finding of this study was that mRNA half-life and microRNA mediated transcript degradation both had relatively minor influence on steady-state mRNA levels, with the authors concluding that "the lion's share of regulatory contribution is at the level of mRNA synthesis and predictable from chromatin alone."[?] An independent study used a similar regression modelling approach to explore chromatin and transcription factor data and again concluded that models built with histone modifications and chromatin accessibility data were almost as accurate as those which also included binding data for 12 transcription factors.[?]

A key study from the ENCODE consortium, that of **?**], used ChIP-seq datasets to predict gene expression in a range of cell types as measured by a variety of experimental techniques. The authors here developed a two-stage model which first attempts to classify each transcription start site (TSS) into an 'on' or 'off' state using a powerful ensemble classifier technique called Random Forests (RF). The second stage of the model used the same range of histone modifications as regressors in a simple linear modelling framework to quantify predicted expression. This approach proved very successful, producing a median Pearson correlation coefficient (PCC) between predicted and empirical expression levels using 10-fold cross-validation of 0.83 across

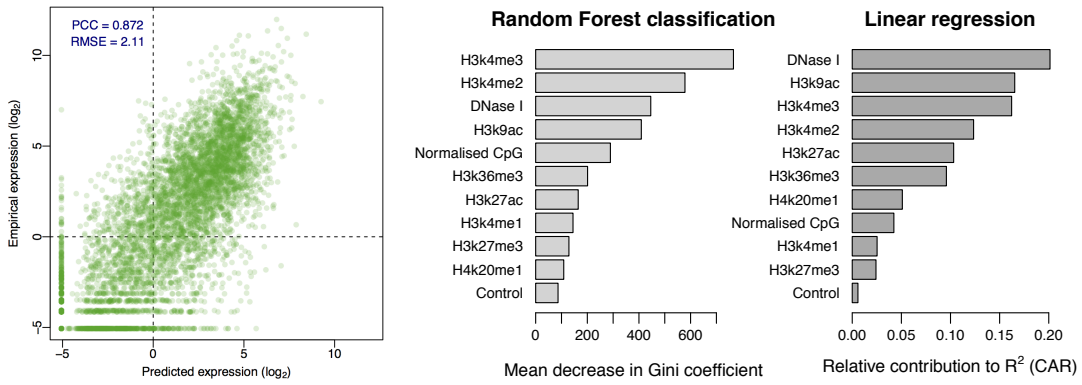

Figure 26: Highly accurate models of expression were built following Dong *et al.* A scatterplot shows the strong correlation between predicted and observed expression levels per transcript (*left*). Variable importance metrics are shown for the "on/off" RF classification step and subsequent linear regression of those loci classified as "on" (*right*).

all cell lines and expression level technologies.[?] Additionally, this study highlighted cap analysis of gene expression (CAGE) as the technology, relative to RNA-seq and RNA-PET, which produced the most predictable expression response. CAGE uses $5'$ capped transcripts to generate short, specific tags which precisely identify TSS positions as well as quantifying the abundance of a given transcript.[? ?]

These recent publications highlight the importance and relevance of advancing our understanding of chromatin biology through a model-based approach. We can extend this idea to the related domain of nuclear architecture, in the hope of understanding the relationships between chromatin and higher order structure in the same way that chromatin features have been related to transcriptional output.

4.2 EXTENDING DONG *et al.*

We reimplemented the published modelling framework of **?**] to replicate their results and analyse the strengths and caveats of their approach.

We were able to reproduce the reported results and generate highly accurate models of transcriptional output based on underlying chromatin features. An example is shown for a predictive model of CAGE transcriptional output in the H1 hESC cell type (Fig. **??**). Note that not all variables used in **?**] were made available for this particular modelling scenario, however the Pearson correlation between predicted and observed expression (0.87) is above the previous study's median value (0.83), and in-line with other models predicting CAGE data (median $PCC \approx 0.87$).[?]

We also re-calculated measures of variable importance for this model of transcription (Fig. **??**). We note some small variations from the example model shown in **?**] which aimed to predict cytosolic CAGE levels recorded in K562 cells. This hints at some

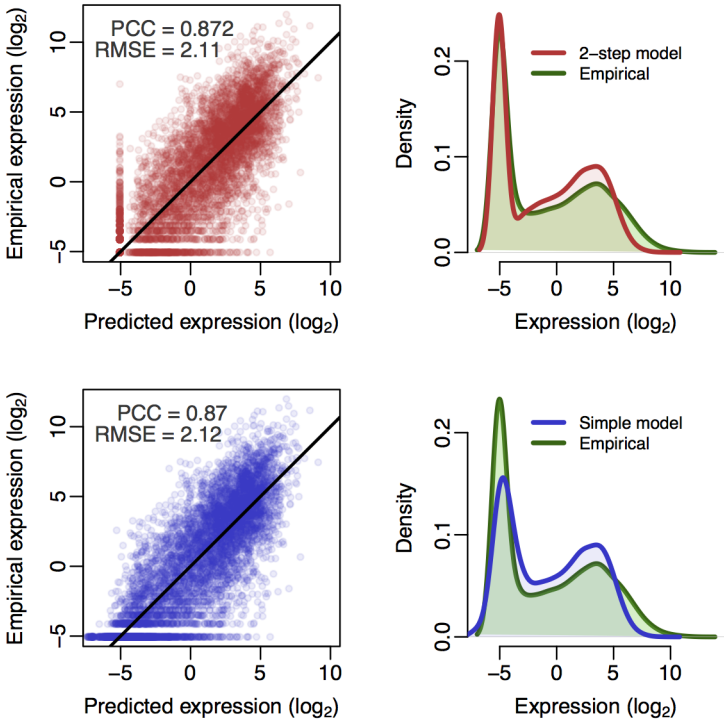

Figure 27: Comparison of a published two-step classification-regression model of transcription with a simple linear regression model. Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for the two-step model of **?**] and simple multiple linear regression are shown (*left*) along with frequency distributions of predicted and observed expression levels (*right*). Scatterplots are annotated with Pearson's correlation coefficient (*PCC*) and the root mean squared error (RMSE); black trendlines describe $y = x$. Overall correlations calculated with 10-fold cross-validation.

degree of some variability between cell type models, though broad similarities also exist such as both models ranking DNase I hypersensitivity as a relatively informative variable (Fig. **??**).

When replicating the modelling approach of **?**], we were surprised to find that the two-step classification then regression (firstly assessing a gene as 'on' or 'off' and then predicting its expression level) added little additional accuracy relative to a simple linear regression model (Fig. **??**). Indeed, it appears the "best bin" technique (explained below, Section **??**) had much greater impact on overall predictive power than the addition of this classification step.

4.2.1 *Bestbin* method

An innovative element of the modelling approach used in **?**] is the 'bestbin' method of matching chromatin measurements to the expression of a given TSS. This strategy first bins normalised signal intensities into 40×100 bp bins encompassing 4 kb around

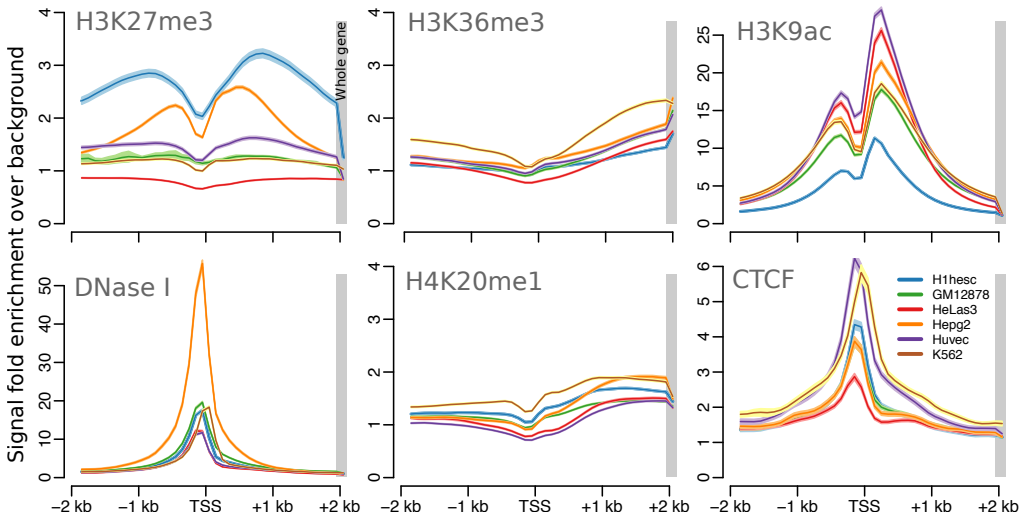

Figure 28: Average input feature profiles over transcription start sites. Mean ChIP-seq signal over input control is shown for 6 factors in 6 human cell types used in **?**]. Each is averaged genome-wide over GENCODE v7 `hg19` defined TSS ±2 kb, and over whole genes (*grey shading*). Ribbons shown 99% confidence intervals of the mean.

the TSS, and adds an additional bin representing the remaining gene body. Then the correlation between the signal of a given mark and the expression of a TSS across all genes is measured and the bin producing the highest correlation is designated as the 'bestbin'. The normalised ChIP-seq signal intensity of this bestbin is then used as an input feature for training the model of transcription. This strategy was shown to increase model performance, measured in terms of the Pearson correlation between predicted and observed expression, by 0.1 in the simple regression model, an increase of almost 13% relative to simply taking the average value across all bins.[?]

The justification for such an approach hinges on the idea that the multitude of input features (mostly histone modifications and DNA binding proteins) have a variety of biological functions, and so the bestbin method is one way of learning these functions in an automated and unbiased way. For example, the histone modification H3K36me3 is understood to be painted across exons that are being actively-transcribed,[? ? ?] thus the genome-wide summary statistic that best captures this function is likely the whole-gene measurement, rather than the level of H3k36me3 at a gene's TSS or upstream promoter. A re-analysis of ENCODE data used in **?**] highlights this kind of variability across input features (Fig. **??**). Some features are clearly enriched directly over the TSS (CTCF, DNase; Fig. **??**) while others show enrichments along the gene body (H3K36me3, H4K20me1; Fig. **??**), and still others show well more complex, asymmetrical "shoulder" patterns (H3K27me3, H3K9ac; Fig. **??**). Bestbin will therefore, to some degree, capture these spatial relationships without *a priori* specification.

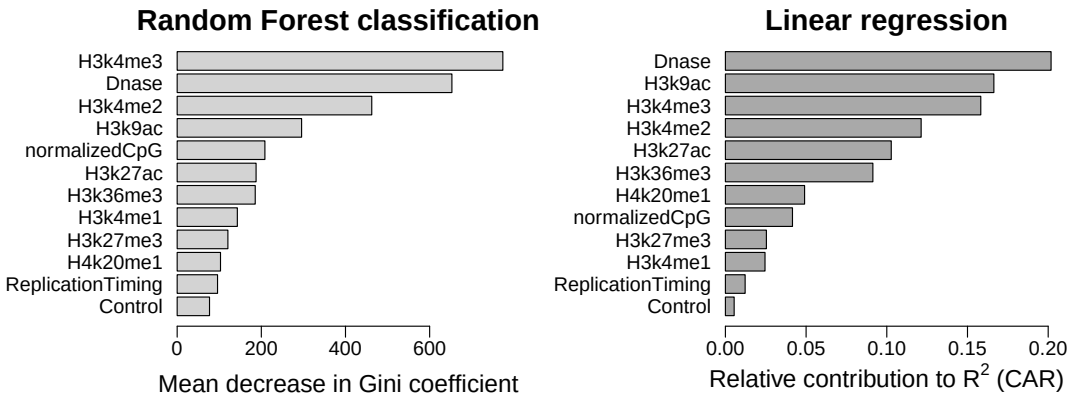

Figure 29: Variable importance metrics for each stage of our reimplementation of a published model for predicting transcriptional output. Variable importance was measured by decrease in Gini impurity for the RF classification step (Methods **??**), and by CAR R^2 decomposition[?] for the linear regression step.

4.2.2 Model exploration

We attempted to improve the accuracy of predicted expression values produced by **?**] through increasing the number of informative regressors. While **?**] included broad coverage of different histone modifications, they did not investigate the impact of higher order chromatin data. For this reason, we matched the TSS positions used in **?**] with previously-published genome-wide replication timing ratios measured in BG02 ESCs.[?] This data is of a somewhat different origin to the transcriptional data in this case (which was recorded in H1 hESC) but replication timing is thought to be largely conserved between cell types, and in particular would be expected to be very similar between two ESC lines.[?]

We then used these values as an additional regressor in both the two-step classification regression model and the simple linear model but saw no significant improvement in either model's accuracy (*data not shown*). Possible reasons for this include that the data were relatively low-resolution (1 Mb) and derived from a imperfectly matched cell line. However, since the existing model is already achieving such accurate results that its original feature set must already be accounting for most of the maximal explainable variance in gene expression given experimental and biological noise. With this in mind, any additional regressors would be expected to yield diminishing returns. Even so, on closer examination, the replication timing data appeared only slightly more informative than the control ChIP-seq input measurements when evaluated with relative importance metrics (Fig. **??**), implying that large-scale chromatin domains have relatively little influence on the expression of the genes resident within them.

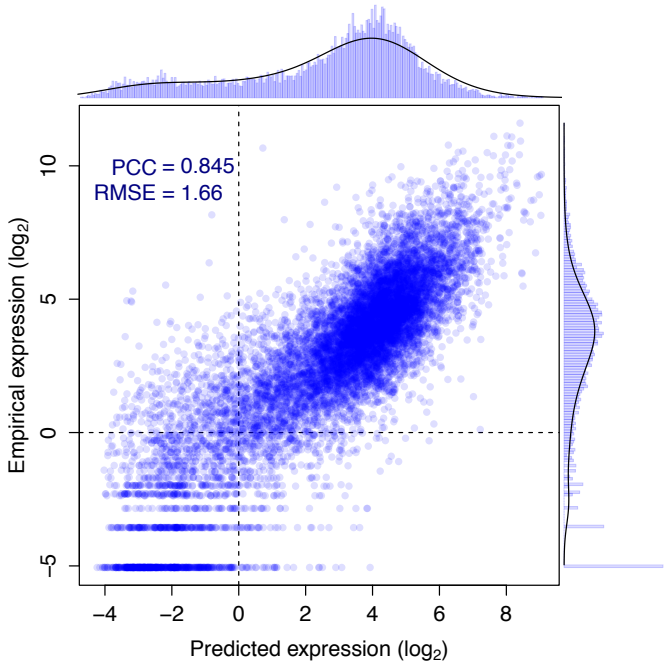

Figure 30: Random Forest predictions of FANTOM5 expression data. RF model predictions are plotted against their empirical values. The marginal distributions of predicted and empirical expression values are shown opposite their respective axes. Summary metrics of Pearson's correlation coefficient (PCC) and the root mean-squared error (RMSE) are also shown (*inset*).

4.3 MODELLING FANTOM5 EXPRESSION DATA

Using FANTOM5 CAGE data[?] and the approach established above (Section **??**), we next attempted to model gene expression at timepoint zero (t_0) of a differentiation timecourse of human H1 embryonic stem cells (H1 hESC) to CD34+ hematopoietic stem cells. Applying this modelling strategy to a novel dataset will allow us to assess the portability of the model design, as well as enabling further analysis of model components such as the bestbin strategy.

We retrieved a number of genome-wide ChIP-seq datasets measured in H1 hESC cells and produced by the ENCODE consortium[?] (Methods **??**). These were matched to transcript annotations to build an input feature set for use in building a predictive model of transcriptional output.

Due to the finding that a two-step (classification–regression) approach added little additional modelling accuracy (Fig. **??**), we employed a single-step design using a Random Forest (RF) regression model.[? ?] With a total of 14 predictors (10 histone modifications, HDAC6, H2A.Z, DNase I and an input control, listed in Methods **??**), we were able to build a highly accurate predictive model of transcriptional output spanning around 11,000 TSS (Fig. **??**).

Model predictions evaluated with 10-fold cross validation show a highly significant correlation with measured CAGE levels ($PCC = 0.845 \pm 1 \times 10^{-4}$, $p < 2 \times 10^{-15}$), and the model is able to explain around 71% of the variance in the expression response (for comparison a linear model resulted in $PCC = 0.825 \pm 3.2 \times 10^{-5}$, $p < 2 \times 10^{-15}$). This result is less impressive than that of **?**] who achieved cross-validated correlation coefficients of up to 0.9, but it is roughly equal to their median test set correlation of 0.83.[?] Though the RMSEs of our predictions, when normalised by the range of observed values, compare more favourably (0.11, compared with Dong *et al.*'s: 0.14).

Our slightly lower predictive power could be explained by our streamlined model design. **?**] implemented a pseudocount optimisation step whereby an additional count was added to each binned signal intensity prior to log transformation to maximise expression correlation. In the model presented above, a fixed pseudocount of 1 was used to avoid introducing an unwarranted positive bias towards higher correlation. We confirmed that a two-step classification–regression design did not improve our model performance metrics; indeed, the PCC and RMSE of a classification–regression framework with this data showed a slight decrease in prediction accuracy ($PCC = 0.834 \pm 0.007$, RMSE $= 1.77$ when applied to the same test and training data used in Fig. **??**).

4.3.1 *Bestbin* location

We again implemented the previously-described 'bestbin' strategy[?] (Section **??**) to objectively select the most-correlated binned signal for each chromatin H1 hESC mark. To explore the implications of this approach, we analysed the stability of chosen bestbins by calculating them on 200 sets of 1000 randomly selected TSS samples, with each sample representing approximately 8% of the complete dataset. Distributions of chosen bestbins across these 200 sub-samples are shown as boxplots (Fig. **??**).

We find that bestbin selections are often highly consistent across sub-samples, indicating there are fairly static informative regions relative to a TSS for each chromatin feature. Furthermore, the selected bestbins match known biological mechanisms; for example the H3K36me3 bestbin is consistently the whole gene measurement (Fig. **??**) and this mark is known to be enriched in actively transcribed exons.[? ? ?] The negative control variable (ChIP-seq input) shows no strong location bias, as expected (Fig. **??**). Other distributions are intriguing but less easily explained, such as those features showing a tight distribution of informative regions slightly downstream of the TSS (H3K9ac, H4Kme2/3 and H3K79me2; Fig. **??**). In the case of H3K9ac, we note that the selected bestbin appears to coincide with the highest summit of its bimodal average profile over all TSS (shown in Fig. **??**).

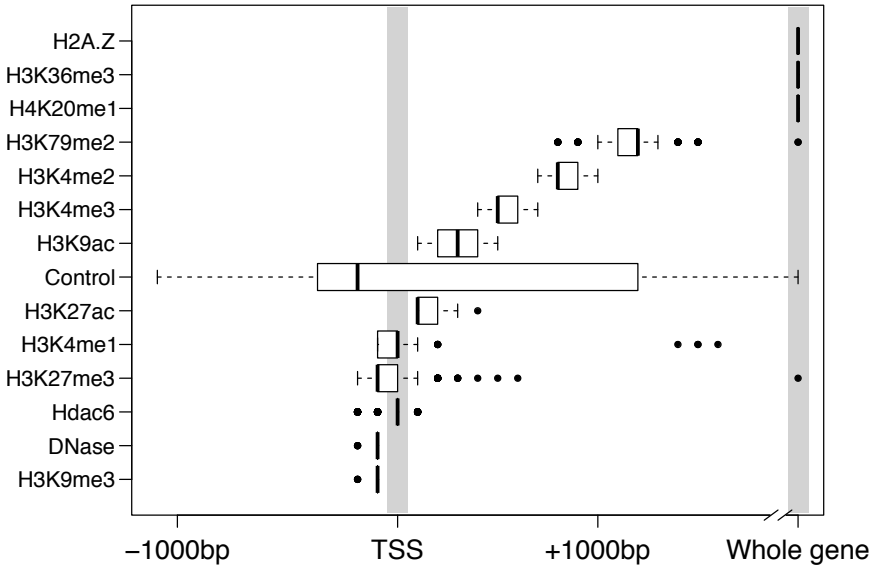

Figure 31: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, DNase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 kb flanking the TSS, but more distal bins were never selected and hence are not shown. 'Whole gene' represents the averaged signal intensity from TSS to transcript end site, as defined by Ensembl Genes v69.

4.4 MODELLING HIGHER ORDER CHROMATIN

Accurate predictive modelling of transcription in a variety of cell types offered several novel insights into the interrelationships between locus-level chromatin features and transcriptional machinery, as well as advancing a quantitative explanation of the degree to which correlated features are informative. It is of interest then, to test whether this approach can be applied to other data, such as the reprocessed higher order chromatin organisation data assembled in this work (Chapter **??**).

Previous publications have identified several correlates which track compartment eigenvector profiles to varying degrees,[? ?] yet to date these relationships have not been quantitively investigated. The above-described modelling framework offers a statistical approach towards understanding the drivers of these observed correlations.

4.4.1 Predictive model

We built Random Forest (RF) regression models (Methods **??**) to predict compartment eigenvector profiles genome-wide in three human cell types. Models were found to have high predictive accuracy, with Pearson correlation between predicted and observed compartment eigenvectors in the range of 0.75–0.82 (Fig. **??**), comparable to that achieved by **?**] in the prediction of transcription.

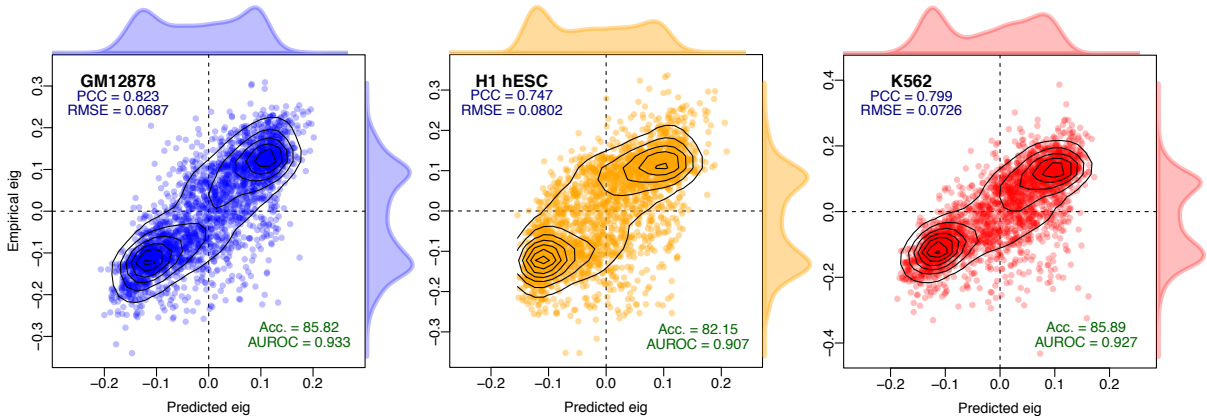

Figure 32: Compartment eigenvector model predictions are highly correlated with observed values. Pearson correlation coefficient (PCC) and root mean-squared error (RMSE) report the degree of success of the regression model, whereas accuracy (Acc.) and area under the receiver operating characteristic (AUROC) give the classification accuracy of binarized outcomes.

Our predictive models were also assessed in terms of classification performance, i.e. did the model correctly assign each block to an A or B compartment. Instead of training a classifier, thereby constructing a second model, we threshold our regression predictions at 0 (Methods **??**). We found our RF models achieved high classification accuracy with $\geq$ 82% of all 1 Mb genomic bins correctly assigned in each cell type (Fig. **??**).

This predictive performance underlines the strong connection between locus-level features and higher order chromatin structure previously noted by **?**] Given such highly-predictive models can be generated, it is then of interest to dissect said models in an attempt to understand the nature of this captured relationship.

4.4.2 Cross-application

High predictive accuracy on cell type specific models could be the result of "overfitting". In machine-learning, overfitting refers to the point at which parameters are being optimised to capture the random errors of a particular sample, on top of any underlying relationship between inputs and predictions, thereby giving an inflated estimate of model performance which would not translate to another feature set with an independent noise profile.[? ?] Commonly Random Forest models are thought not to suffer from overfitting, but while it is true that this is not an issue when increasing the size of a forest, the technique itself can indeed overfit, particularly in the instance of fully-grown regression trees whose leaf nodes can contain a single case.[?]

To test if overfitting was responsible for our high modelling accuracy, we cross-applied models learnt in one cell type to unseen input data from each of the other

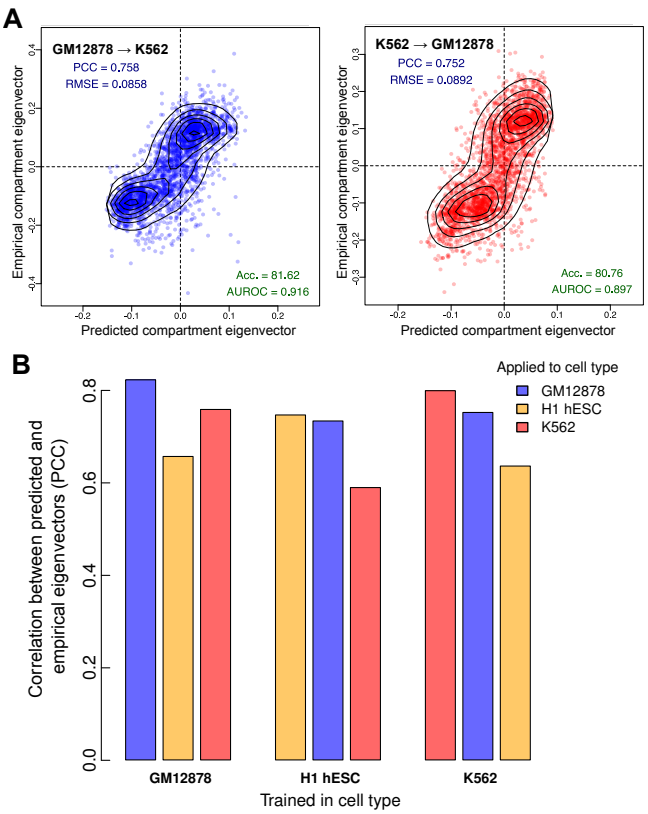

Figure 33: Models of higher order chromatin structure learned in one cell type can be cross-applied to others. Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. (A) The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross (PCC = 0.75). Inset metrics are the same as those shown in Figure **??**. (B) In each case, predictive accuracy decreased on cross-application but there remains significant agreement between predicted and empirical values.

two cell types under study. If predictive accuracy is a lot lower on unseen data, this lends evidence to the idea that our models may overfit to their respective cell types. Conversely, it could be the case that biologically-distinct mechanisms are in place that differ between cell types, preventing a simple cross-application.

We found cross-application between cell types was possible and resulted in similarly-high levels of accuracy to within cell-type cross-validation (Fig. **??**). This gives good evidence not only that these models are not overfitting to cell-type specific noise, but also that there exist broad rules linking chromatin conformation and locus-level features. Model performance under cross-application suggests that there are enough commonalities for compartment profile predictions to transcend the cell-type specific biology inherent to an embryonic stem cell or differentiated cell type.

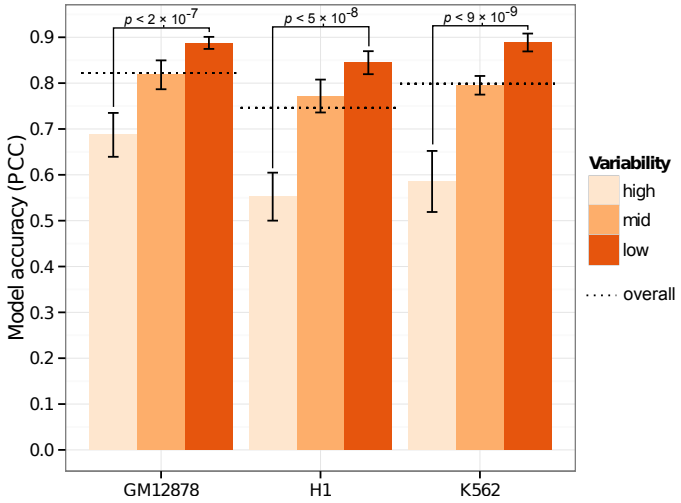

Figure 34: Genomic regions that vary across cell types are modelled less successfully than static regions. Genome-wide compartment eigenvectors were partitioned into thirds according to their median absolute deviation (MAD) across the three cell types under study (Methods **??**). This bar chart compares the accuracy of models fitted independently to each third, according to the Pearson correlation coefficient (PCC) between predicted and observed values.

4.4.3 Between-cell variability

Given much of the higher order chromatin organisation is conserved between the three cell types used in this work (Fig. **??**), a testable hypothesis is that these conserved regions are drivers of cross-applicability between cell types. Under this hypothesis we might also expect those genomic regions which vary most across cell types to be more difficult to predict.

Consistent with this we found the most variable regions across cell types were the most difficult to predict through our RF modelling framework (Fig. **??**). In each cell type, the third of the genome with the most consistent compartment eigenvectors across cell types could then most accurately be modelled in each cell type, and conversely the most variable third showed significantly depleted predictability (Fig. **??**). This result suggests these variable regions could either be those which are noisiest, where the eigenvector is least able to capture compartment structure, or where cell-type specific biology is influencing compartment structure in ways not captured by our input feature set and low resolution modelling pipeline. Results presented in Section **??**, showing that regions of variable structure are enriched for cell type specific enhancers and transcription, is suggestive of this latter explanation.

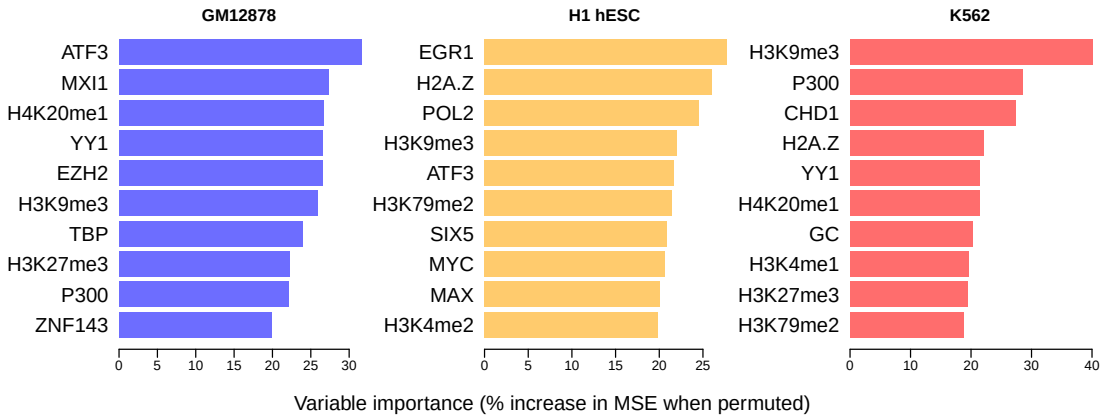

Figure 35: Variable importance per cell type specific model. Variable importance for each Random Forest model was measured in terms of percentage increase in mean squared error when permuted (Methods **??**) and the top 10 ranking variables are shown.

4.4.4 Variable importance

Having built accurate predictive models, we next dissect the relative variable contributions made from our range of input features and compare these across cell types. An overview of the top 10 most highly-ranked features in cell type specific models shows some agreement but also notable differences between cell types (Fig. **??**)

Only one input feature, H3k9me3, is present in the top 10 most important variables of each model (Fig. **??**). H3k9me3 is one of the few features to be negatively correlated with compartment eigenvectors, hence offers orthogonal information to the majority of other, positively-correlated input variables (Fig. **??**; Section **??**). Of those important variables shared between two cell type models, H3k27me3 is also a repressive mark and deposited by polycomb repressive complex 2 (PRC2)[?] while H2A.Z is a histone variant again linked to polycomb-regulated genes and essential for embryonic development.[?] Furthermore EZH2, the catalytic subunit of PRC2,[?] is also included in the feature set and is highly ranked in the GM12878 cell type model (Fig. **??**). Other interrelated and important variables include MYC and MAX, which are found in the top 10 influential variables in H1 hESC, and MXI1, found to be an informative variable in GM12878. Recent results suggest MYC binds open chromatin as a transcriptional amplifier in embryonic stem cells,[? ?] with MAX and MXI1 acting as antagonistic co-regulators.[?] These biological relationships between variables may help explain the observed differences between models: different representatives of correlated clusters of input variables are likely being selected in each model (this is explored in Section **??**).

To assess the significance of observed intersections (Fig. **??**), the variable selection process could be modelled with, for example, a multivariate hypergeometric distribution or via simulation. Simulation was used here for simplicity: each intersection

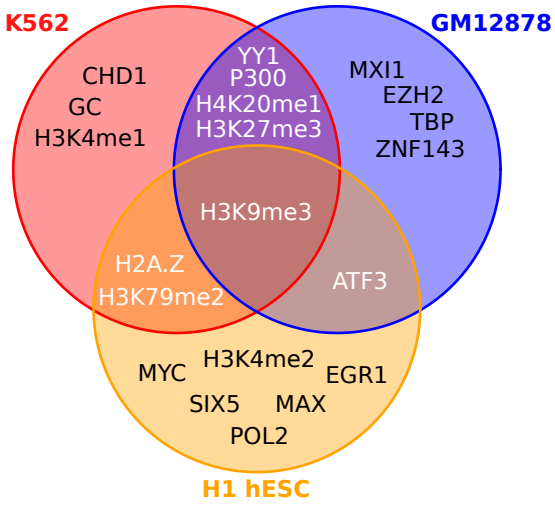

Figure 36: Intersections of the top 10 ranked variables in cell type specific models. Venn diagram showing the intersections between sets of ten most influential variables calculated in each cell type specific Random Forest regression model of compartment eigenvector (Fig. **??**).

was calculated under 10,000 variables draws with uniform distribution and empirical p-values were then calculated accordingly. Under the assumption that variables are ranked independently in each cell type, drawing at least one variable in all three cell types would be expected by chance ($p = 0.6$). Similarly, the overlaps between pairs of cell types is within the range of expectation (probability of 7 or more variables appearing in exactly two sets: 0.39). Hence these data suggest the top 10 most influential variables are not significantly more alike across the three cell-type specific models than expected by chance, however 10 is an arbitrary cutoff, and many of the rankings are based on small differences in variable importance, thus could be unstable between multiple generations of stochastic RF models.

In addition to rankings, raw variable importance metrics can also be compared between cell-type specific models (Fig. **??**). Through this analysis we found that variables such as CTCF have a relatively small but highly consistent variable importance across the three cell type specific models, whereas other features like ATF3 are highly influential in one cell type but not the other two. Absolute differences in these figures should not be over interpreted and will be affected to some degree by data quality, eigenvector calculation and other sources of noise. Nevertheless there are observations which may reflect biological phenomena, such as the higher relative importance of P300 in both haematopoietic cell line models, potentially reflecting its activity as a histone acetyl transferase that regulates haematopoiesis,[?] compared with the more consistent influence of CTCF in each model, an insulator and transcription factor widely known as a regulator of genome architecture (discussed in Section **??**).

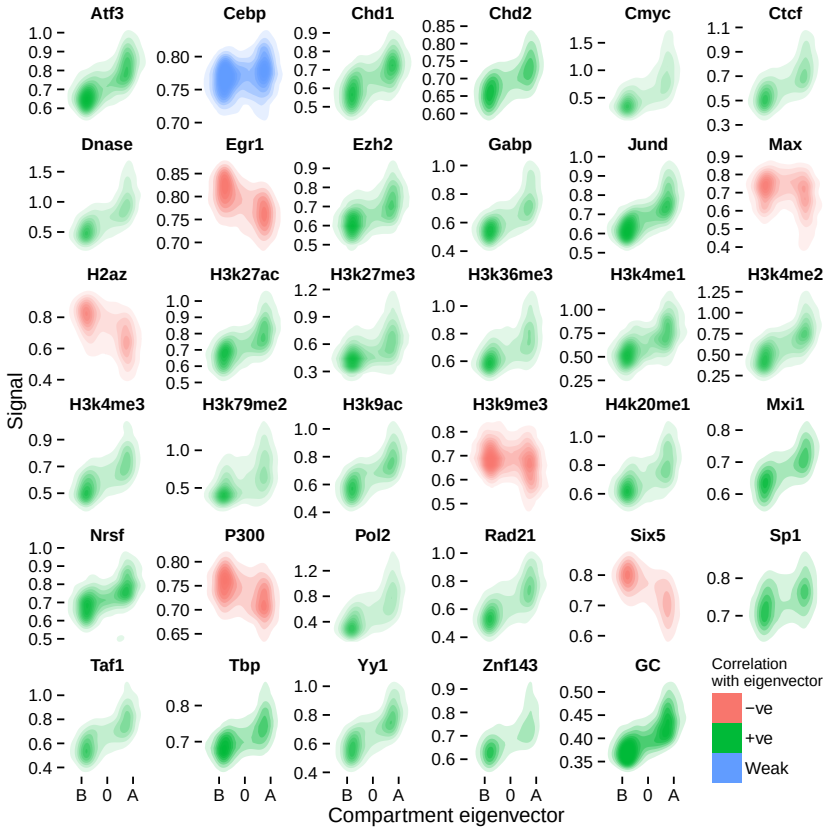

Figure 37: Correlations of individual features with compartment eigenvector in the H1 hESC cell type. Two-dimensional kernel density estimates show the density of points in a scatterplot of compartment eigenvector (x-axis) against each input feature individually (y-axes). Features with a PCC against eigenvector of above or below 0.1 are coloured as positive or negative, respectively.

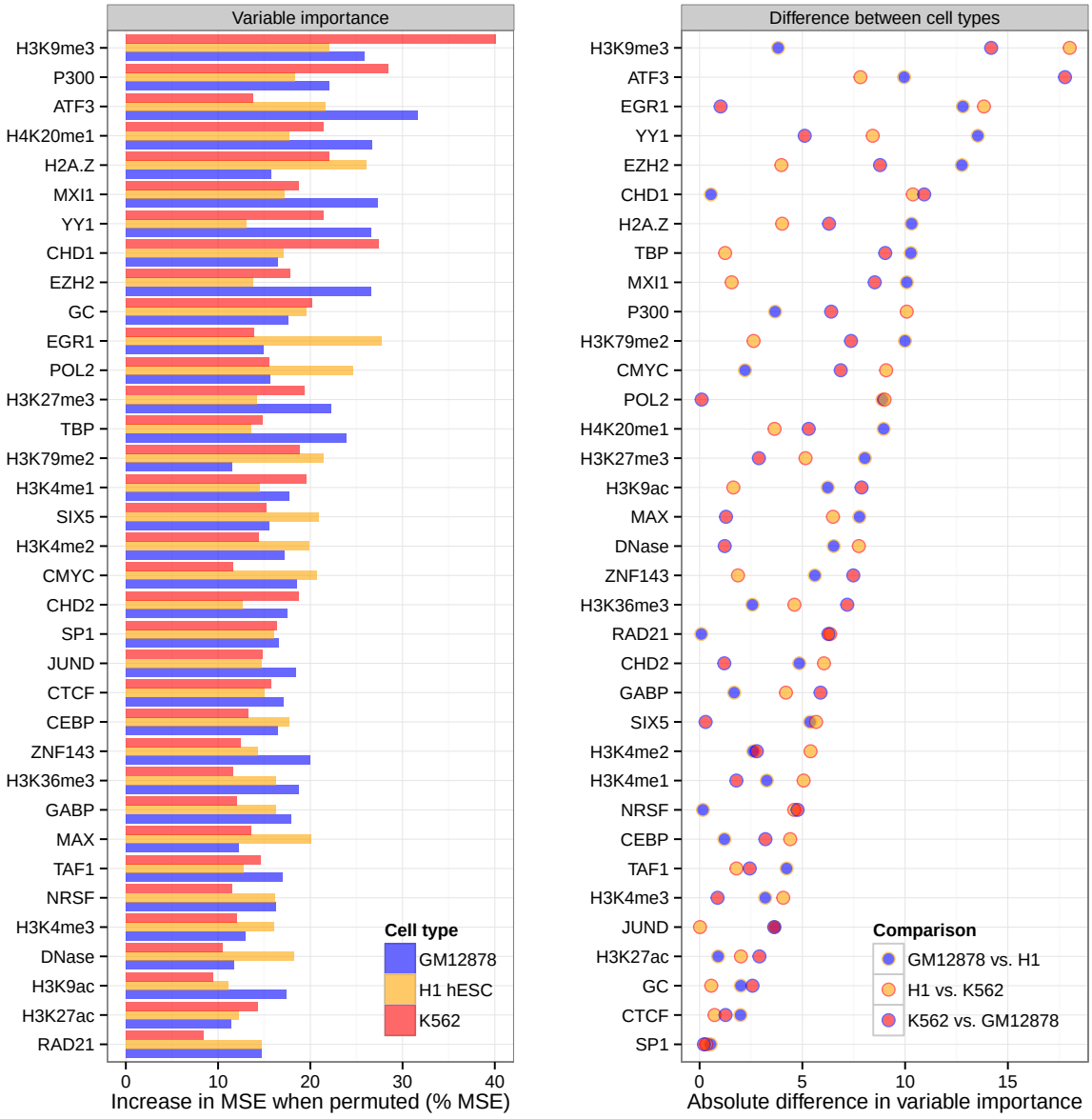

Figure 38: Comparison of variable importance between three cell type specific Random Forest models. Variable importance for each Random Forest model was measured in terms of percentage increase in mean squared error when permuted (Methods **??**). Results are shown sorted by mean variable importance (*left*) and by largest absolute difference in pairwise comparisons (*right*).

4.4.5 Correlating input features

We have an *a priori* expectation of multicollinearity in our feature set, for example between those that each broadly correlate with transcriptional activity (including POL2, H3K36me3 and sequence GC content). To explore these relationships, we performed unsupervised clustering of our feature sets in each cell type (Fig. **??**).

We found pervasive multicollinearity across our feature sets, with the majority of input variables in each model falling into a persistent "active" cluster containing regions with high DNase hypersensitivity, POL2 binding and histone modifications H3K36me3 as well as GC content (Fig. **??**).

Outliers are also present. H3K9me3, noted for high variable importance in each model (Fig. **??**) and the only feature ranked within the top 10 in each model (Fig. **??**) is a clear outgroup in the H1 hESC and GM12878 correlation heatmaps, and in K562 forms a stable cluster only with the P300 transcription factor (Fig. **??**). This suggests H3K9me3 is providing orthogonal information to many of the other input variables, and likely explains its high variable importance.

4.5 TECHNICAL CONSIDERATIONS

4.5.1 Resolution

Thus far models were built at 1 Mb resolution, but if we are capturing true biological relationships we would expect these to hold at higher or lower resolutions. To test this, models trained at 1 Mb resolution were applied to feature sets binned at 100 kb, an order of magnitude higher resolution.

Model accuracy when applied to higher resolution input features proved to be similarly high, with empirical PCC being 88 to 95% as high as that at 1 Mb native resolution (Fig. **??**). We found that eigenvectors at higher resolution, beyond 100 kb, do not necessarily reflect A/B compartmentalisation.

Note however, there is some indirect leakage between test and training set when 100 kb bins have been used in aggregate in learning the 1 Mb models. Nevertheless, sustained accuracy is evidence that our models are not resolution-sensitive, and could likely be applied to higher resolutions than the 1 Mb predominantly used in this work.

4.5.2 Other modelling approaches

Random Forest (RF) was *a priori* chosen as an appropriate and powerful modelling tool for this work, but other methods could have been used and should be compared.

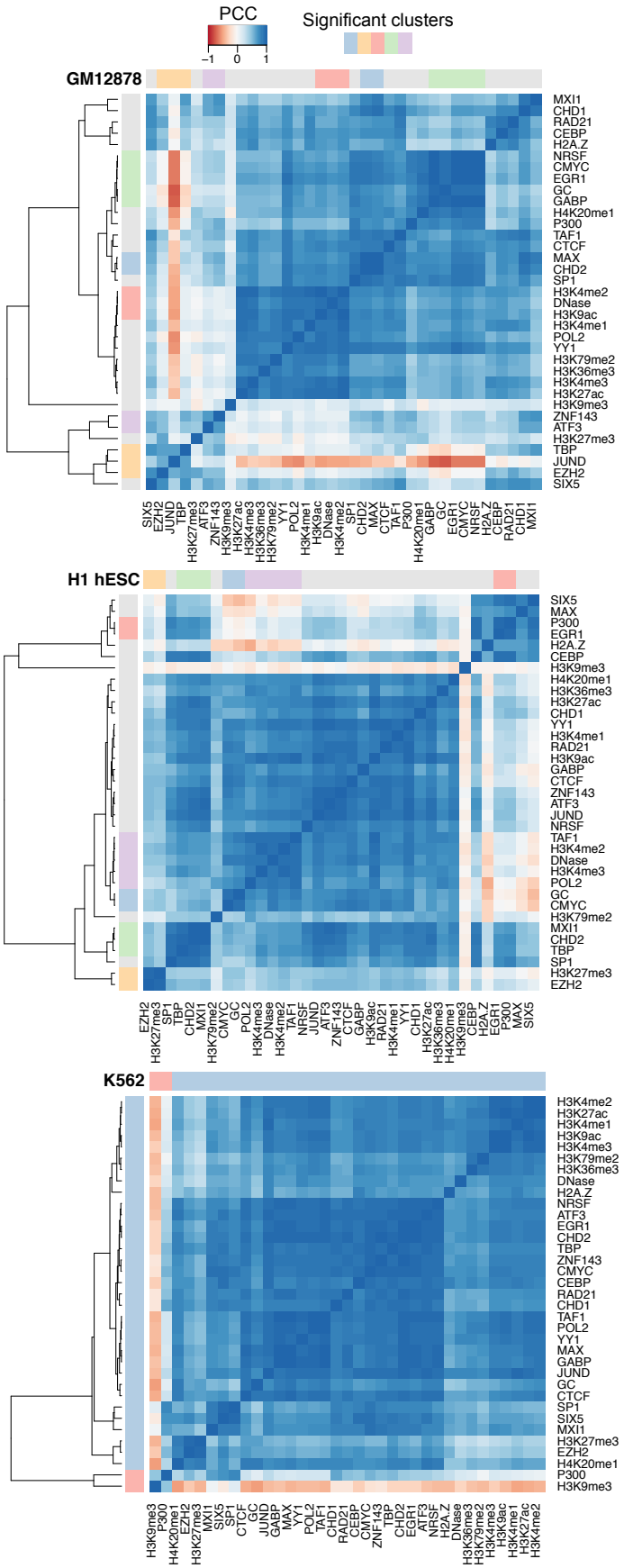

Figure 39: Correlation heatmaps of the 35 features used to model compartment eigenvectors. The Pearson correlation coefficient (PCC) of genome-wide 1 Mb bins of each feature were pairwise correlated with each other. The features were also clustered using hierarchical clustering. The significance of these clusters was determined through multi-scale bootstrap resampling, with those clusters that were stable across different sizes of resampling deemed significant (Methods **??**). Such clusters are labelled with coloured blocks in heatmap sidebars.

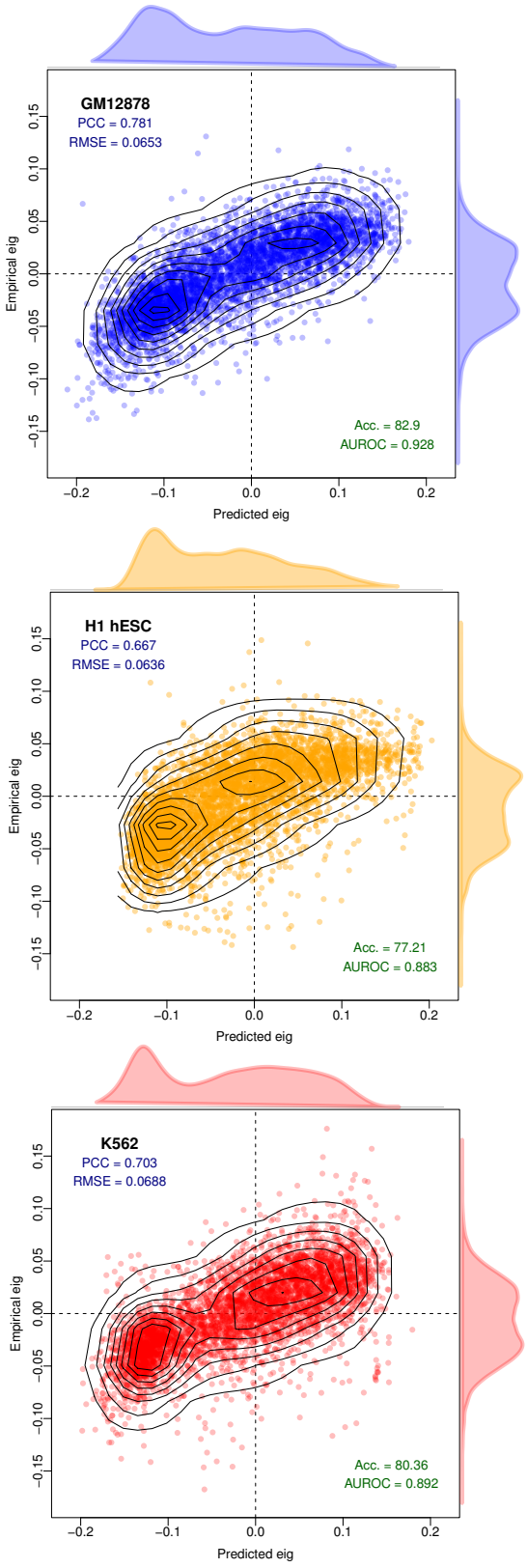

Figure 40: Models trained at 1 Mb resolution can be applied to higher resolution datasets. Despite having been trained on lower resolution training sets, the Random Forest models generated can successfully predict compartment eigenvectors at higher resolution (100 kb, a 10× zoom).

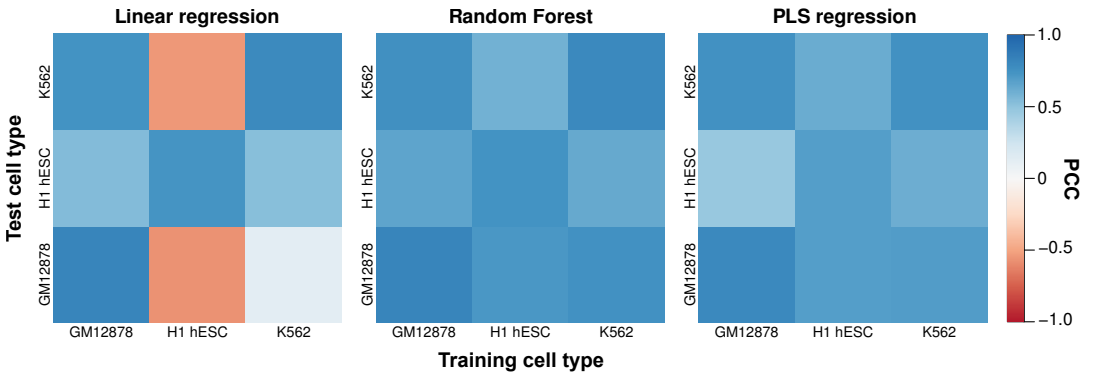

Figure 41: Comparison of Random Forest performance with other modelling approaches. Heatmaps show the Pearson correlation coefficient between predicted and observed compartment eigenvectors genome-wide for three regression techniques: multiple linear regression (LM), Random Forest (RF) and partial least squares (PLS). Results are summarised in Table **??**.

Table 4: Performance comparison of different modelling techniques. Comparison of mean Pearson correlation coefficient between predicted and observed compartment eigenvectors for three different modelling approaches: LM: linear regression; RF: Random Forest regression; PLS: partial least squares regression. Correlations were averaged per cell type over three cell types (cell type specific) and in the six possible crosses (cross-application) shown in Fig. **??**.

	LM	**RF**	**PLS**
Cell type specific	0.787	0.790	0.750
Cross-application	0.139	0.689	0.641

Here we compare our RF approach with two other options: multiple linear regression and partial least squares regression (Methods **??**).

Our results confirm RF as a suitable and powerful approach for modelling our relationships of interest in this work (Fig. **??**), with both the highest cell-type specific performance (PCC between predicted and observed = 0.790) and on cross-applications (mean PCC = 0.689).

Multiple linear regression assumes linear relationships between model parameters and input features and allows for simple, normally-distributed errors. Surprisingly, this simple approach is capable of accurate cell-type specific predictions (mean PCC = 0.787; Table **??**), likely due to the high raw correlation between the inputs and dependent variable. However this simple approach fails to cross-apply between cell types (mean PCC = 0.139; Table **??**) indicating problems with overfitting. This can be remedied through variable selection procedures, however a strength of the RF approach is that this step is not necessary, and pre-selection of model variables may result in a sub-optimal end result.[?]

Partial least squares regression is a technique that uses dimensionality reduction to engineer a lower-dimension and orthogonal feature set. Hence this method is well-suited to collinear inputs, such as the set of variables used in this work (e.g. Fig.

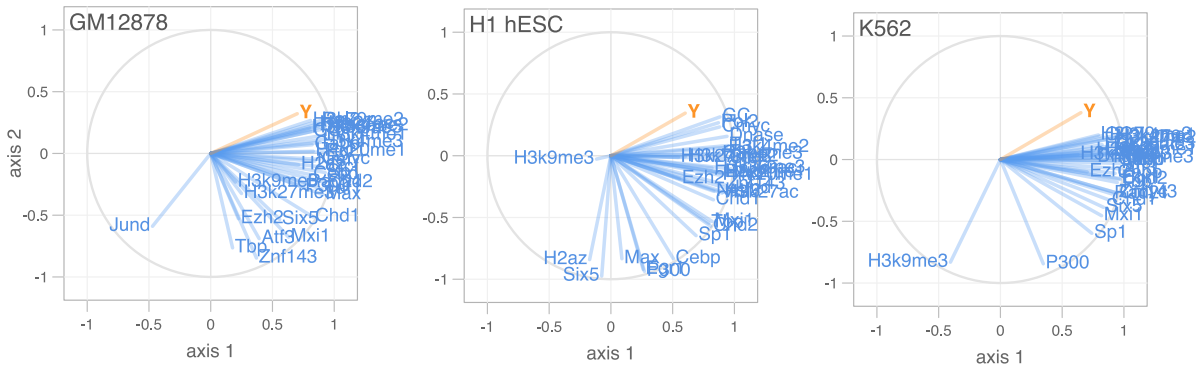

Figure 42: Circle of correlations of variables compared with PLS axes. Model variables are plotted against the first two components used in PLS regression models per cell type. *Y* represents our compartment eigenvector.

??). As expected, PLS regression provides highly accurate cell type specific predictions (mean PCC = 0.750; Table **??**) and performs well during cross-application (mean PCC = 0.641; Table **??**), though in both cases produces slightly inferior results to RF models (Fig. **??**).

PLS uses a type of dimensionality reduction, which offers another way to explore the inter-relationships between our feature set. Plotting input features against these lower-dimension components can give a revealing insight beyond simple correlations (e.g. Fig. **??**). Figure **??** shows a "circle of correlations", where features are plotted onto polar co-ordinates against the first two PLS components (Methods **??**). Nearby variables in the scatterplot are positively correlated, and the vector length from the circle centre is proportional to the variable's representation in the model. Negatively correlated variables point in opposite directions while uncorrelated variables are orthogonal to each other.[?] We therefore see the known multicollinearity represented as groupings of overlapping variables in each cell type, with a smaller number of orthogonal and negatively correlated variables in each cell type (Fig. **??**).

4.5.3 Non-independence

As recognised through our use of hidden Markov models (Methods **??**), eigenvector values for consecutive bins along a chromosome are non-independent yet thus far predictive models have not considered this inter-dependence.

This is for two reasons: firstly non independence could be thought of as an artefact of bin-sizing (we have elected to use regular, fixed binning beneath the scale of compartments themselves whereas another approach could use variable bin sizes, for example per compartment, TAD or restriction fragment); secondly using information of a bin's surroundings may obscure by proxy the chromatin features which would

otherwise prove predictive. As an example, knowing that bin x_{i-1} and bin x_{i+1} are in compartment state A would allow us with high confidence to predict the state of bin x_i, but without learning anything of the region's relationship with its encompassed histone modifications and bound factors.

4.6 PARSIMONIOUS MODELS FROM EXPANDED FEATURE SETS

Strongly predictive models can be useful tools to reason about a complex system, however from a researcher's perspective there also exists a trade-off between predictive power and parsimony. Namely simpler models with fewer inputs may be more interpretable and of wider utility, for example they could be applied to cell types with less ChIP-seq data available than those used in this work. For this reason we explore parsimonious models with reduced feature sets, with an aim to build simpler models of chromatin state while retaining, if possible, similar levels of predictive accuracy.

On the other hand, the 35 variables used thus far as model inputs are not the complete set available in each cell type, but only the subset of those assayed in all three cell types under study. The ENCODE consortium has produced a significantly greater number of datasets[? ?] in each cell type which have thus far gone unused. Here we explore models of higher order chromatin structure, in some cases built from over 100 variables, and then generate parsimonious models using optimal subsets guided by statistical techniques that penalise model complexity.

4.6.1 Stepwise regression

Multiple linear regression is a simple and analytically well-described modelling framework which is amenable to regularisation through a variety of methods. A simple approach is to start with a complete model and serially remove and/or add variables, then calculate a metric (here we use the Bayesian information criterion, BIC) which weighs the the model likelihood against model complexity. This process is iterated until the metric reaches a (local) minimum, thus creating a more parsimonious model which retains predictive accuracy and should be less prone to overfitting. Stepwise regression also aids interpretation by selecting representative features from collinear clusters.[?] A detailed explanation of this feature selection procedure can be found in Methods **??**. It should be noted that despite its continued widespread usage, several statistical issues have been identified with the stepwise procedure for model selection.[? ?]

In terms of model performance alone, stepwise regression gives the highest predictive accuracy on a held-out validation set in each cell type specific model of compartment eigenvector (Table **??**), however it must be said that differences in model

Table 5: Performance comparison of full and optimised RF and LM models. PCC between predicted and empirical compartment eigenvectors is shown for a range of modelling scenarios, including multiple linear regression (LM) and Random Forest (RF) approaches. For model selection, two methods are used: stepwise BIC-regularised linear models and LASSO regression; in each case those same features were then also used in building a separate RF for comparison.

	GM12878			H1 hESC			K562		
	n	LM	RF	n	LM	RF	n	LM	RF
All features	115	.836	.828	71	.744	.755	187	.811	.813
Matched subset	35	.827	.823	35	.740	.747	35	.796	.799
LASSO ℓ_1	23	.823	.836	23	.734	.750	39	.779	.811
Stepwise BIC	21	.840	.831	13	.746	.738	27	.819	.810

performance across all comparisons are modest. These results do show that even expanded feature sets of up to 187 input features add little explanatory power beyond that of much less complex models with 20 or fewer input variables (Table **??**).

4.6.2 LASSO regression

A more modern technique for regularisation of linear models is the least absolute shrinkage and selection operator (LASSO). In brief, the LASSO is a form of ℓ_1 regularisation that penalises the sum of absolute values of standardised regression coefficients. By penalising total absolute values, rather than squared values as in ℓ_2 regularisation, coefficients can be shrunk to 0 thereby removing terms from the model.[? ?] Thus LASSO combines the coefficient shrinkage of techniques like Ridge regression with a type of feature selection as seen in stepwise regression. A more rigorous description of this method is given in Methods **??**.

Again we can perform a simplistic comparison of model performance using LASSO regression and other techniques (Table **??**). LASSO retrieves comparable numbers of informative variables to the stepwise regression technique in each cell type, and again removes the majority of input features from expanded sets as redundant or relatively uninformative.

Of those variables with a non-zero coefficient at the optimally-selected tuning parameter λ (Methods **??**), the ten largest in each cell type are shown (Fig. **??**). Similarities can be observed with variable importance from previous 35 input models (Fig. **??**), including the large (negative) coefficient for EGR1 in the H1 hESC model as well as that of P300 in K562 (Fig. **??**). Also of note is that Zinc finger protein 143 (ZNF143) appears among the largest model coefficients in two of the three cases (Fig. **??**). Recently ZNF143 was found to be a novel chromatin looping factor which connects promotors and *cis*-regulatory elements[?] and previous studies have found it to be enriched over chromatin domain boundaries.[? ?] Here however, ZNF143 has a

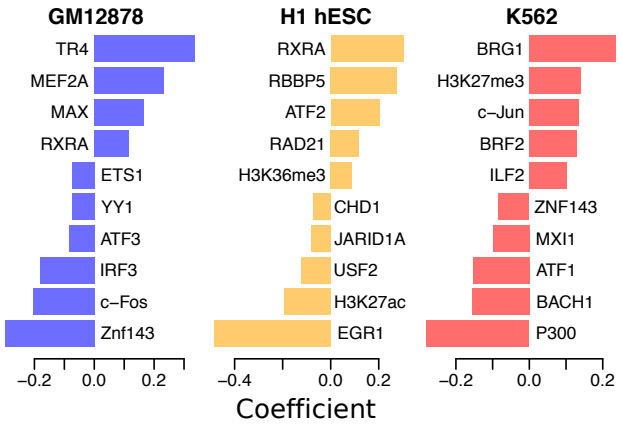

Figure 43: Ten largest LASSO coefficients in models derived from expanded feature sets. Those coefficients with the largest absolute standardised value are plotted for each cell type specific LASSO model.

negative coefficient in models of compartment eigenvector, indicating some additional role in heterochromatin B compartments.

Of interest is the appearance of a group of factors known to collectively form the heterodimeric activator protein-1 (AP-1), these include c-Fos, c-Jun and ATF1–3; all are spread across the most highly-ranked variables in each model of chromatin organisation (Fig. **??**). The AP-1 complex has been shown to have DNA bending properties,[?] and recently FOS and JUN members were associated with long-range chromatin interactions[?] suggesting an under-explored role for this complex in genome organisation.

4.6.3 Regularised Random Forest

Random Forest (RF) comparisons are included for comparison in Table **??** where RF models were built using those features selected by procedures based on linear regression. Thus the linear regression-based feature selection acts as a "filter" method for feature selection, fully independent of the RF learning algorithm. A more coherent approach might be an "embedded" method, where a regularisation procedure is integrated with the learning algorithm.[? ?]

While RF is a much younger technique than linear models, a framework for Regularised Random Forests (RRF) has recently been described[?] and implemented in the R package `RRF`.[?] The RRF algorithm uses the idea that at each node in a tree, unused variable should only be included if they offer a significant information gain over those available variables which have already been used in the tree. This differs from the standard RF algorithm where splitting decisions at each node are entirely independent of each other (Methods **??**).

We found that this algorithm was unable to perform feature selection on our highly collinear feature set, instead leaving full or almost full feature sets in each case (*data not shown*) and so providing equal results to a standard RF model using expanded feature sets (Table **??**). Potentially this problem could be investigated using a rapidly-advancing set of techniques known collectively as "deep learning". These cutting-edge machine learning methods are capable of learning both a predictive model and concurrently how input features should best be represented within this model, often using multiple layers of connected neural networks (for reviews, see **? ?**).

5 CHROMATIN DOMAIN BOUNDARIES

5.1 INTRODUCTION

A succession of studies have defined chromatin domains of different types, for example: A and B chromosome compartments;[?] topologically associating domains (TADs);[?] contact and loop domains;[?] physical domains;[? ?] and others.[?] The existence of these domains necessitates "boundary regions" either between consecutive domains or bookending more separated domains, however the functional relevance of said boundary regions is still open to debate.

In their study of topological domains, **?**] identified average enrichments over TAD boundary regions in both human and mouse for various features including CTCF and PolII. Boundaries were also enriched for signs of active transcription, such as with the histone modification H3k36me3.[?] These results, coupled with an observable enrichment for promoters at domain boundaries, have lead to the theory that boundaries may act as an additional layer of transcriptional control.[?] However an alternative theory is that if chromatin domains represent co-regulatory regions as is widely thought,[? ? ?] boundaries themselves could be mere side-effects and as such of limited biological interest.

An obvious experiment to resolve these opposing theories would be to delete a predicted boundary region and test for local changes in both contacts and expression. Such an experiment was performed on a region of the human X-chromosome containing the genes encoding the dosage-compensation long non-coding RNAs *Xist* and *Tsix*, which are separated by a TAD boundary.[?] This study found that while histone modifications within the body of a TAD could be removed without affecting overall domain structure, deletion of a boundary did have an effect and led to increased intradomain contacts.[?] Surprisingly however, the two domains did not completely merge, lending credence to the alternative theory that TADs may be centrally constrained, rather than by their borders.[?]

A recent experiment used CRISPR genome editing to link TAD boundary changes with limb development disorders,[?] indicating that boundary changes could provide an underlying explanation for pathogenic non-coding structural variants.[?] Similarly, domain boundaries on *C. elegans* X-chromosomes were found to be weakened following the disruption of condensin binding sites.[?] Together these studies suggest a complex scenario whereby TAD boundaries are an important structural feature, yet do not fully explain domain partitioning.

Many questions remain about chromatin boundaries. For example, are the enrichments reported in **?**] persistent across cell types and how do they compare across organisation strata, such as compartments and TADs? Through computational analysis of the set of boundaries re-called from published datasets, we can investigate these questions and probe boundary enrichments across a broad array of locus-level chromatin features.

5.2 BOUNDARY ANALYSIS

The mammalian genome is organized into topologically associating domains (TADs), predominantly self-interacting chromatin domains, with boundary regions reportedly associated with pronounced peaks and troughs of particular features within 500 kb of the predicted boundary.[?] Exploration of this phenomenon using a set of 24 mouse ESC chromatin features (and a smaller number of human ESC features) revealed enrichment peaks of CTCF, H3K4me3 and H3K36me3, as well as a pronounced dip in H3K9me3, suggesting that high levels of transcription may contribute to boundary formation.[?] However, the peaks and dips of these features lacked any estimates of statistical significance. It was also unclear whether other features might show unusual patterns in TAD boundary regions, and how the constellation of features involved might vary between cell types. Moreover, the features associated with boundaries separating A and B compartments calculated from Hi-C eigenvectors have not been studied to our knowledge. The datasets assembled here, consisting of 35 matched chromatin features across three cell types, allow us to conduct the first comparative study of the constituents of human TAD and compartment boundary regions.

5.2.1 TAD boundaries

We derived TAD boundaries from uniformly reprocessed Hi-C data (Chapter **??**) according to established methods (see Methods **??**) for all three cell types under study. We then sought evidence for significantly enriched or depleted features at TAD boundary regions using a conservative approach (a nonparametric statistical test and Bonferroni multiple testing correction, see Methods **??**).

Our findings confirmed the previously reported peaks (CTCF and POL2) and dip (H3K9me3) in ESC data, but also revealed substantial heterogeneity between cell types and some novel boundary features. CTCF binding was found enriched at TAD boundaries across all cell types, but other features, including H3K27me3 and H3K4me3, show dramatic peaks of enrichment in H1 hESC cells that are not seen consistently in other cell types (Fig. **??**). Although the dip in H3K9me3 at TAD boundaries is seen in all cell types, the extent of the depletion varies and is weakest in H1 hESC

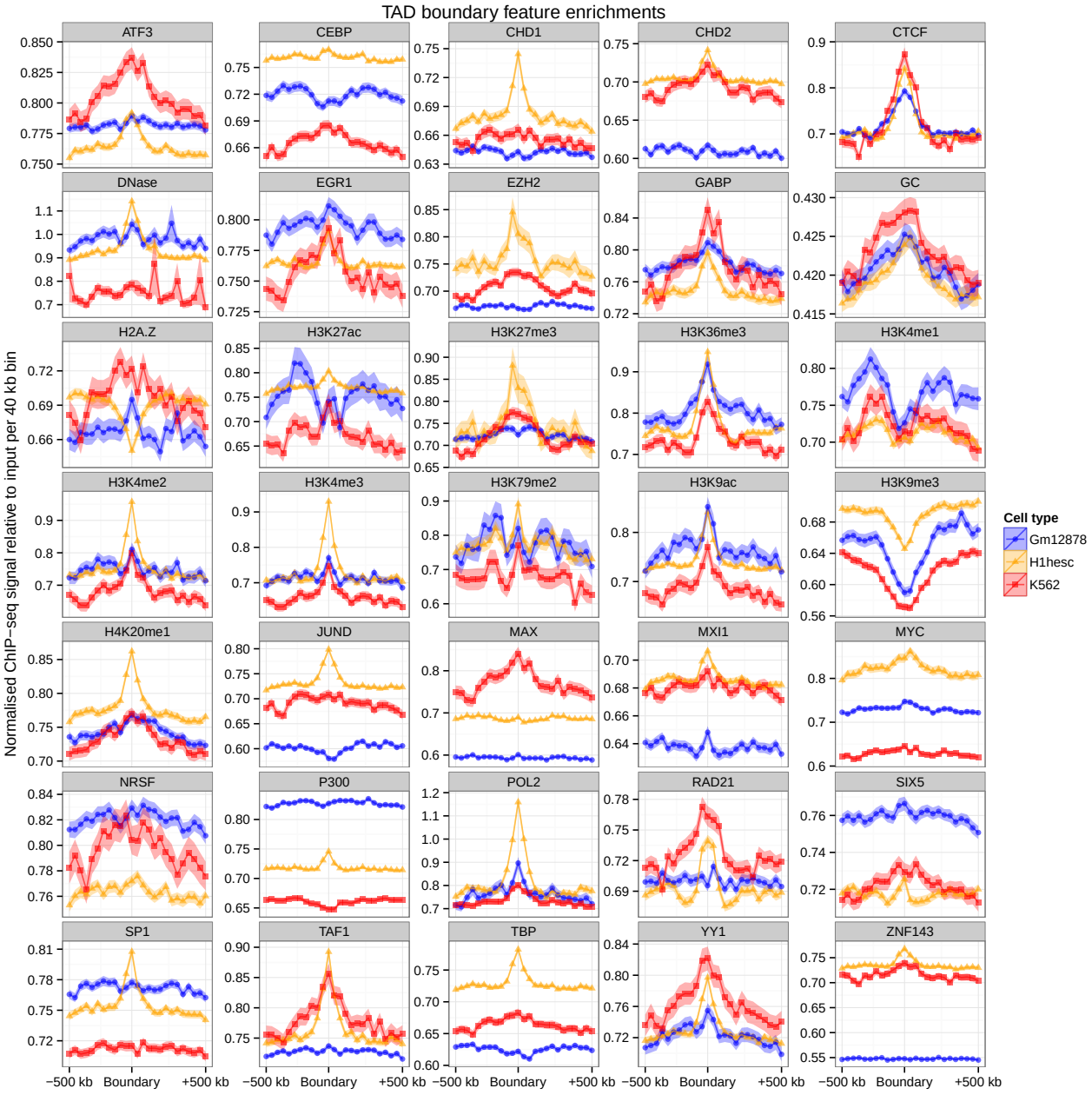

Figure 44: TAD boundary enrichments and depletions. 36 features were averaged over 1 Mb windows centred on TAD boundaries genome-wide (25 × 40 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

cells. We also note an apparent enrichment of H4K20me1 over TAD boundaries in H1 hESC cells, a modification previously implicated in chromatin compaction.[?] Finally we observe consistent increases in GC content at TAD boundaries, at a scale that is difficult to explain in terms of those much smaller GC-rich features such as binding motifs or CpG islands (Fig. **??**). The statistical significance of these enrichments and depletions is considered in Section **??**.

5.2.2 Compartment boundaries

Where neighbouring genomic regions occupy contrasting A and B nuclear compartments, the disparity implies the presence of a boundary region. We identified putative compartment boundaries using an HMM to infer the state sequence of A/B compartments across the genome based on observed principal component eigenvectors (Section **??**). Analogously to the TAD boundary analysis we then sought significant enrichments or depletions in our set of chromatin features over these compartment boundaries.

Compartment boundaries display similar spectra of enrichments to previously studied TAD boundaries[?] but at lower resolution, reflecting the different scales of these levels of organisation (Fig. **??**). Peaks associated with active promoters (POL2, TAF1, H3K9ac) are again evident. Enrichments of CTCF and YY1 are again seen at compartment boundaries, as they were for TAD boundaries, in each cell type under study. In addition, compartment boundaries show enrichments of H3K79me2, which is known to play a critical role in cellular reprogramming.[?] This histone modification has also recently been shown to mark the borders of small (hundreds of bp) regions of open chromatin,[?] hinting at similarities in chromatin boundaries at very different scales.

Certain features show intriguing contrasts between cell types. For example, the histone variant H2A.Z shows a clear enrichment over K562 compartment boundaries, but not the other two cell types (Fig. **??**). Compartment boundaries also show enrichment for the cohesin complex subunit RAD21 in the two hematopoietic cell types, and cohesin is another factor implicated in modulating nuclear architecture in partnership with CTCF.[?] Various other enrichments of modest effect size can also be seen (Fig. **??**). In contrast with TAD boundaries, the composition of compartment boundaries appears least complex in H1 hESC, relative to the other two cell types (cf. Fig. **??**).

5.2.3 Significance testing of boundary associations

Domain boundary associations with other chromatin features are most frequently presented through the statistics-free "average-o-grams" similar to those presented

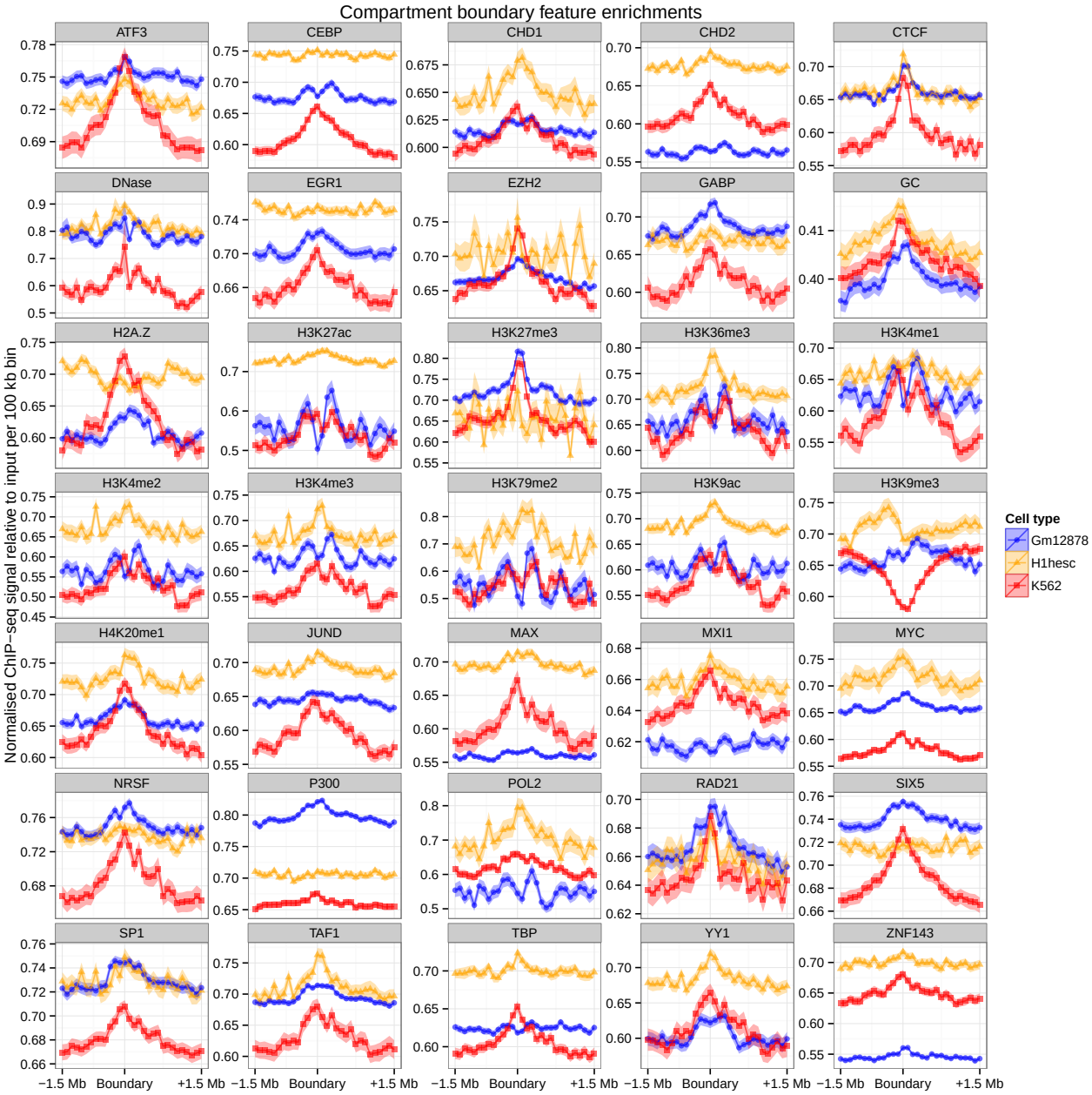

Figure 45: Compartment boundary enrichments and depletions. 36 features were averaged over 3 Mb windows centred on compartment boundaries genome-wide (30 × 100 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

in this work (Figs. **??**, **??**). We then went on to perform a quantitative test of the associations between these features with compartment and TAD boundaries. In brief, we compare the values for each normalised signal directly over a boundary bin with those peripheral bins in each $\approx$ 1 Mb window (for TADs). We then perform a non-parametric, rank-based significance test to look for statistically significant enriched or depletions of each feature. Details of this procedure are given in Methods **??**.

Results of this significance testing, along with selected boundary profiles, are presented in Figure **??**. Overall we find that compartment and TAD boundaries are associated with overlapping spectra of highly significant enrichments and depletions of chromatin features across cell types. Across all tests (Fig. **??**), we frequently find boundary enrichments for DNA binding proteins implicated in chromosome architecture (e.g. CTCF, YY1, RAD21), but also note broad enrichments for classes of input features associated with active transcription (e.g. POL2, TBP, H3K9ac).

Reflecting their different scales, we find enrichment and depletion profiles typically spanning regions of up to 500 kb for TAD boundaries but those over compartment boundaries often span more than a megabase (Fig. **??**). We expect part of the reason for this discrepancy is the resolutions at which domains were called: TADS are resolved to 40 kb bins while compartment boundaries fall between megabase-sized bins. A further consideration is that larger numbers of TAD boundaries were called in the H1 hESC cell line, due its more deeply-sequenced Hi-C library (Section **??**), giving greater statistical power to detect enrichments thus resulting in smaller p-values (Fig. **??**). The lower resolution megabase bins used in compartment calling do not suffer from this issue.

5.2.4 CTCF and YY1

Significant boundary enrichments for both CTCF and YY1 are evident in all cell types at both compartment and TAD boundaries (Fig. **??**), which is intriguing given the evidence that YY1 and CTCF cooperate to affect long distance interactions.[?] Co-binding of CTCF with YY1 has also been shown to identify a subset of highly conserved CTCF sites.[?] This colocation may also therefore be a contributing factor in the establishment of TAD boundaries, which appear to be broadly conserved across mammals.[?]

To test this, we split our sets of TAD boundaries into those possessing ChiP-seq peaks (region peaks as called by the ENCODE data processing pipeline[?]) for CTCF, YY1, both CTCF and YY1 (overlapping peaks) and neither. We then tested each boundary subset for genome-wide enrichments of the other features in our dataset (Fig. **??**). Unexpectedly, we found that those boundaries marked by YY1 but without overlapping CTCF peaks were generally most strongly-enriched for other features in our dataset. This result potentially highlights YY1 as an under-appreciated

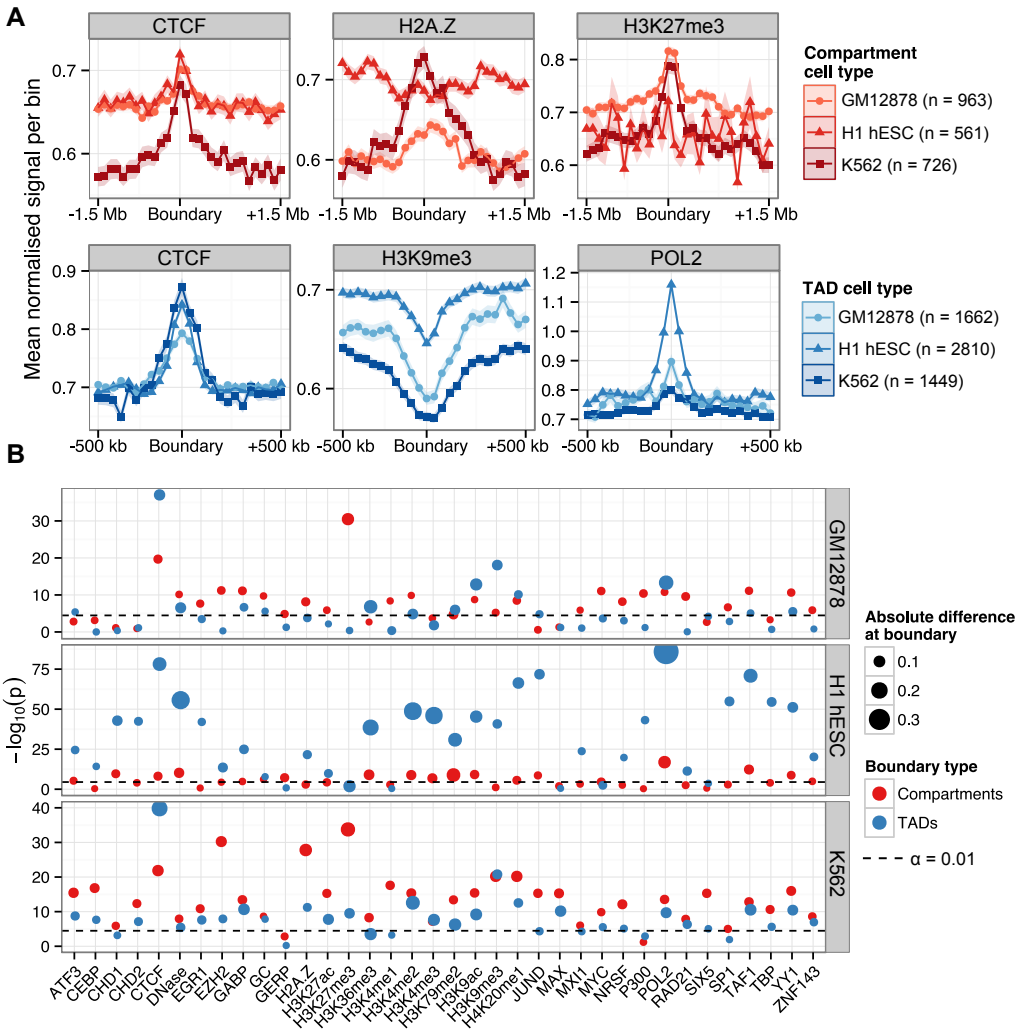

Figure 46: Compartment and TAD boundary enrichment summary in three human cell types. (A) Selected profiles for locus-level features are shown for compartment boundaries (CTCF, H2A.Z and H3K27me3) and TAD boundaries (CTCF, H3K9me3 and POL2), as a mean normalized ChIP-seq signal relative to input chromatin per bin (± 1 standard error). TAD boundaries were examined over 40 kb bins over the 1 Mb flanking each boundary; compartment boundaries were examined in 100 kb bins over 3 Mb. (B) The significance of enrichment or depletion ($-\log_{10}(p)$ two-tailed Mann–Whitney test) of a feature was calculated as the boundary bin relative to the ten most peripheral bins (five either side). Points are scaled by the absolute mean difference in signal over the boundary relative to the mean of peripheral bins.

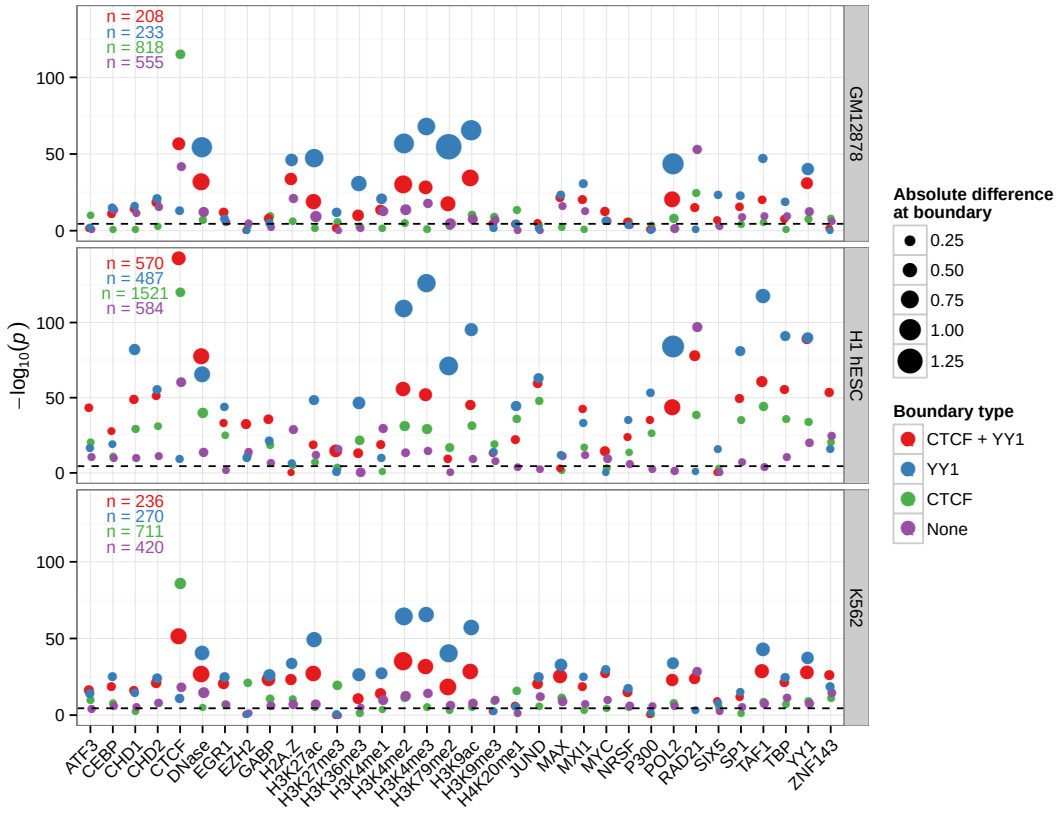

Figure 47: Distinct enrichments of CTCF and YY1 boundaries. The significance of TAD boundary enrichments and depletions are shown (as in Fig. **??**) for boundaries split into classes based on the presence or absence of ChIP-seq peaks within boundary bins. CTCF and YY1 groups are those boundaries with at least one ENCODE region peak[?] for their respective features, while CTCF + YY1 is the group of boundaries which had one or more overlapping peaks for these two factors. Boundaries in the "none" group had neither a CTCF or YY1 region peak called (but can still be enriched for their respective features in terms of raw signal).

contributor to boundary demarcation, particularly relative to the well-studied CTCF. We also found that boundaries lacking both CTCF and YY1 peaks showed instead the strongest enrichments for RAD21 in each cell type (Fig. **??**), reinforcing previous findings that describe the distinct influences of CTCF and cohesin in organising chromatin structure.[? ? ?]

5.2.5 Repeats

?] identified short interspersed element (SINE) repeats as being enriched over TAD boundaries and suggested roles for these repeats in altering genome organisation, in line with prior evidence.[? ?] For example, SINE elements are thought to be responsible for spreading CTCF binding sites through mammalian genomes during evolution.[?] Analysis of recent high-resolution Hi-C data again reported a SINE B2 link with CTCF loops in mice,[?] and examples have been reported of human genes whose expression has been altered by CTCF sites inserted through Alu repeats.[?] Together these results suggest repeats could be a key component in the makeup of domain boundaries.

To investigate this, we used the `RepeatMasker`[?] software package to call repeat classes and families in the `hg19` and `mm10` genome assemblies. Counts for each annotated feature were then averaged over boundaries as described previously (Methods **??**).

At the level of repeat class, we corroborate the findings of **?**] that the majority of repeat classes show no enrichment or depletion at TAD boundaries, and we find that this also holds for compartment boundaries (Fig. **??**). A notable exception is the short interspersed element (SINE) repeat class which appears to be enriched at TAD boundaries in each cell type. Testing the significance of this observed peak confirms this to be the case, with SINEs significantly enriched at TAD boundaries in each cell type, and borderline significant enrichments can also be observed at compartment boundaries (Fig. **??**).

We also find long interspersed elements (LINEs) are significantly depleted over TAD boundaries in two cell types, and borderline significant in the third, though with modest effect sizes in each case (Figs. **??**, **??**). DNA repeats appear to be enriched at both boundary types (Fig. **??**), however these observations do not surpass our significance threshold ($\alpha = 0.05$) after multiple testing correction (Fig. **??**).

Repeat classes can be broken into smaller repeat families. **?**] reported that the Alu repeat family of the SINE repeat class (or SINE B2 in mouse) is enriched over TAD boundaries. Again we can reproduce this finding and extend the analysis to compartment boundaries, where we do not see a significant enrichment for Alu repeat elements (Fig. **??**). Surprisingly almost all other repeat families show no significant departure from their expected levels over TAD or compartment boundaries.

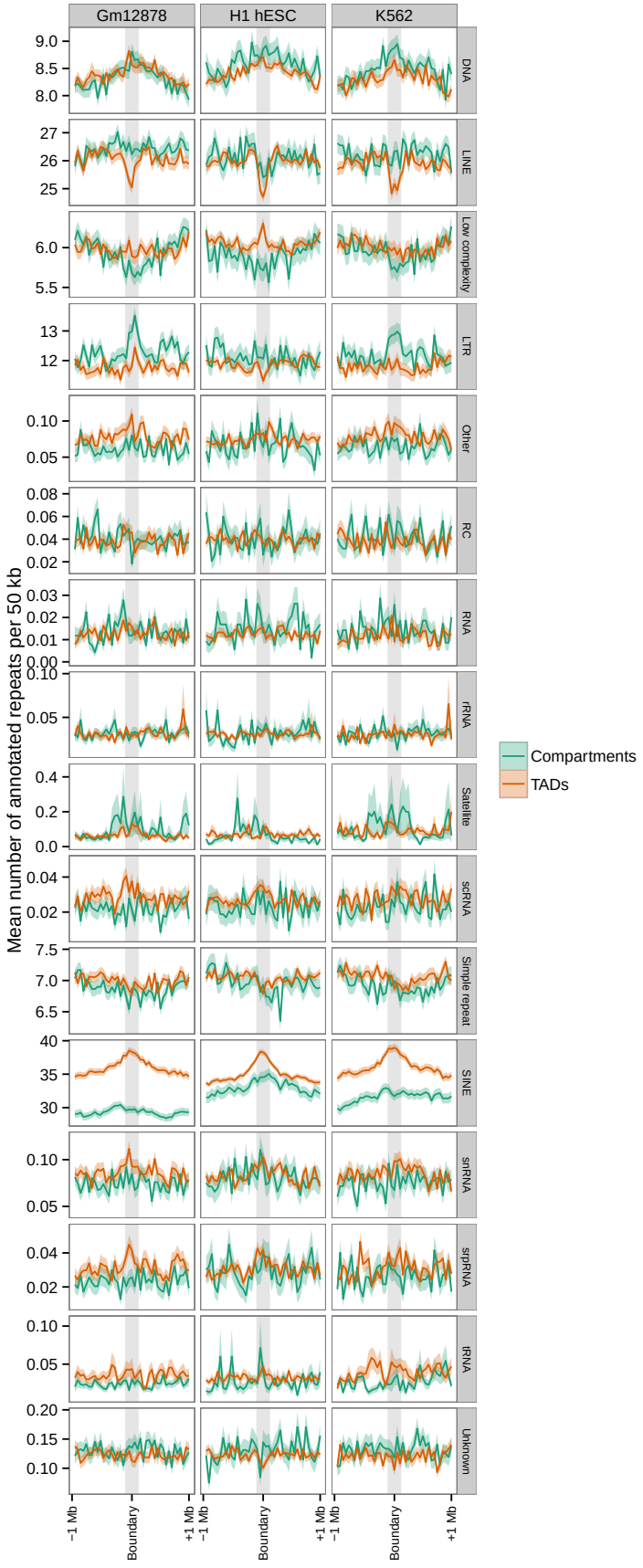

Figure 48: Repeat class average-o-grams over all TAD and compartment boundaries. `RepeatMasker` repeat annotations are counted per 50 kb for 1 Mb either side of all TAD and compartment boundaries. The mean counts per bin genome-wide are plotted with ±95% confidence intervals.

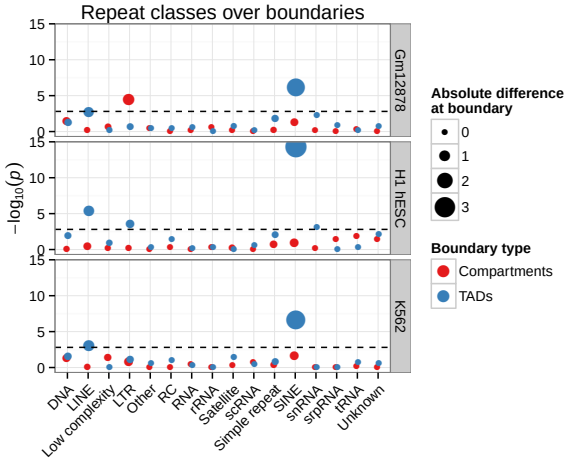

Figure 49: Significance and effect sizes of repeat class enrichments and depletions over boundaries. Boundary profiles (as shown in Figure **??**) were each tested for enrichment or depletion of each repeat class at the boundary bin relative to peripheral non-boundary bins (see Methods **??**). Bubble area is proportional to the raw effect size of an enrichment or depletion. The Bonferroni-corrected significance threshold is highlighted with a dashed line.

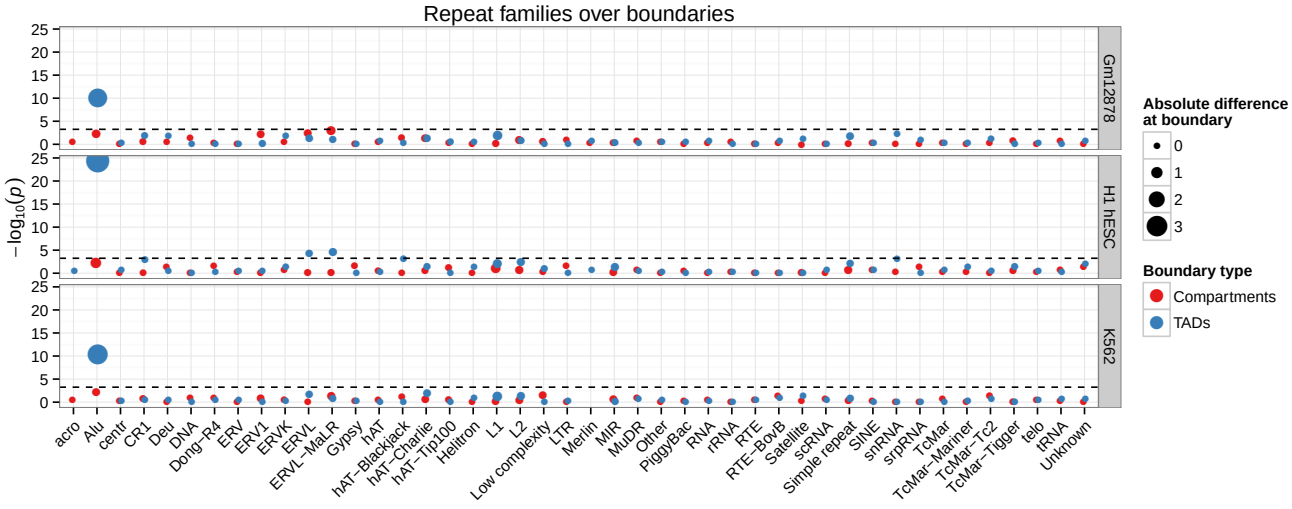

Figure 50: Significance and effect sizes of repeat family enrichments and depletions over boundaries. As per Figure **??** but for a more specific repeat classification.

5.3 TAD BOUNDARY PREDICTION

We have shown TAD and compartment domain boundaries to be reliably and significantly marked by a variety of features. Compartment boundaries are successfully predicted as a side-effect of modelling the continuous compartment profile eigenvector (Section **??**), reflecting global patterns of transcriptional activation and repression, however a related measure of activity and repression, or other comparable profile we might want to predict, does not exist for TADs. Instead then, in this section we attempt to predict TAD boundaries within a class-balanced classification framework.

5.3.1 Learning boundary classification

Results presented in this thesis describe the spectra of chromatin marks over TAD boundaries (Figs. **??**, **??**), thus it is of interested to test if we can build a predictive model (in the manner of Chapter **??**) that can call boundary regions from these marks alone. Such a model, if successful, could have broad utility in domain prediction in metazoan organisms where Hi-C data is not available.

A straightforward approach to this modelling task is to build a supervised classifier that learns the associations between two classes of genomic region: those labelled TAD boundaries and those which are not. To this end, we again turn to a Random Forest (RF) model, due to its many attractive properties discussed previously (Methods **??**). Our input feature set is made up of the same 35 matched features used in models of compartment eigenvector (Section **??**), with the addition of Alu repeat element counts (Section **??**). Domain calls are those produced by the **?**] TAD calling algorithm (Methods **??**), therefore TAD boundaries were resolved to 40 kb bins. We class TAD boundary bins as boundary true positives (TP), and select matched bins 500 kb upstream as boundary true negatives (TN) for our training set.

To build parsimonious and accurate models (as discussed in Section **??**), we used the `AUC-RF` algorithm.[?] This is a form of stepwise model selection which optimises feature subset selection relative to the area under the receiver operating characteristic (AUROC), a metric which captures both the specificity and sensitivity of a classifier. The AUC-RF algorithm was applied to a training set of 80% of boundaries per cell type, with predictions assessed on out-of-bag (OOB) data as each forest was constructed. Selected models were then applied to the remaining held-out test set of 20% of TAD boundaries, with their matched non-boundary bins (full details are given in Methods **??**).

Predictive performance of these models is shown as ROC plots (Fig. **??**) and in each case an AUROC of around 0.67–0.71 was achieved. In practice, this means that each classifier has around a 70% probability of ranking a random boundary region more highly than a random non-boundary region.[?] According to a commonly-

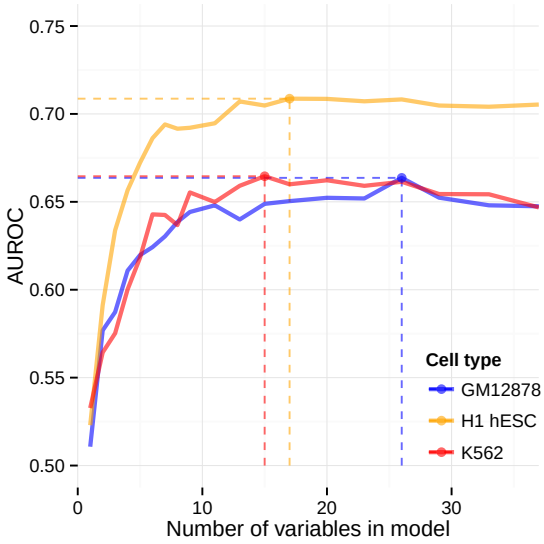

Figure 51: The AUC-RF stepwise algorithm finds that variable subset which maximises the AUROC on out-of-bag data. The AUROC is calculated at each stage of a stepwise method for selecting the best subset model from a full featureset. Here AUC-RF selected the following combinations: GM12878: 26 vars, 0.66 AUC; H1 hESC: 17 vars, 0.71 AUC; K562: 15 vars, 0.66 AUC. The AUC-RF algorithm is described in Methods **??**.

used AUROC rule of thumb, this performance falls between the ranges of "poor" to "moderate" classification accuracy.[?]

Despite this sub-optimal classification accuracy, it is still of interest to analyse the variable importances in each cell type model. Strikingly we find that CTCF stands out as the most informative variable in each classifier by some margin (Fig. **??**). This is in agreement with our results that CTCF showed among the highest and most consistent enrichments at TAD boundaries (Fig. **??**). RAD21 is also highly ranked in each case and this ties in with our previous results suggesting orthogonal boundary enrichments for either CTCF or RAD21 (Section **??**). Surprisingly YY1 does not feature as highly ranked in any model despite our observed consistent enrichments at TAD boundaries (Fig. **??**). In fact YY1 was pruned from the optimal H1 hESC and GM12878 models by the AUC-RF algorithm due to its low variable importance (*data not shown*). This could be due to a redundancy between information provided by YY1 and the CTCF variable as, for example, co-binding of YY1 and CTCF is thought to occur at sites of long-range chromatin interactions.[?]

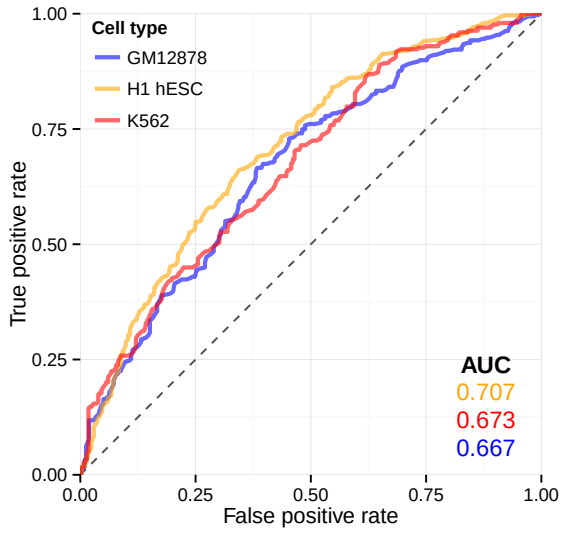

Figure 52: Receiver operating characteristics for TAD boundary classifications. ROCs are shown for three cell type specific classifiers of TAD boundary bins. The area under each cure is also shown (*inset*).

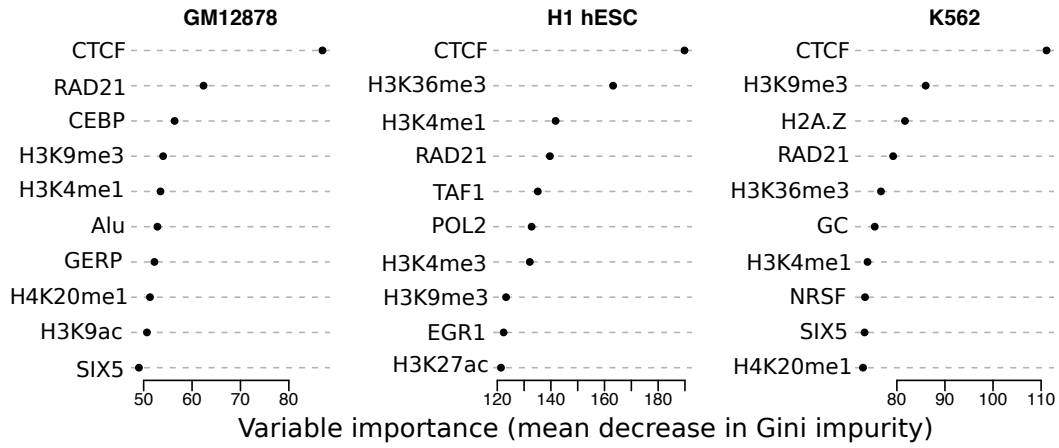

Figure 53: Variable importance for boundary classification Random Forest models. The ten most important variables in each boundary classification model are shown and ranked by their variable importance as defined by mean decrease in Gini impurity (Methods **??**).

5.4 METATAD BOUNDARIES

Our collaborators in the Pombo lab (Max Delbrück Center, Berlin) proposed the concept of "metaTADs": sequential aggregations of adjacent and strongly-interacting TADs that form a hierarchy of domain organisation covering each chromosome.

MetaTADs are constructed simply by performing constrained hierarchical clustering based on inter-domain contacts. That is, those two neighbour TADs that have the largest number of inter-TAD contacts are linked to form a metaTAD and this process is recursed until all TADs on a chromosome are joined into a single tree which fully describes the hierarchical nature of domain organisation (*manuscript under revision*).

My contribution to this work was to explore these newly-described metaTAD structures and perform boundary analysis as was done with TADs and compartments (Section **??**). A testable hypothesis, for example, is that boundaries of larger metaTAD structures could display greater enrichments for boundary-defining features.

5.4.1 MetaTAD boundary comparison

Due to the manner in which metaTADs were constructed by our collaborators, by sequential aggregation of TADs (Methods **??**), boundary comparisons between TADs and metaTADs are not completely straightforward. Every metaTAD boundary is, at a lower level, also a TAD boundary hence for a meaningful comparison we compared only a subset of metaTAD boundaries. Selection of this threshold involved a trade-off between maintaining a sufficient sample size of metaTADs and minimising the overlap between metaTAD and TAD boundaries to increase the discriminative power of our comparison (Fig. **??**).

From the calibration plot (Fig. **??**), a lower bound metaTAD size cut-off of 10 Mb was selected for comparison with TAD boundaries. This left a reasonable sample size of 263 metaTAD boundaries, while reducing the overlap with the set of all TAD boundaries to approximately 5% (Fig. **??**). We also used an upper bound for size selection, based on observations by our collaborators that interactions between metaTADs larger than around 40 Mb were no higher than expected background signal (*data not shown*). In practice, almost all boundaries making up metaTADs larger than 10 Mb are also present in those larger than 40 Mb, but as hierarchical clustering continued up to the whole chromosome level, this upper bound may exclude a small number of edge-case peripheral TADs which aggregated into chromosome-wide metaTADs without evidence of heightened intra-TAD interactions.

Next a comparison between metaTAD boundaries for metaTAD size (s) in the range 10 Mb $< s <$ 40 Mb was performed. Our collaborators generated several ChIP-seq datasets, including for CTCF and three PolII variants, as well as expression data in the form of CAGE (Methods **??**). We calculated the average profiles of each of these

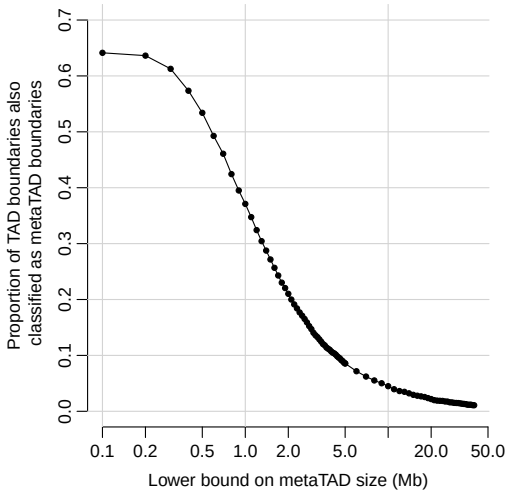

Figure 54: Calibration of metaTAD size selection bounds. As the lower bound on metaTAD size is increased, the proportion of all TAD boundaries which are also metaTAD boundaries decreases.

features over regions surrounding the set of all metaTAD and TAD boundaries (Fig. **??**). These average profiles show heightened enrichment for PolII variants, CTCF and DNase, with non-overlapping 95% confidence intervals of the mean over the boundary bin. Profiles also suggest increased enrichments of gene density and the histone modification H3K27me3 at metaTAD boundaries relative to TAD boundaries (Fig. **??**). The co-incidence of metaTAD boundaries and lamina associated domains (LADs) is explored further in Section **??**.

Increased enrichment at metaTAD boundaries relative to TAD boundaries lends evidence to the functional importance of metaTADs, and suggests boundaries become increasingly well-demarcated at higher levels of organisation. However, if this is a genuine biological phenomenon, we may expect the trend not just to be observable in a comparison between two selected sets, but to increase monotonically as we ascended the metaTAD hierarchy from TADs to chromosomes.

To test this, we reran the metaTAD boundary analysis (Fig. **??**) but at a range of metaTAD size cut-offs. Results of this analysis are shown in Figure **??**. Generally we find increasing enrichments in metaTAD boundaries relative to TAD bounds through the range of lower bound cutoffs from 0 to 20 Mb, possibly with a slight decreased effect size at the highest cutoff of 30 Mb, where the sample size of boundaries decreases to just 62 (Fig. **??**). This analysis strengthens the evidence for heightened functional enrichment of metaTAD boundaries (Figs. **??**, **??**) and suggests the metaTAD aggregation procedure is capturing boundaries of increasing strength, in terms of enrichment of boundary associated features, through the metaTAD hierarchy. Though the nature and significance of these boundary enrichments are currently open to debate, there

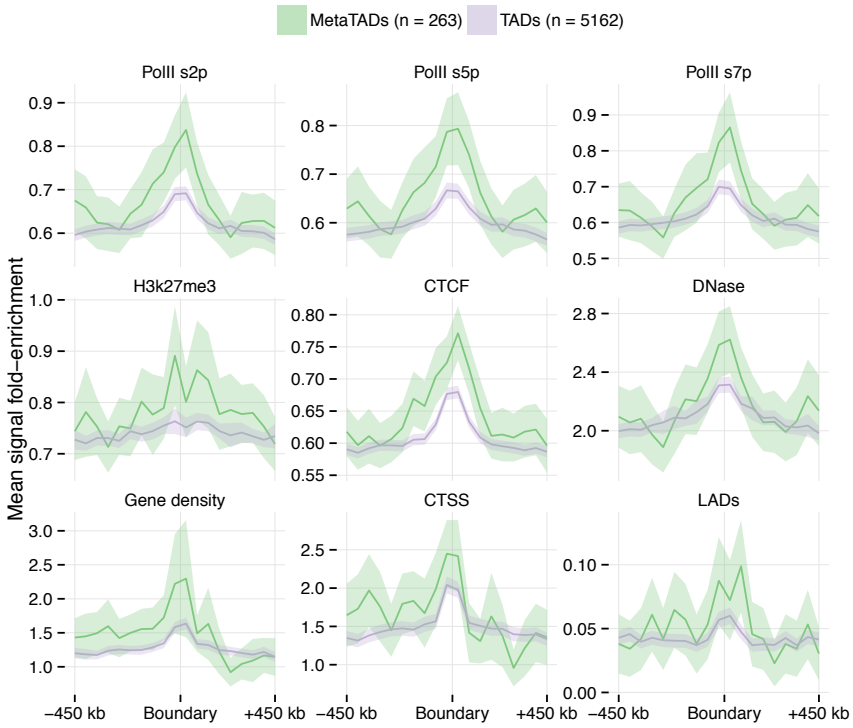

Figure 55: **Large metaTADs show greater enrichments than TADs for an array of boundary features.** Genome-wide profiles of epigenomic features and gene densities averaged over all TAD and metaTAD (10 – 40 Mb) boundaries (ribbons show 95% confidence intervals of the mean).

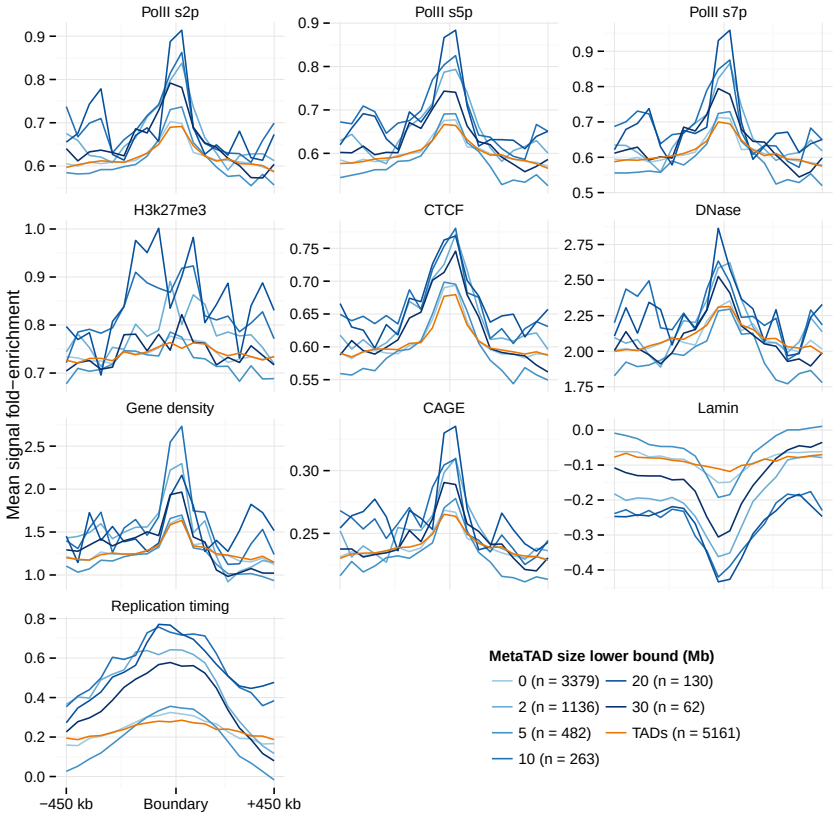

Figure 56: MetaTAD boundary enrichments and depletions with varying size selection. Boundary average-o-grams are shown for TAD and metaTAD boundaries with variable lower bound cut-offs. Sample sizes for each threshold are shown in the legend (*inset*).

exist precedents where greater enrichments over boundaries have been invoked as evidence that a novel TAD calling algorithm outperforms previous efforts.[? ?]

5.4.2 Lamina associated domains

In the previous section we report a colocation of metaTAD boundaries and lamina associated domains (LADs), and at a greater level than that observed with smaller TADs (Fig. **??**). This hints at an association between metaTADs and LADs which merits further investigation.

High resolution LAD data in mouse embryonic stem cells were retrieved from **?**] in the form of continuous measures of lamin-B1 association produced by the DamID technique, known to reflect proximity to the nuclear lamin.[?] This measure of lamina association was then processed in windows around each metaTAD and TAD boundary, and profiles were combined to form a heatmap (Fig. **??**).

Boundaries were seriated in order to separate out those that coincide with a LAD boundary, indicated by a transition in lamina association values. This was achieved

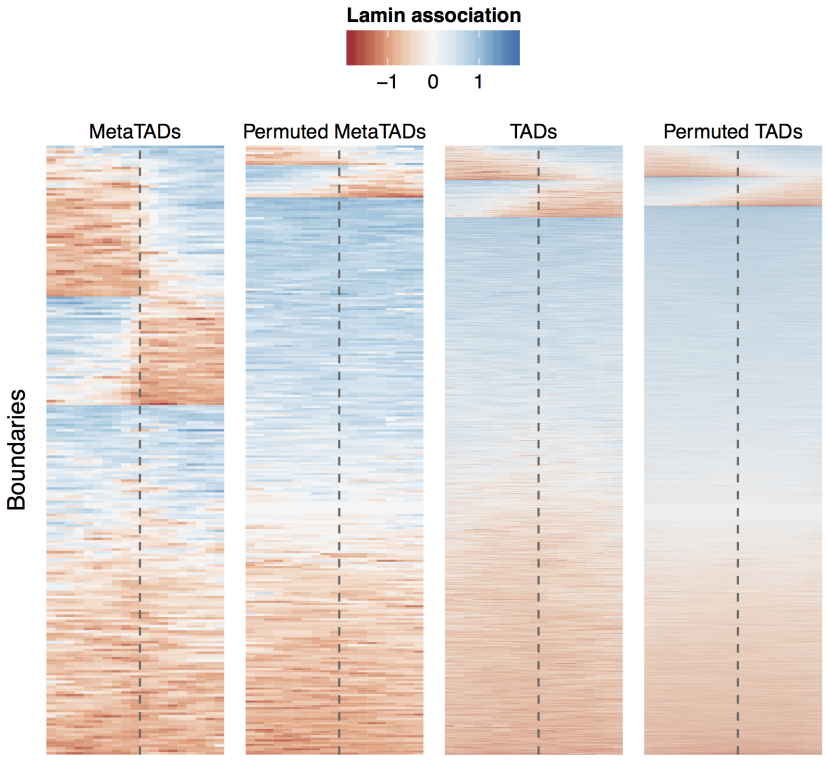

Figure 57: MetaTAD boundaries align with those of lamina associated domains. Heatmaps of LaminB1 association microarray probe intensity values over MetaTAD boundaries (from domains of size 10 – 40 Mb) and TAD boundaries, are displayed beside examples of circularly-permuted boundaries (Methods **??**). Profiles are shown ±450 kb from each boundary.

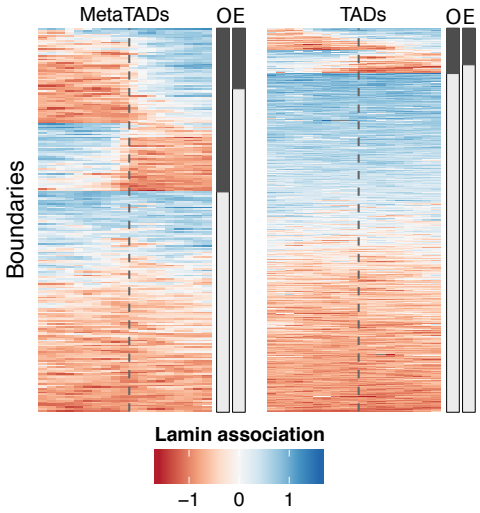

Figure 58: MetaTAD boundaries co-occur with LAD boundaries more often than is expected by chance. Heatmaps of lamina association over MetaTAD boundaries are shown as in Figure **??**. Sidebars labelled O and E reflect observed and expected proportions of metaTAD / LAD boundary overlaps (Methods **??**).

by fitting a linear regression model across the lamina association profile of each boundary, then using a coefficient cutoff heuristic to select those that represent a boundary transition (Methods **??**). Using this approach, we found a markedly large proportion of metaTAD boundaries (43%) co-occur at a LAD boundary (Fig. **??**). The same comparison using TAD boundaries found a coincidence of just 12%. However LADs are large domains and there were over 5,000 TADs called using these Hi-C data by our collaborators, thus in absolute numbers of boundaries this is still represents a large overlap.

To test the significance of these observations, we apply a permutation-based statistical test where our observed coincidences are compared with those produced by 1,000 circular (per-chromosome) permutations (Methods **??**). We found that the metaTAD and LAD boundary coincidence is around a 2.7-fold increase above null expectation (observed: 42.6%; expected: 15.8%; empirical p-value: $p < 1 \times 10^{-4}$; Fig. **??**). Meanwhile TAD boundaries were found to have a smaller, yet still significant, 1.2-fold increase in coincidence with LAD boundaries relative to a null model (observed: 11.8%; expected: 9.5%; empirical p-value: $p < 1 \times 10^{-4}$; Fig. **??**).

As with other enrichments (e.g. Fig. **??**), we went on to verify that this result was not specific to the choice of boundary size cutoff. Recall that as metaTADs are derived from TADs, selecting metaTADs within a size range is a trade-off between minimising overlap between boundaries assigned to TADs and metaTADs to enable more powerful comparison, while retaining a sufficiently large sample size of metaTAD boundaries (Section **??**). In the case of LAD–metaTAD coincidence, we again find this result is insensitive to the selection of domain size cutoff, and indeed that there is evidence for

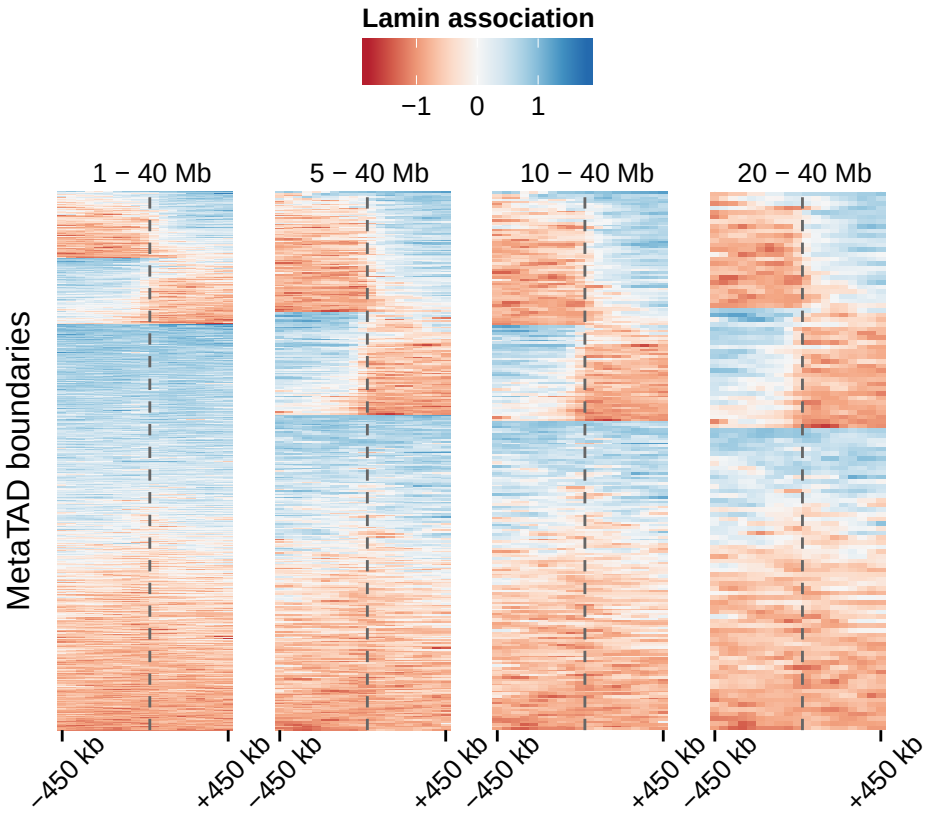

Figure 59: MetaTAD–LAD coincidence increases at higher levels of the metaTAD hierarchy. Heatmaps of lamina association over MetaTAD boundaries are shown as in Figure **??**. Panels show the size selection cutoffs for metaTAD domains considered in each case.

an increasing proportion of co-occurrence as we ascend the metaTAD hierarchy (Fig. **??**).

This result suggests again that metaTADs seems to offer useful perspectives onto higher order genome organisation. In this case, it appears TADs will often neatly aggregate within LADs and together these constitute what we observe as a metaTAD.

5.4.3 Boundaries over a time series

For the first time, our collaborator's applied the Hi-C technique over a differentiation time course from mouse embryonic stem cells, to neural progenitors and finally fully-differentiated neuron cells. Successive expression measures were also taken alongside this Hi-C data in the form of CAGE data, produced by the FANTOM5 consortium.[?] Together these datasets offer a unique perspective onto how higher order genome organisation varies with expression during differentiation.

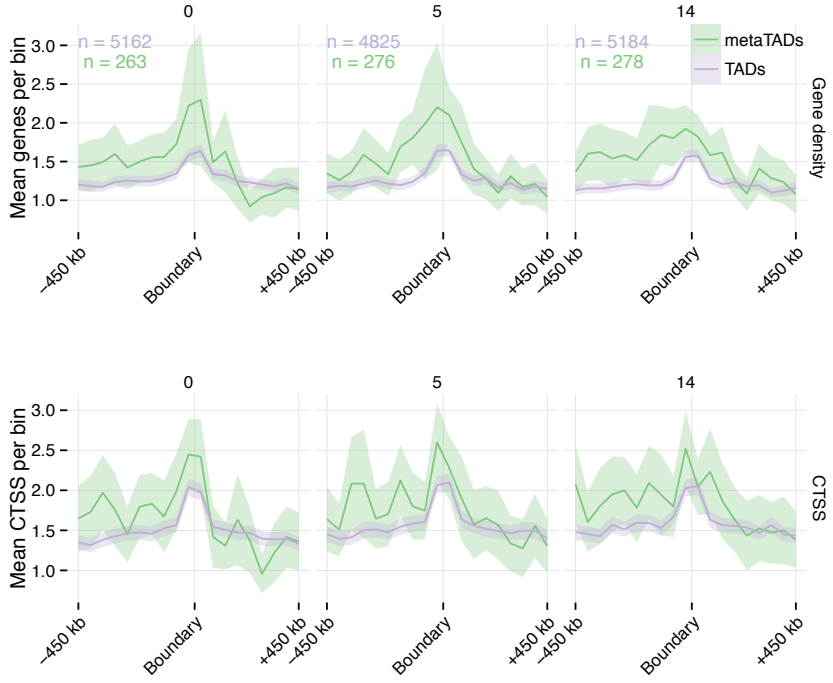

Figure 60: Observed enrichments persist over a time series. CAGE-defined active TSS (CTSS) were counted per 50 kb bin across each TAD and MetaTAD (10 – 40 Mb) boundary and averaged (ribbons show 95% confidence intervals of the mean). Gene densities refer to mean counts of annotated genes per bin, with an overlap of at least 250 bp (Methods **??**).

Collaborators explored changes in both the overall tree structure between timepoints, and aggregate expression changes between TADs and metaTADs at successive timepoints. They identified rewiring events in the metaTAD tree over the timecourse, and found corresponding changes in expression (*data not shown*). Given metaTAD structures appear to differ and matched CAGE data exists for each timepoint, it was of interest to test how observed boundary enrichments for gene expression (Figs. **??**, **??**) might vary over this timecourse.

We find actively-transcribed CAGE-defined TSS (CTSS) to be consistently enriched over the shifting boundaries through the differentiation timecourse (Fig. **??**). We coupled this with a static measure of gene density in order to distinguish expression changes from genic overlap, however both series show similar patterns so it does not seem that boundary expression varies at a global scale over this timecourse. At each timepoint, peak heights over boundary bins suggest modestly stronger enrichments at metaTAD boundaries relative to TAD boundaries, as seen with other features (Fig. **??**).

5.5 OTHER BOUNDARIES

5.5.1 Giemsa bands

A recent analysis of Hi-C datasets examined the hierarchy of nuclear compartment and TAD organisation in human HeLa cells across the cell cycle. They found that interphase and metaphase chromatin structure are highly distinct, such that the TADs and compartments observed here (e.g. Fig. **??**) are effectively abolished in metaphase.[?] This raises the question of how the structural organisation seen in (and often shared between) interphase cells is inherited through the cell cycle.

Human Giemsa metaphase banding (G-band) pattern data have been integrated with the human genome assembly, and although such data are widely used, they are also necessarily of low resolution.[?] These G-band patterns are constant over human cell types at metaphase, but all traces of interphase higher order structure were reported to be absent at metaphase.[?] We would therefore not necessarily expect an agreement between G-bands, labelled in metaphase cells, and nuclear compartments, called from cell populations which were mostly in interphase.

We examined the genome wide concordance of interphase compartment boundaries with metaphase G-band boundaries, relative to an expected distribution derived by permutation (Methods **??**). We found a significant, though modest, excess of compartment boundaries within close proximity of G-band boundaries, such that 13.9% of compartment boundaries are within 500 kb of a G-band boundary (expectation = 10.5%, K-S test: $D = 0.076$, $p < 3 \times 10^{-12}$). This is seen for compartment boundaries calculated for all three cell types independently (a full comparison for GM12878 is shown in Figure **??**).

The genome wide overlap of compartment A and B regions with particular G-band classes is nonrandom, and suggests a greater correspondence than that of simple boundary distances. Regions assigned to compartment A are significantly over-represented within lighter staining (especially G-negative) bands, while compartment B regions are over-represented in the most darkly staining (G-positive) bands (Fig. **??**). Approximately 40% of the genome jointly occupies interphase compartment A as well as the lightly-stained `gneg` or `gpos25` metaphase G-bands, or occupies the interphase B compartment in addition to the well-stained `gpos75` or `gpos100` bands at metaphase (Fig. **??**). Similar trends are seen across all three cell types as expected given the high correlations seen in A/B compartment profiles between cell types (*data not shown*).

This agreement is not wholly unexpected given the known characteristics of G-negative and G-positive bands, with contrasting gene density, GC content and replication timing[?] that is strongly reminiscent of the contrasts between interphase A and B compartments.[?] Despite showing an association, these data agree with experimental evidence that many domain boundaries are not well preserved between interphase and metaphase. However there is evidence for relationships between the broad structural

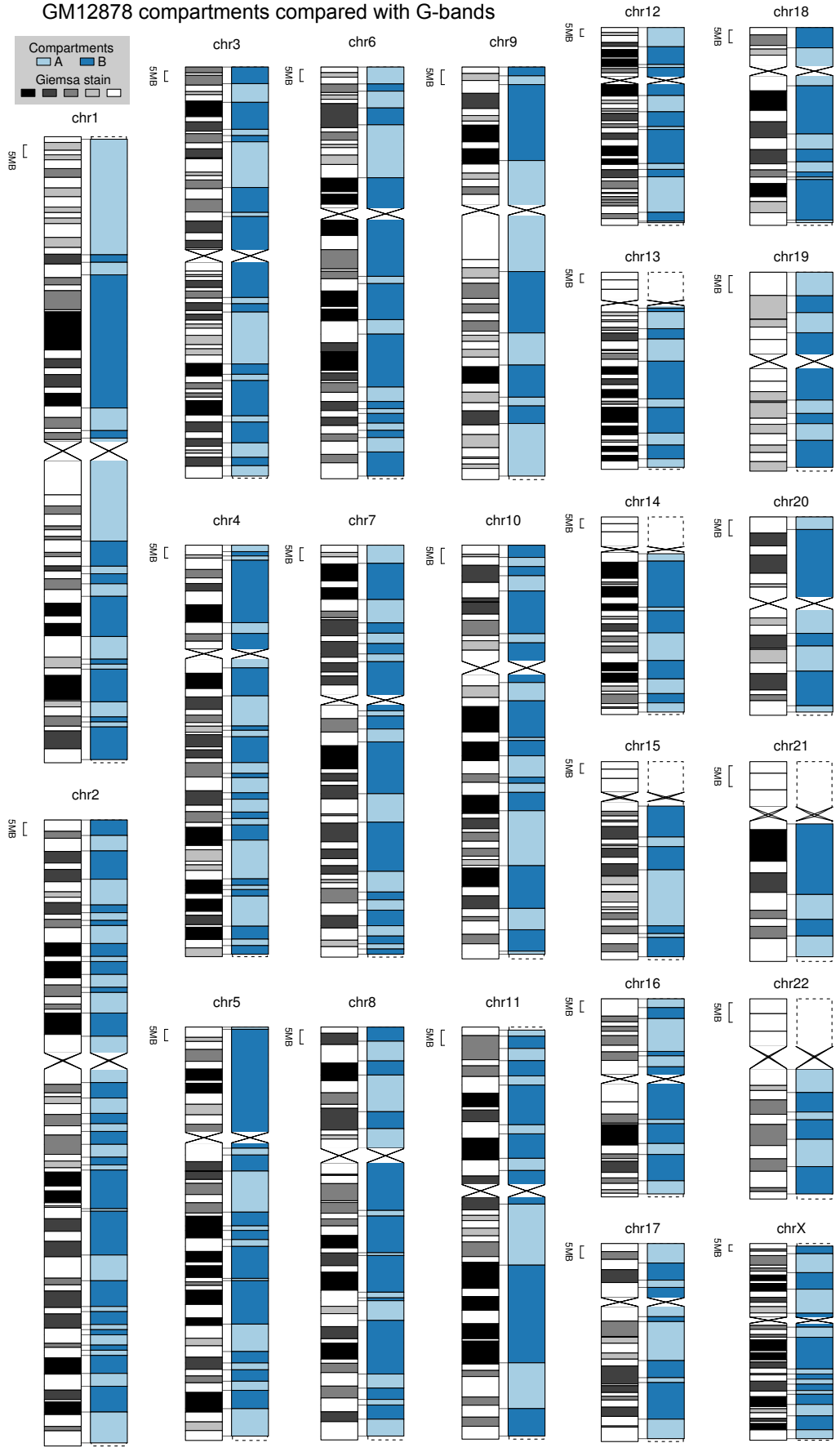

Figure 61: Genome-wide agreement between Giemsa bands and A/B compartments. G-bands, assayed at metaphase, often correspond with interphase A/B compartments across all chromosomes. Data for cell type GM12878 is shown.

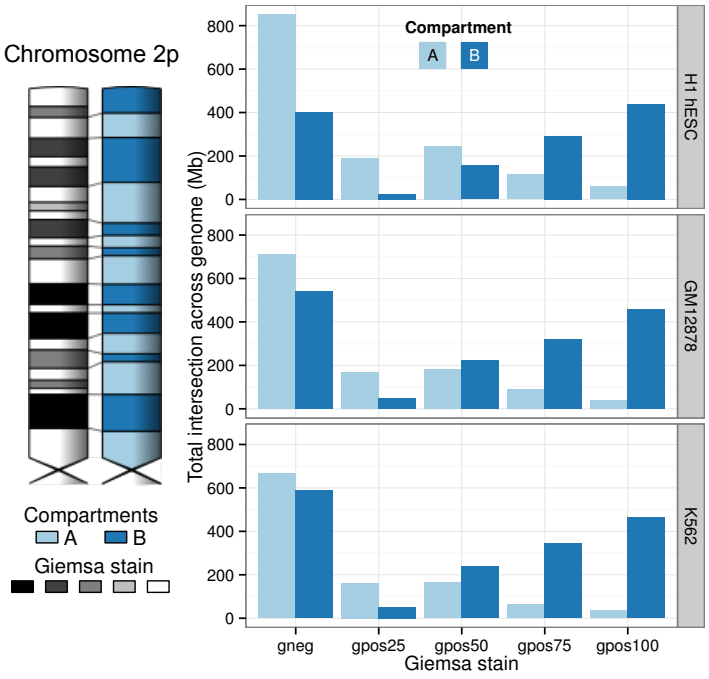

Figure 62: Giemsa–stain bands correspond to A/B compartments. The correspondence between G-bands and A/B compartments is broken down into the five levels of Giemsa stain. A compartments largely match *gneg* staining, while *gpos75* and *gpos100* are enriched in B compartments.

categories of compartments and G-bands which may reflect similarities in the degree of compaction throughout the cell cycle.

6 LOCAL CHROMATIN CONFORMATION

6.1 INTRODUCTION

The Hi-C assay provides a genome-wide overview of chromatin conformation, however the broad scope of this all-versus-all assay places inherent limits on the resolution at which individual interactions can be analysed. For a closer look at chromatin conformation within a region of interest, alternative C-based assays such as 3C, 4C and 5C can be employed alongside classical microscopy techniques like FISH.

Here I discuss a collaborative project within the IGMM (with members of Prof. Robert Hill's laboratory) involving the use of 4C and 5C data to zoom in on a particularly well-studied locus involved in limb development: the Sonic hedgehog (*Shh*) gene and its distal *cis*-regulatory element named ZRS.

6.2 THE SHH LOCUS

Anterior-posterior patterning in the developing limb is regulated in mammals by the Sonic hedgehog morphogen, encoded by the *Shh* gene.[?] Specifically, the *Shh* gene is expressed within a confined region of developing limb buds named the "zone of polarising activity" (ZPA). Its expression within this region is known to be regulated by a well-studied enhancer, the "ZPA regulatory sequence" or ZRS.[?] ZRS is located almost 1 Mb downstream of its target *Shh* promoter in humans (nearer 800 kb in mouse), and is located in intronic regions of another gene, LMBR1 (Fig. **??**).[? ?] Expression of *Shh* within the ZPA is tightly controlled, initiating in mice at developmental stage E9.5 and terminating at E12.5.[?] As such, single point mutations and short insertions within the ZRS enhancer have been linked to various limb deformities, including pre- and post-axial polydactyly.[? ? ?] For example, a heritable point mutation in the ZRS enhancer is the cause of polydactyly in "Hemingway cats", a large group of domestic cats with extra toes that reside at the former home of Ernest Hemingway.[? ?]

To further investigate the dynamics of the *Shh* locus, our collaborators in the Hill lab (IGMM, University of Edinburgh) have developed a model system which allows inducible *Shh* expression in a non-expressing 14fp cell line derived from the developing murine limb bud. Treatment of this cell line with the histone deacetylase inhibitor trichostatin A (TSA) then leads to detectable *Shh* expression, and increased levels

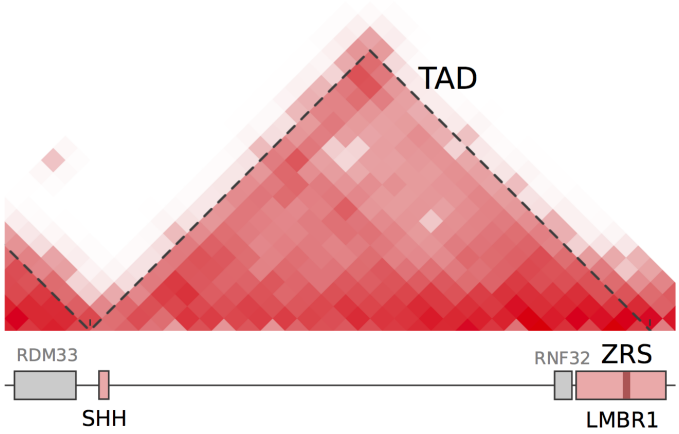

Figure 63: ZRS–*Shh* contacts occur within a stable TAD. An approximately 1 Mb region of the mouse genome is shown below a matched section of a Hi-C contact map (derived from previously published data[?]). A clear TAD can be identified spanning from the *Shh* gene to the ZRS, dashed lines show TAD boundaries called by **?**]

of the histone activation mark H3K27ac at the ZRS (*unpublished data*). However, the question remains whether this TSA treatment is fundamentally altering local chromatin structure—that is, bringing together the ZRS enhancer with its target *Shh* promoter—or whether ZRS and *Shh* are in contact in both the active and non-expressing cell lines in a poised state. Previous 3D-FISH experiments have shown the ZRS–*Shh* contact to be associated with *Shh* expression in the developing limb bud, though it is not detected in every cell.[? ?] Instead only a proportion of cells in the ZPA are *Shh*-expressing at a given time, and it is thought that the ZRS–*Shh* colocalisation is most likely to occur within these expressing cells.[?]

My part in this collaboration was to analyse 3C-seq (also known as 4C), then 5C data generated by our collaborators over the *Shh*–ZRS region in mouse. Experimental design and wet-lab procedures were performed by members of the Hill lab.

6.3 4C OF THE ZRS

4C experiments were performed by collaborators using the ZRS region as a bait sequence, or "viewpoint", such that ZRS contacts were measured with all other HindIII restriction fragments genome-wide. Thus the 4C technique allows us to assay changes in the ZRS–*Shh* contact relative to the totality of other chromatin interactions involving ZRS.

The 4C experimental design involved two control experiments. The first used cells derived from whole limb bud at developmental stage E11.5, thereby containing some *Shh* expressing cells as a positive control. The negative control was a mouse mandibular cell line (MD) which does not express *Shh*. Expression status in each case

was experimentally verified (*data not shown*). For the perturbation experiment, 4C was performed in the *Shh*-inducible 14fp cell line, both with and without trichostatin A (TSA) treatment.

6.3.1 4C pipeline

The 4C analysis pipeline, starting from de-multiplexed sequencing reads (`fastq` files) as produced by our in-house sequencing facilities using an Ion Torrent Ion Proton™ sequencer, can be summarised as:

1. Trim known bait sequence using `cutadapt`,[?] select only those reads where known viewpoint-associated sequence was present
2. Map reads to the mouse reference genome (build `mm9`) using `bowtie2`[?] with the `very-sensitive` flag
3. Filter alignments with a MAPQ score > 30 to select for high-confidence alignments using `samtools`[?]
4. Normalise contacts using the `r3cseq` R package and assign FDR q-values to interactions, with the aim of finding those significantly over-represented relative to expectation (Methods **??**)

6.3.2 ZRS–*Shh* interaction following TSA treatment

The results of a comparison between 4C experiments in TSA treated and untreated 14fp cells is shown in Figure **??**. In it we see a striking and highly significant ZRS–SHH contact in the treated sample (q-value $< 5 \times 10^{-10}$), with a weaker but still significant contact in the adjacent restriction fragment in the untreated sample (q-value $< 5 \times 10^{-5}$).

Comparing these results with controls shows a detectable and significant ZRS–*Shh* contact in each case, regardless of *Shh* expression status (FDR q-value < 0.05); Fig. **??**). This is in agreement with previous evidence suggesting ZRS contacts *Shh* constitutively.[?] However the TSA– 14fp cell line also shows a large number of off-target contacts, potentially indicating a lack of specificity in the ZRS–*Shh* contact, or a range of alternative contacts occurring throughout the cell population.

Unpublished experimental results show that following TSA treatment, *Shh* expression increases over a period of 24 hours until it reaches that seen in the limb, this steady increase is also mirrored by an increase in the level of H3K27ac histone mark over ZRS (Hill lab, *personal communication*). For this reason, 4C of ZRS was also performed a full 24 hours after TSA treatment, as well as the 18 hour treatment analysed above (e.g. Fig. **??**). These two experiments give largely similar results (Fig. **??**), and

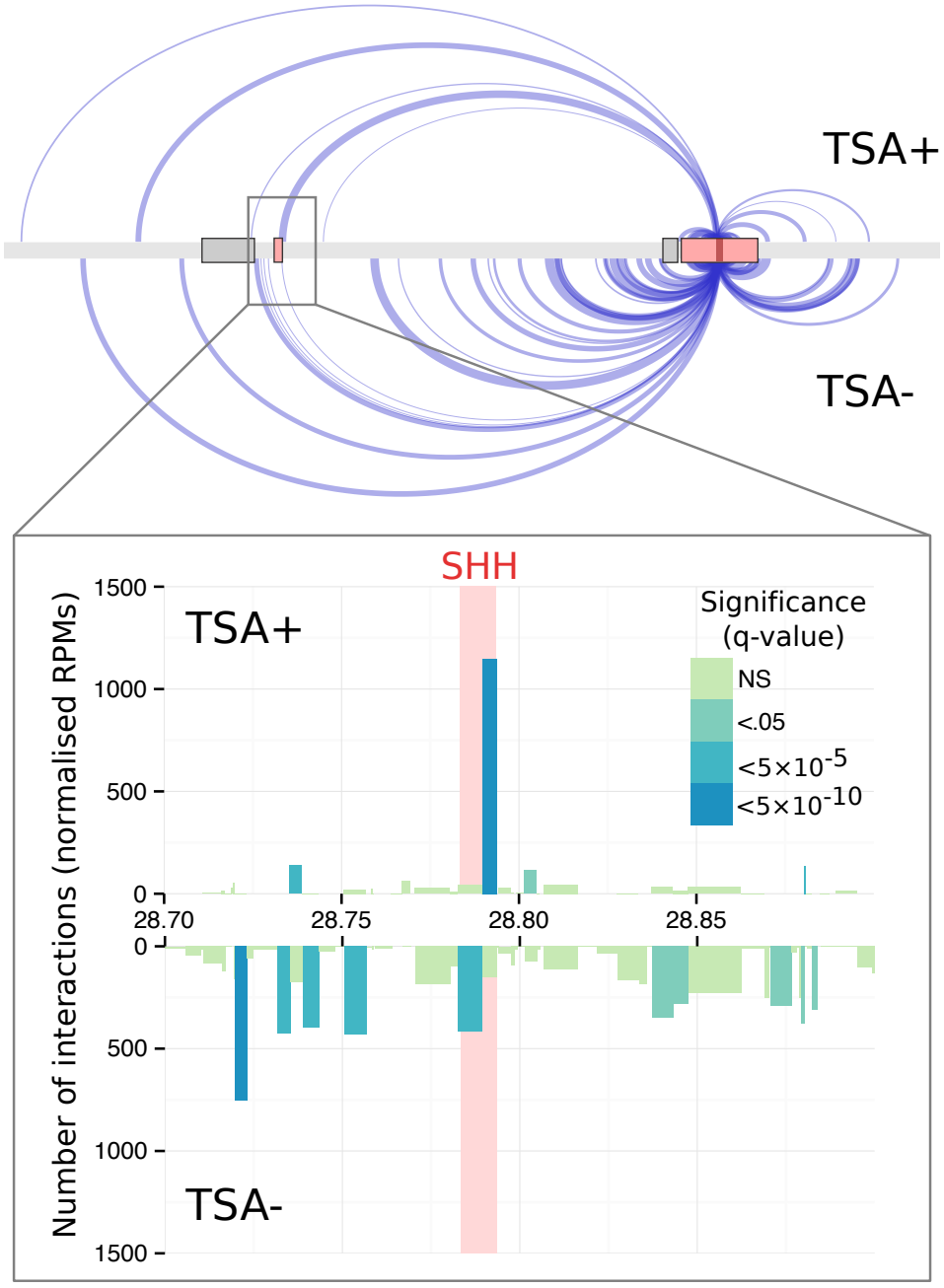

Figure 64: TSA treatment induces a strong ZRS–*Shh* interaction. 4C interactions are shown as edges from source node (ZRS enhancer bait fragment) to targets along an approximately 2 Mb region of chromosome 5. Edge width is proportional to the number of interactions, only highly significant interactions are shown (FDR q-value $< 5 \times 10^{-5}$; Methods **??**). Zoomed region shows the number of interactions of the bait region with *Shh* in both untreated and TSA treated (after 24h) samples. Each green–blue rectangle is a restriction fragment, coloured by FDR q-value indicating the significance of the interaction above expected levels.

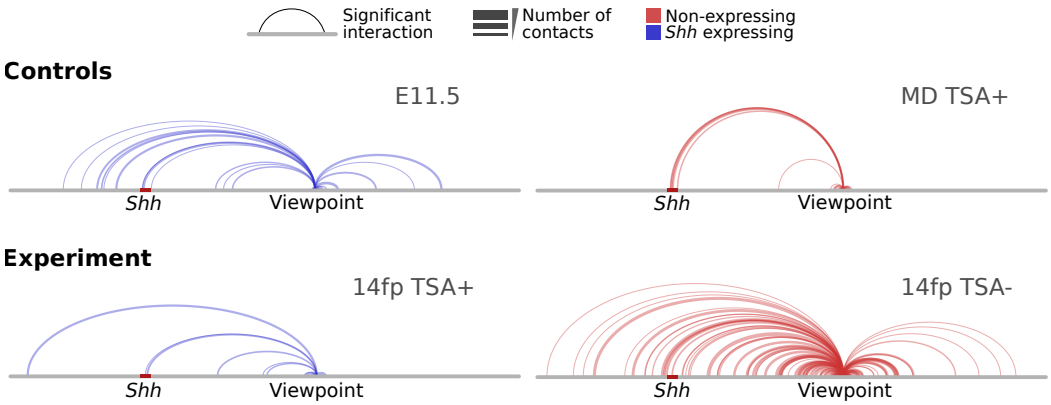

Figure 65: TSA treatment in 14fp cells results in a more specific ZRS–*Shh* contact. Arc plots are shown for two control experiments: the *Shh* expressing E11.5 whole limb bud and non-expressing mandibular cell line (MD). Also shown is the 14fp inducible cell line, with and without treatment with trichostatin A (TSA) after 18 hours. Arcs link significant interactions (q-value < 0.05) and arc widths are proportional to the normalised number of reads recorded for the interaction (Methods **??**).

the ZRS–*Shh* interaction frequency is highly significant in each case, particularly 24 hours after treatment ($TSA\text{-} : q < 1.5 \times 10^{-5}$; $TSA{+}^{18h} : q < 1.4 \times 10^{-8}$; $TSA{+}^{24h} : q < 7.8 \times 10^{-35}$).

Additional FISH data produced by our collaborators shows approximately equal levels of compaction in this region in both TSA treated and untreated 14fp cells (*data not shown*). This information in combination with the 4C results reported here (Fig. **??**) support a hypothesis that as these two loci border a TAD (Fig. **??**), they frequently contact each other regardless of *Shh* expression state. It could also be the case that TSA treatment brings about a more specific, functional ZRS–*Shh* contact in 14fp cells which is coupled with expression of the *Shh* gene.

6.3.3 Assay diagnostics

The 4C protocol used by our collaborators in this work was that of **?**] In it, the authors advise some statistical tests to ensure the quality of the experiment results. Among these were:[**?**]

1. Sequencing reads should be found to have high duplication rates of 95% or greater.
2. 50% or more of all reads should map to the chromosome on which the bait region is located.

Additionally, the 4C procedure was adapted for specific in-house sequencing instruments (an Ion Torrent Ion Proton™ sequencer as opposed to Illumina™ technology) and as such required diagnostics to confirm the experimental data was accurate.

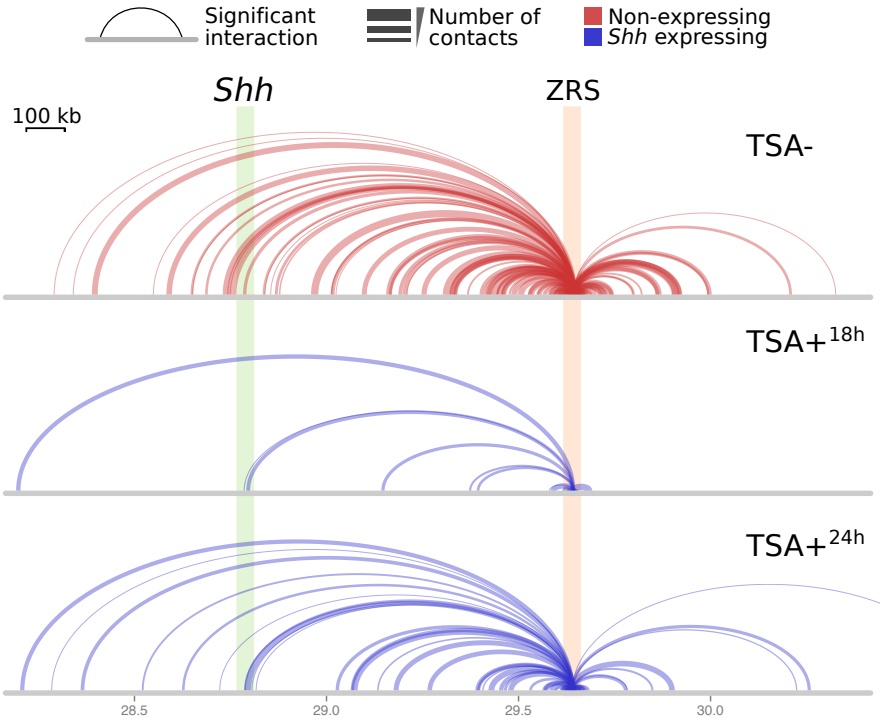

Figure 66: A stable ZRS–*Shh* interaction is coupled with reduced extraneous contacts. Arc plots are shown for an untreated, non-expressing 14fp cell population (*TSA-*) and following TSA treatment after 18 and 24 hours. Arcs link significant interactions (q-value < 0.05) and arc widths are proportional to the normalised number of reads recorded for the interaction (Methods **??**).

Table 6: 4C sequencing library statistics. 4C experiments are summarised as total number of reads in each experiment and the percentage of those reads labelled "duplicates". Note in 4C these duplicates are not artifactual and instead result from large numbers of contacts nearby to the viewpoint.

	14fp TSA-	14fp TSA+ 18h	14fp TSA+ 24h	E11.5	MD TSA+
Reads (million)	10.0	8.8	24.2	10.7	12.2
Duplicated (%)	62.8	74.2	84.4	80.2	72.8

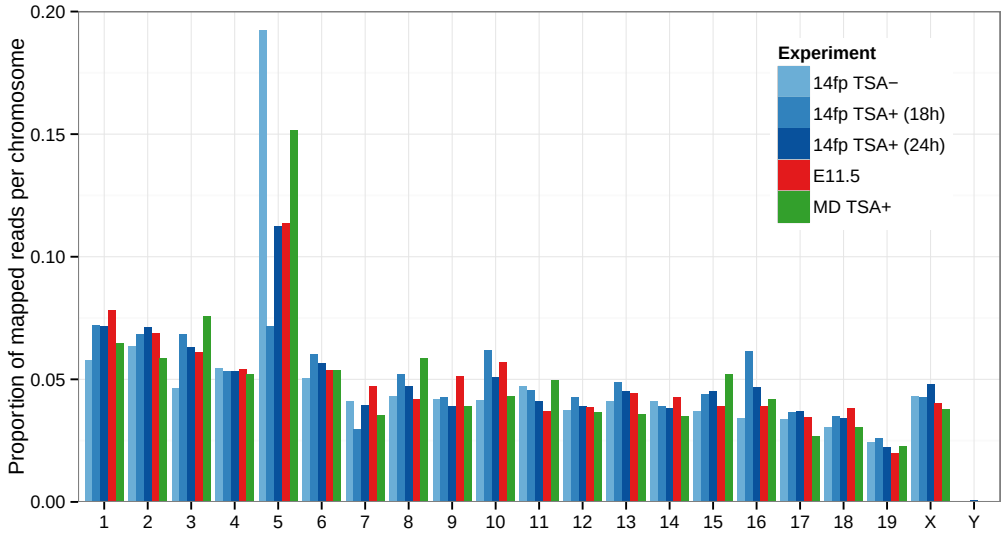

Figure 67: The bait chromosome is enriched for 4C sequencing reads. Chromosome 5 is visibly enriched for 4C reads as it contains the ZRS bait region (or viewpoint). Experiments include treated and untreated 14fp cells (TSA-, TSA+^{18h}, TSA+^{24h}) as well as positive control (E11.5) and negative control (MD TSA+).

Sequence duplication levels were measured with `FastQC`[?] and are shown in Table **??**. We find slightly lower than expected levels of duplication, ranging from 62.8% to 84.4%. This suggests that while the assay does appear to be working, there may be extraneous noise and non-bait interactions in the sequencing library.

Unfortunately we found the proportion of reads mapped to the bait region chromosome, chromosome 5 in this case, fell below the prescribed level of 50%. Across 4C experiments, we find instead that between approximately 10–20% of all reads mapped to the bait chromosome (Fig. **??**), except for the 18 hour TSA+ treatment experiment which shows only around 7%. While this is still a clear enrichment over non-bait chromosomes, relative to their lengths, it suggests the assay results suffer from either increased *trans*-contact noise or decreased *cis*-contact enrichment around the bait region.

Lower than expected levels of both sequence duplication and bait chromosome enrichment suggest loss of signal around the bait region itself. This is the area where we expect both very high levels of duplication (identical restriction fragment pairings

between nearby genomic locations) and where a majority of all sequencing reads should originate, driving the overall chromosome enrichment seen in Figure **??**. The precise reason for the discrepancy is unclear but suggests the results of the assays performed by collaborators may have a lower signal-to-noise ratio than has previously been achievable in 4C experiments.[?] Potentially the signal-to-noise ratio could be improved by utilising a double cross-linking procedure such as that used in **?**]

6.4 POLYMER MODELLING

Chromosome conformation capture assays permit the exploration of genome organisation, but such data are commonly analysed using one or two-dimensional representations. A growing set of algorithms looks instead to rebuild the three-dimensional trajectory of a chromatin fibre, using Hi-C or 5C data as input (e.g. **? ? ? ? ? ? ? ? ?**). Intuitively, in each method the interaction frequency between two regions is idealised as inversely proportional to their physical distance (and adjusted according to various other constraints). Where these methods differ is in their approaches to performing this spatial transformation, and in solving the subsequent optimisation problem. We chose the `AutoChrom3D` method[?] for use in this work (described in Section **??**) as the algorithm can accept 5C input and model polymers at high resolutions of up to 8 kb.

6.4.1 `AutoChrom3D` method

The procedure implemented in `AutoChrom3D` can be summarised as:[?]

1. The chromatin fibre is represented as beads-on-a-string, with $N_{beads} = \lceil \frac{L}{R} \rceil$ (where L is the length of the region and R the resolution)
2. A local compaction parameter is calculated using a sliding window of each 50 adjacent beads (intra-window contacts are averaged and compared to those over the whole region under study)
3. Interaction frequency between beads of a given genomic distance is modelled as a Poisson-distributed random variable and noisy or unstable contacts, considered in the context of neighbouring beads, are filtered
4. This filtered set of interaction frequencies are then normalised using the previously-calculated compaction parameter to give an $N_{beads} \times N_{beads}$ matrix of interaction strengths
5. Interaction strength is converted to spatial distance through two linear transformations based on experimental observations of nuclear occupancy and regional flexibility[?]

Table 7: Measurement distances between ZRS and *Shh* in each inferred 3D structure. Distances are given in arbitrary units. *Shh* spans two beads of the polymer model, hence two distances are calculated in each case (d_1, d_2). RMSD is the minimised root mean squared deviation between the two structures and is given as a relative unitless quantity. The radius of gyration (gyradius) is also shown.

		Distance		RMSD	**Gyradius** (μm)	
		TSA-	TSA+		TSA-	TSA+
88fp	d_1	5.4	5.1	1.701	0.244	0.244
	d_2	4.1	3.9			
MD	d_1	6.2	3.3	2.377	0.217	0.205
	d_2	4.8	2.0			

6. Cartesian co-ordinates are then calculated via non-linear constrained optimisation of pairwise spatial distances using LINGO[?]

6.4.2 Modelling the *Shh* region with 5C

5C data was generated by our collaborators over the same ZRS–*Shh* region as was assayed with 4C (Fig. **??**; Section **??**) with the aim of developing a multi-point perspective on local chromatin conformation beyond that available from 4C data.

We used this 5C experimental data in combination with the `AutoChrom3D` three-dimensional inference algorithm[?] in an attempt to compare polymer trajectories in TSA treated and untreated 88fp mouse cells, a similar and complimentary cell line to that used in earlier 4C experiments (14fp). As a control, 5C was also performed on mandibular (MD) cells, with and without TSA treatment, which do not express *Shh*. Prior to structural modelling, the `my5C` program was used to generate normalised 5C interaction frequencies.[?]

We find that TSA treatment of 88fp cells does appear to slightly reduce the distance between *Shh* and ZRS in inferred 3D structures (Fig. **??**), however this difference is overshadowed—to our surprise—by that observed in the non-expressing MD cell line. This latter mandibular cell line undergoes a large structural transition which brings the *Shh* gene and ZRS into close proximity. Measurements between these elements for each structure are shown in Table **??**.

We also report a greater overall structural shift following TSA treatment in the MD cell line, with an RMSD between the two structures of 2.377 arbitrary units, relative to 1.701 between TSA+ and TSA- 88fp cells. The radius of gyration, unchanged in 88fp, is also decreased in the MD cell line following TSA treatment, indicating the region becomes more compact following TSA treatment (Table **??**).

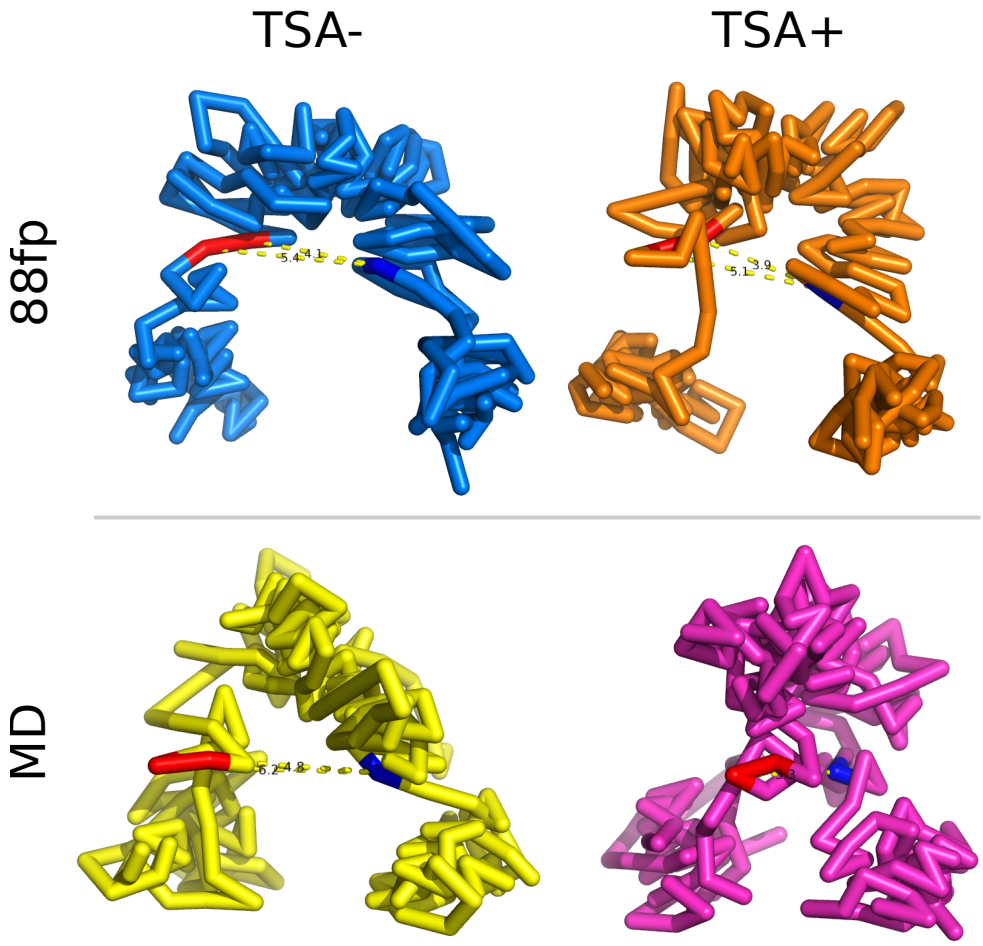

Figure 68: Inferred polymer trajectories of the ZRS–*Shh* region following TSA treatment in two cell lines. 3D structures are shown for 5C experiments assaying the region around *Shh* (*red*) and ZRS (*blue*) in an *Shh*-expressing limb bud cell line (88fp) and a non-expressing mandibular cell line (MD). Labelled measurements are given in Table **??**. Structures were predicted by `AutoChrom3D`[?] using 210 × 8 kb beads per polymer.

6.4.3 Repeat simulations

We have shown what appears to be a structural shift in the ZRS–*Shh* locus by 3D modelling predictions (Section **??**). It is of interest to assess the stability and reproducibility of these results through repeat simulations of the polymer trajectory. At this point it is unclear whether the ZRS–*Shh* bound state represents a firm consensus over the cell population, or an alternative structure with similar optimisation energy to that of the non-contacting state.

We re-ran simulations of the 3D chromatin fibre in the ZRS–*Shh* region a total of five times (Fig. **??**). In each case, the algorithm generates the known ZRS–*Shh* TAD as a compacted domain bookended by the two loci under study. This sanity check shows that the results are broadly compatible with our *a priori* expectation of the region's structure given the 2D heatmap representation of 5C data (Fig. **??**).

Repeat simulations indeed appear to recreate the induced ZRS–*Shh* contact in the mandibular cell line (MD) following TSA treatment (Fig. **??**). This is again surprising, as the MD cell lines do not express *Shh* and so were included as a negative control, with no expected changes in local chromatin structure following TSA treatment. In repeat simulations of the 88fp cell line, a close analogue of the 14fp cell line used in 4C, we see relatively little change in distance between *Shh* and ZRS (Fig. **??**).

We quantified these distances by measuring from the single bead containing ZRS to the two beads which overlap the *Shh* gene (Fig. **??**). While these are not biological replicates, just repeat simulations, we find the distance shift in MD cells is statistically significant at the level of $\alpha = 0.05$ for both bead distances (Mann-Whitney: $d_1 : p < 0.012$, $d_2 : p < 0.012$). Distances in the 88fp cell line were not significantly different following TSA treatment (Fig. **??**).

Qualitatively, there could be some observable structural dynamics caused by TSA treatment in 88fp cells. It appears potentially that part of the *Shh*–ZRS TAD becomes more loosely-packed at the ZRS side. This can be seen most clearly in two of the five simulations of the TSA treated 88fp polymer models (Fig. **??**). Given the function of TSA as a histone deacetylase inhibitor, and unpublished results showing it causes an increase in H3K27 acetylation over the ZRS, we speculate that additional acetyl groups around this locus could be causing greater repulsion between histones leading to a less-compacted structure. Potentially then, the ZRS is transitioning to a more accessible state despite no change in its physical distance relative to *Shh*.

The main result, that TSA treatment induces a *Shh*–ZRS contact in mandibular cell lines but not in limb bud, is difficult to explain and runs contrary to our expectations. 4C experiments performed over the same region reported ZRS–*Shh* contacts (Fig. **??**) but polymer models using 5C found instead that these two loci remain relatively separated with or without TSA treatment (though still much closer than expected relative to their genomic distance; Fig. **??**). One explanation for this could be the filtering method used by `AutoChrom3D` (Section **??**). Highly improbably contacts are

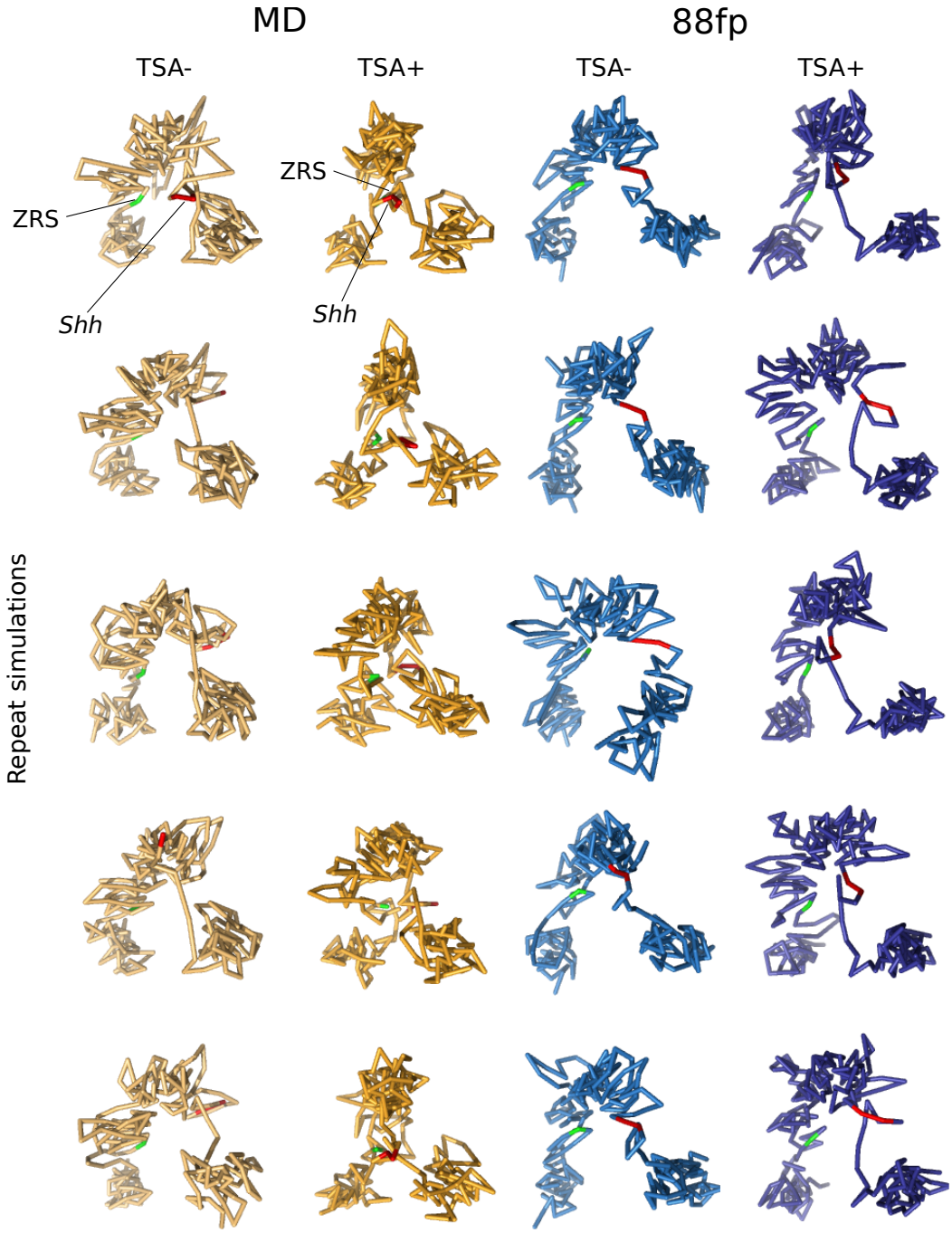

Figure 69: Repeat simulations of 3D polymer trajectories in the ZRS–*Shh* region. 3D structures are shown for 5C experiments assaying the region around *Shh* (*red*) and ZRS (*green*) in an *Shh*-expressing limb bud cell line (88fp) and a non-expressing mandibular cell line (MD). Structures were aligned as whole molecules to the uppermost replicate in each column.

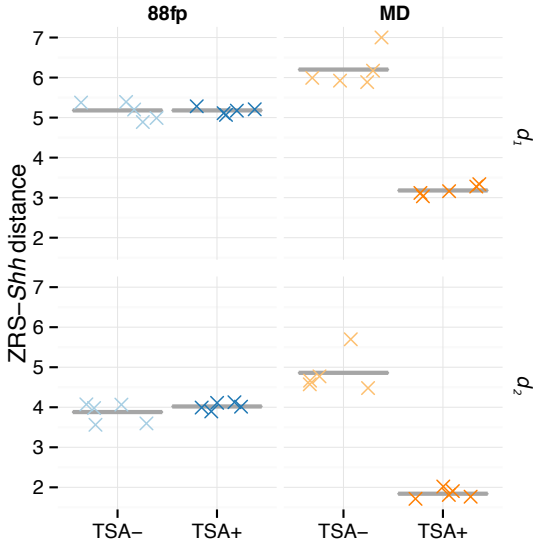

Figure 70: ZRS–*Shh* distance measurements from repeated 3D polymer simulations. Measurements were taken from 5 replicate 3D simulations (shown in Figure **??**). Distances are given in arbitrary units. *Shh* spans two beads of the polymer model, hence two distances are calculated in each case (d_1, d_2).

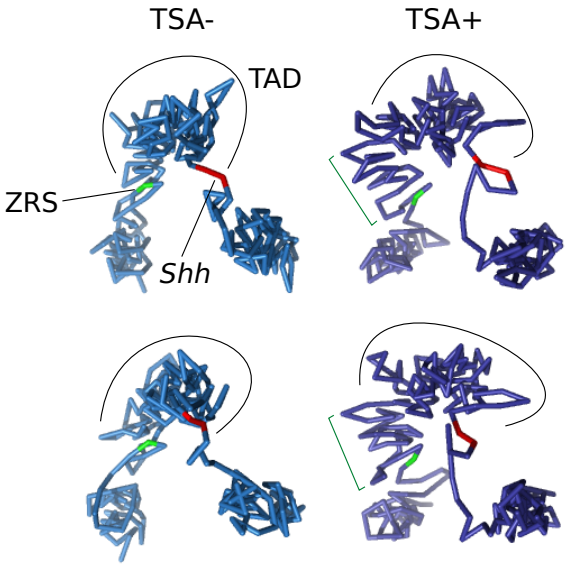

Figure 71: Polymer models showing partial TAD decompaction following TSA treatment. Two of the simulations from Figure **??** are shown here with additional annotation. In the TSA treated 14fp samples (TSA+) there is potentially evidence for a looser packing of the chromatin around the ZRS (*dark green*).

filtered before structural prediction to prevent errors or artefacts leading to aberrant structures. In this case, a genuine instance of long-range *cis*-regulation may end up being down-weighted or removed before polymer modelling. Alternatively this may be an example of where, as has been noted at high-resolution, the results of 5C as formaldehyde cross-linking efficiencies cannot be interpreted as spatial distances.[?]

Additional follow-up experiments are underway to further explore the dynamics in this region.

7 DISCUSSION

7.1 MODELLING HIGHER ORDER CHROMATIN ORGANISATION

Prior to the results presented in this thesis, much of the research into computational modelling of chromatin has been focused either on learning functional chromatin states from histone modifications and transcription factors (e.g. **? ? ? ? ? ? ? ? ? ?**), spanning small regions on the order of hundreds of basepairs, or alternatively on the inference of the overall three-dimensional chromatin fibre trajectory based on conformation data (e.g. **? ? ? ? ? ? ? ? ?**). In this work we attempt an intermediate approach, in which we use locus-level chromatin information to model higher order characteristics of nuclear architecture, such as chromosomal compartments and topological domains.

Our data show that accurate predictions of Hi-C derived chromosome compartment eigenvectors using locus-level chromatin features alone are entirely achievable (Section **??**). Generalisation across cell types further suggests that chromosome compartments could be inferred for those cell types without any available Hi-C data but with available ChIP-seq for a handful of chromatin features. For example, the NIH Roadmap Epigenomics project has generated histone modification data in hundreds of cell lines, tissues and developmental stages.[? ?] If the novel models in this work were adapted to use matched inputs, this would allow comprehensive comparisons of inferred chromosome compartments across a diverse range of conditions and cell types. In the same vein, chromosome compartments are known to be related to and recapitulate other aspects of higher order chromatin organisation, including replication timing domains, nuclear lamina associated domains and nucleolus association domains.[? ? ?] We therefore suggest a similar modelling approach could prove successful for each of these domains of interest. An exciting idea is that an integrative model capable of identifying these LADs and NADs could forward this information to a subsequent three-dimensional reconstruction algorithm, which could then use this information to generate a comprehensive, *in situ* perspective on nuclear architecture.

We had less success with the prediction of TAD boundaries (Section **??**). One reason for this is that the TAD calling algorithm used in this work[?] (Methods **??**), though a published and widely used method, produces observably flawed domain calls in some contexts. In addition the sensitivity of this method is proportional to the sample sequencing depth, which varied across our three human Hi-C datasets. Another consideration is that we resolved TAD domains to 40 kb bins, far removed from the approximately 15 basepair CTCF motifs which can generate physical domains. Indeed, given the recent release of some very-deeply-sequenced Hi-C datasets,[?] an

improved method of predicting domains might start from individual ChIP-seq peaks and consider pairs of correctly-orientated CTCF motifs. In addition, any predictive model of such domains would do well to consider the hierarchical nature of chromatin organisation (exemplified by metaTADs, Section **??**) rather than seeking simple linear discretisation of chromatin fibre into non-overlapping domains. Finally, we note that an accurate predictive model of lower levels of domain organisation, be they TADs or smaller physical domains, could likely recapitulate, on aggregate, broader domains such as compartments and metaTADs, culminating in a multi-scale model of nuclear architecture from the levels of kilobases up to entire chromosomes.

7.2 DOMAIN BOUNDARIES: FUNCTIONAL OR INCIDENTAL?

Chromatin domains have been described at multiple scales, from 5 Mb chromosome compartments[?] down to 185 kb contact domains[?] in human cells. Across all domains, many questions remain about how they are constructed and maintained. Two competing ideas are that boundary elements, akin to the classic chromatin insulators, block intra-domain contacts and the spread of heterochromatin and hence create chromatin domains; however, another suggestion is that boundary regions are rather less important and in fact an unavoidable consequence of adjacent self-interacting domains, which are perhaps instead held together through internal enhancer–promoter interactions and other contacts. The importance of boundary elements has implications for the re-establishment process of domains during the cell cycle, for example, where it has been shown that domains are entirely absent during mitosis but then re-established in early G1 phase through an as-yet-unknown mechanism.[? ?] If boundary elements bring about domains, this may hint that key boundary-binding factors are retained through mitosis, else restored through sequence motifs. The alternative, rebuilding domains through internal contacts, would require a highly-reproducible and deterministic mechanism of reconnecting specific functional interactions in sequence.

In favour of functional boundary elements, both knockdown of CTCF[?] and deletion of a specific boundary element[?] have been shown to increase inter-TAD contacts, suggesting boundaries do indeed contribute to domain delineation. In this thesis we report an array of boundary enrichments and depletions (Section **??**), which at minimum suggests some directed biological process is in effect at boundaries. Nonetheless not all observed boundary enrichments and depletions are expected to have a detectable function; it has been shown for example that removal of the H3K27me3 mark had no effect on domain boundaries.[?] One potential functional consequence of boundaries could be that genes positioned adjacent to or over a domain boundary might be most amenable to dynamic regulation, for example by associating or disassociating from the nuclear lamina. Enrichments for gene promoters have been noted at domain boundaries in this work (e.g. Section **??**) and in previous studies.[?] Alternatively, this

boundary enrichment could be due to promoter–promoter looping inducing domain boundaries.[? ? ?]

The incidental boundary hypothesis is supported by data showing that deletion of specific boundary elements, while increasing intra-TAD interactions, is insufficient to cause adjacent domains to completely merge,[?] suggesting the presence of other factors mediating domain stability. In addition, the majority of CTCF binding sites—currently thought to be the principal architects of domain boundaries—fall within TADs rather than at their boundaries (approximately 85% of human CTCF sites are non-boundary[?]). This strongly suggests CTCF binding alone is insufficient to bring about a domain boundary. Further it has been shown that the majority of enhancer–promoter contacts are tissue invariant,[?] hence if functioning as anchors of structural domains, these constitutive contacts could account for the high levels of domain conservation reported previously[? ? ? ?] and in this work (Chapter **??**).

As is the case with many biological phenomena, the question of whether boundary regions or internal contacts are responsible for chromatin domains is reductive, and it seems likely that both boundary insulation and intra-TAD contacts work together to maintain chromatin domains.

7.3 DOMAIN EVOLUTION

In this work we find an array of chromatin features that, on average, are statistically associated or excluded from TAD or compartment boundaries (Section **??**). Among these are features with a long history of studies implicating them in chromatin organisation, including CTCF and cohesin subunit RAD21. We also report enrichments for Alu repeat elements (Section **??**) but no other repeat classes. Alu repeats and CTCF are linked by evidence that CTCF binding sites have in the past been dispersed through waves of retrotransposon expansion.[? ?] This suggests a model for the evolution of topological domains, whereby purifying selection removes those inserted CTCF sites which disrupt desirable regulatory environments, while those which bring-about efficient "regulon" structures are favoured. Newly-released comparative Hi-C and CTCF datasets[?] offer an opportunity to investigate this proposed evolutionary model.

7.4 ON CAUSALITY

Throughout this thesis we have probed correlative relationships, including those between chromatin features and either expression (Section **??**), higher order chromatin structure (Section **??**), or domain boundaries (Section **??**). However even the most predictive correlations make no comment on the underlying chain of causality. Whether

genome organisation is a cause or consequence of the functions of underlying genetic elements remains an open question.[?]

Two different approaches could be used to address the causality question. A standard rejoinder is to design wet-lab experiments, for example extending Hi-C studies to perturbation or differentiation time courses, such as that performed by collaborators in Chapter **??**. However, another approach is to first develop theoretical models which, under simulation, recapitulate observed data, and then to use these models to generate testable hypotheses about the effects of specific perturbations. This latter approach is exemplified in a study by **?**] where the authors applied physical polymer modelling to deconvolute population-level 5C data into single-cell conformations. The model suggests that population-level averages are explained by transient contacts in each cell, rather than persistent loops. Subsequently these models were able to accurately predict the effects of a genetic deletion of a CTCF site separating the *Tsix* and *Xist* TADs.[?]

The models built in this thesis could also be applied to predicting the effects of experimental perturbations. For example, an experiment decreasing the tri-methylation of H3K9, perhaps through down-regulation of SETDB1 or SV39H1, might be expected to lead to heterochromatic regions becoming more permissive and allow the transcription of marked tandem repeat sequences.[?] Our models further suggest the effect would be most pronounced in K562 cells (Section **??**). A previous experiment analysed the effects of losing H3K9me3 in SETDB1 knockout mice and found increased expression of a number of endogenous retroviruses,[?] but whether these expression changes were also coupled with alterations in chromosome compartment was not tested. Performing such an experiment over a number of timepoints could help to establish whether transcriptional machinery drives genomic regions to an active compartment or *vice versa*.

7.5 INSIGHTS INTO GENOME ORGANISATION

Overall our results agree with a functional model of genome architecture whereby a majority of the genome is arranged into large static compartments (Section **??**), be they Lamina associated, nucleolus associated or central and accessible chromatin. Indeed, it seems plausible that such large, constitutive anchor points may be enough to generate a significant amount of concordance in nuclear architecture between cell types.[?] These broad similarities are coupled with local structural changes in different cell lines (Section **??**, Chapter **??**), allowing cell type specific regulation of loci through "looping out", detachment from the nuclear lamina and other conceivable mechanisms of structural variation. Whether these local changes are driven by DNA-binding proteins and chromatin remodellers or by functional contacts such as enhancer–promoter interactions remains unclear.

7.6 SUMMARY

Work presented in this thesis began with the collection and uniform reprocessing of publicly-available genome-wide Hi-C datasets (Chapter **??**). While many studies present only their own novel data, we demonstrated the utility in making use of that which is already openly-available. We compared this chromosome conformation data across three human cell types of diverse origin (human embryonic stem cell H1 hESC, transformed lymphoblastoid cell line GM12878 and the chronic myelogenous leukemic line K562), and found strong conservation of higher order chromatin structure. Where we found regions of variable structure between cell types, these were enriched for cell type specific enhancer and transcriptional activity, and also showed dramatic changes in their long-range contact profiles. These results demonstrate the close relationship between genome structure and function across three human cell types.

In Chapter **??**, we reproduced and extended a predictive model of transcriptional output, before returning to our reprocessed Hi-C data to employ a similar machine learning and model dissection paradigm. Our models of compartment eigenvectors showed high predictive accuracy and in doing so learned general associative rules between locus-level chromatin features and chromosome compartments. Probing variable importance within these models revealed some differences consistent with the biology of the cell type in which a model was learned, whereas other dissimilarities appeared to be the result of collinear clusters within our feature space (Section **??**).

We also examine boundary composition across cell types and at varying levels of higher order chromatin structure, including TADs, chromosome compartments and those of a newly-proposed layer linking the two: metaTADs (Chapter **??**). Led by these observed enrichments and depletions, we report modest success with the prediction of TAD boundaries in the absence of Hi-C. Higher-resolution chromatin conformation capture data and improved domain calling algorithms will undoubtedly enable more powerful boundary-predictive models in the near future, which in turn could allow broad comparisons of inferred higher order chromatin structure without the application of costly and time-consuming genome-wide C-methods.

In summary, we show that integrative modelling of large chromatin dataset collections can generate useful insights into nuclear architecture and seed testable hypotheses for further study. As this thesis neared completion, another study was published on the prediction of chromosome compartments;[?] while just a month earlier, a separate publication reported a predictive model of TAD boundaries built from histone modifications.[?] These very recent studies, those presented throughout this thesis, and others no doubt soon to emerge, are proving machine learning and statistical analyses to be powerful and vital apparatus for advancing our understanding of higher order chromatin organisation.

APPENDICES

Table A1: Gm12878 functional enrichments in regions of variable structure. FE: fold enrichment; FDR: false discovery rate.

Category	Term	Count	%	FE	*p*-value	FDR
GOTERM_CC_FAT	GO:0005882 intermediate filament	36	4.20	4.90	6.42E-15	8.95E-12
GOTERM_CC_FAT	GO:0045111 intermediate filament cytoskeleton	36	4.20	4.79	1.35E-14	1.87E-11
SP_PIR_KEYWORDS	keratin	31	3.62	5.64	1.72E-14	2.47E-11
INTERPRO	IPR007951:PMG	11	1.28	25.11	9.80E-14	1.56E-10

Table A2: H1 hESC functional enrichments in regions of variable structure. FE: fold enrichment; FDR: false discovery rate.

Category	Term	Count	%	FE	*p*-value	FDR
PIR_SUPERFAMILY	PIRSF003152:G protein-coupled olfactory receptor, class II	116	10.55	6.64	3.25E-68	4.41E-65
INTERPRO	IPR000725:Olfactory receptor	116	10.55	6.53	7.58E-63	1.21E-59
SP_PIR_KEYWORDS	olfaction	116	10.55	6.40	2.07E-61	2.97E-58
GOTERM_MF_FAT	GO:0004984 olfactory receptor activity	116	10.55	6.13	1.30E-60	1.97E-57
GOTERM_BP_FAT	GO:0007608 sensory perception of smell	117	10.64	5.96	1.91E-59	3.35E-56
GOTERM_BP_FAT	GO:0007606 sensory perception of chemical stimulus	118	10.73	5.37	1.71E-54	3.01E-51
KEGG_PATHWAY	hsa04740:Olfactory transduction	108	9.82	4.94	8.72E-51	1.03E-47
SP_PIR_KEYWORDS	sensory transduction	125	11.36	4.58	2.61E-48	3.74E-45
INTERPRO	IPR017452:GPCR, rhodopsin-like superfamily	131	11.91	4.03	1.40E-44	2.24E-41
INTERPRO	IPR000276:7TM GPCR, rhodopsin-like	131	11.91	4.02	1.68E-44	2.68E-41
PIR_SUPERFAMILY	PIRSF800006:rhodopsin-like G protein-coupled receptors	131	11.91	3.63	5.04E-43	6.85E-40
GOTERM_BP_FAT	GO:0007600 sensory perception	138	12.55	3.54	4.78E-41	8.40E-38
SP_PIR_KEYWORDS	g-protein coupled receptor	136	12.36	3.62	1.69E-40	2.42E-37
GOTERM_BP_FAT	GO:0050890 cognition	143	13.00	3.23	5.34E-38	9.38E-35
SP_PIR_KEYWORDS	transducer	137	12.45	3.39	1.48E-37	2.12E-34
GOTERM_BP_FAT	GO:0050877 neurological system process	163	14.82	2.72	3.85E-34	6.76E-31
GOTERM_BP_FAT	GO:0007186 G-protein coupled receptor protein signaling pathway	148	13.45	2.77	1.36E-31	2.40E-28
SP_PIR_KEYWORDS	receptor	172	15.64	2.31	3.72E-26	5.33E-23
GOTERM_BP_FAT	GO:0007166 cell surface receptor linked signal transduction	188	17.09	2.06	8.02E-24	1.41E-20
SP_PIR_KEYWORDS	cell membrane	198	18.00	1.86	5.96E-19	8.52E-16
UP_SEQ_FEATURE	topological domain:Extracellular	227	20.64	1.72	1.26E-17	2.20E-14
UP_SEQ_FEATURE	topological domain:Cytoplasmic	250	22.73	1.52	1.13E-12	1.98E-09
UP_SEQ_FEATURE	disulfide bond	211	19.18	1.56	9.11E-12	1.60E-08
SP_PIR_KEYWORDS	disulfide bond	214	19.45	1.52	6.20E-11	8.88E-08
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	285	25.91	1.41	7.31E-11	1.28E-07
GOTERM_CC_FAT	GO:0005886 plasma membrane	255	23.18	1.37	1.26E-09	1.77E-06
SP_PIR_KEYWORDS	glycoprotein	289	26.27	1.37	1.83E-09	2.61E-06
GOTERM_CC_FAT	GO:0016021 integral to membrane	328	29.82	1.27	9.34E-09	1.31E-05
SP_PIR_KEYWORDS	transmembrane	317	28.82	1.31	1.37E-08	1.96E-05
UP_SEQ_FEATURE	transmembrane region	314	28.55	1.31	2.03E-08	3.56E-05
GOTERM_CC_FAT	GO:0031224 intrinsic to membrane	333	30.27	1.24	7.49E-08	1.05E-04
SMART	SM00355:ZnF_C2H2	69	6.27	1.86	4.23E-07	5.43E-04
UP_SEQ_FEATURE	zinc finger region:C2H2-type 5	55	5.00	2.08	5.12E-07	8.99E-04
UP_SEQ_FEATURE	zinc finger region:C2H2-type 4	57	5.18	2.01	8.49E-07	0.0015
INTERPRO	IPR013087:Zinc finger, C2H2-type/integrase, DNA-binding	59	5.36	1.94	1.73E-06	0.0028
UP_SEQ_FEATURE	zinc finger region:C2H2-type 2	58	5.27	1.95	1.73E-06	0.0030
SP_PIR_KEYWORDS	membrane	372	33.82	1.21	2.69E-06	0.0038
INTERPRO	IPR015880:Zinc finger, C2H2-like	69	6.27	1.78	4.09E-06	0.0065
UP_SEQ_FEATURE	zinc finger region:C2H2-type 8	44	4.00	2.14	4.13E-06	0.0073
UP_SEQ_FEATURE	zinc finger region:C2H2-type 3	58	5.27	1.90	4.43E-06	0.0078
UP_SEQ_FEATURE	zinc finger region:C2H2-type 7	46	4.18	2.06	5.87E-06	0.0103
INTERPRO	IPR007087:Zinc finger, C2H2-type	67	6.09	1.75	9.19E-06	0.0147
UP_SEQ_FEATURE	zinc finger region:C2H2-type 6	48	4.36	1.99	9.83E-06	0.0173

Table A3: K562 functional enrichments in regions of variable structure. FE: fold enrichment; FDR: false discovery rate.

Category	Term	Count	%	FE	*p*-value	FDR
PIR_SUPERFAMILY	PIRSF038651:G protein-coupled olfactory receptor, class I	26	7.08	24.94	7.86E-30	8.99E-27
GOTERM_MF_FAT	GO:0004984 olfactory receptor activity	40	10.90	6.12	7.39E-20	1.01E-16
INTERPRO	IPR000725:Olfactory receptor	39	10.63	6.18	3.00E-19	4.29E-16
SP_PIR_KEYWORDS	olfaction	39	10.63	6.15	4.55E-19	6.09E-16
GOTERM_BP_FAT	GO:0007608 sensory perception of smell	39	10.63	5.48	1.19E-17	1.94E-14
SP_PIR_KEYWORDS	sensory transduction	44	11.99	4.60	8.72E-17	1.44E-13
GOTERM_BP_FAT	GO:0007606 sensory perception of chemical stimulus	39	10.63	4.89	6.32E-16	1.09E-12
KEGG_PATHWAY	hsa04740:Olfactory transduction	38	10.35	4.58	6.87E-16	7.22E-13
INTERPRO	IPR017452:GPCR, rhodopsin-like superfamily	43	11.72	3.72	2.96E-13	4.23E-10
INTERPRO	IPR000276:7TM GPCR, rhodopsin-like	43	11.72	3.72	3.10E-13	4.43E-10
SP_PIR_KEYWORDS	transducer	46	12.53	3.26	4.97E-12	6.65E-09
SP_PIR_KEYWORDS	g-protein coupled receptor	44	11.99	3.35	6.34E-12	8.48E-09
PIR_SUPERFAMILY	PIRSF800006:rhodopsin-like G protein-coupled receptors	42	11.44	3.26	6.34E-12	7.26E-09
GOTERM_BP_FAT	GO:0007600 sensory perception	45	12.26	3.18	1.10E-11	1.80E-08
GOTERM_BP_FAT	GO:0050890 cognition	46	12.53	2.87	1.87E-10	3.07E-07
UP_SEQ_FEATURE	zinc finger region:C2H2-type 10	27	7.36	4.64	1.94E-10	3.10E-07
UP_SEQ_FEATURE	zinc finger region:C2H2-type 1; degenerate	17	4.63	8.23	2.35E-10	3.77E-07
GOTERM_BP_FAT	GO:0007186 G-protein coupled receptor protein signaling pathway	51	13.90	2.63	2.87E-10	4.70E-07
UP_SEQ_FEATURE	zinc finger region:C2H2-type 11	25	6.81	4.91	3.32E-10	5.31E-07
UP_SEQ_FEATURE	zinc finger region:C2H2-type 9	28	7.63	4.30	4.58E-10	7.33E-07
UP_SEQ_FEATURE	zinc finger region:C2H2-type 12	23	6.27	5.27	5.15E-10	8.24E-07
SMART	SM00349:KRAB	26	7.08	4.36	7.67E-10	8.65E-07
UP_SEQ_FEATURE	zinc finger region:C2H2-type 15	17	4.63	7.40	1.17E-09	1.88E-06
UP_SEQ_FEATURE	zinc finger region:C2H2-type 7	30	8.17	3.84	1.33E-09	2.13E-06
INTERPRO	IPR001909:Krueppel-associated box	26	7.08	4.20	3.15E-09	4.49E-06
UP_SEQ_FEATURE	domain:KRAB	25	6.81	4.37	3.38E-09	5.41E-06
UP_SEQ_FEATURE	zinc finger region:C2H2-type 14	17	4.63	6.32	1.19E-08	1.90E-05
UP_SEQ_FEATURE	zinc finger region:C2H2-type 13	19	5.18	5.50	1.19E-08	1.91E-05
UP_SEQ_FEATURE	zinc finger region:C2H2-type 8	27	7.36	3.73	1.86E-08	2.98E-05
UP_SEQ_FEATURE	zinc finger region:C2H2-type 6	29	7.90	3.42	3.22E-08	5.15E-05
INTERPRO	IPR001089:Small chemokine, C-X-C	7	1.91	29.85	4.94E-08	7.06E-05
INTERPRO	IPR002473:Small chemokine, C-X-C/Interleukin 8	7	1.91	27.72	8.52E-08	1.22E-04
GOTERM_BP_FAT	GO:0050877 neurological system process	48	13.08	2.21	2.61E-07	4.27E-04
INTERPRO	IPR018048:Small chemokine, C-X-C, conserved site	7	1.91	22.83	3.35E-07	4.79E-04
INTERPRO	IPR002337:Haemoglobin, beta	5	1.36	55.44	5.04E-07	7.20E-04
INTERPRO	IPR013087:Zinc finger, C2H2-type/integrase, DNA-binding	30	8.17	2.77	1.34E-06	0.002
SMART	SM00355:ZnF_C2H2	33	8.99	2.48	1.77E-06	0.002
SP_PIR_KEYWORDS	receptor	52	14.17	2.00	2.39E-06	0.003
UP_SEQ_FEATURE	zinc finger region:C2H2-type 5	27	7.36	2.90	2.39E-06	0.004
UP_SEQ_FEATURE	zinc finger region:C2H2-type 3	29	7.90	2.70	3.79E-06	0.006
GOTERM_MF_FAT	GO:0047760 butyrate-CoA ligase activity	5	1.36	38.47	3.81E-06	0.005
INTERPRO	IPR007087:Zinc finger, C2H2-type	33	8.99	2.43	5.58E-06	0.008
PIR_SUPERFAMILY	PIRSF002522:CXC chemokine	6	1.63	20.55	6.13E-06	0.007
SP_PIR_KEYWORDS	oxygen carrier	5	1.36	35.19	6.39E-06	0.009
INTERPRO	IPR015880:Zinc finger, C2H2-like	33	8.99	2.39	7.71E-06	0.011
GOTERM_BP_FAT	GO:0007166 cell surface receptor linked signal transduction	59	16.08	1.78	9.41E-06	0.015
UP_SEQ_FEATURE	zinc finger region:C2H2-type 16	11	3.00	6.18	1.14E-05	0.018

PIR_SUPERFAMILY	PIRSF500045:hemoglobin, vertebrate type	5	1.36	29.97	1.16E-05	0.013
UP_SEQ_FEATURE	zinc finger region:C2H2-type 17	10	2.72	7.02	1.20E-05	0.019
UP_SEQ_FEATURE	disulfide bond	77	20.98	1.62	1.27E-05	0.020
UP_SEQ_FEATURE	topological domain:Extracellular	75	20.44	1.62	1.99E-05	0.032
PIR_SUPERFAMILY	PIRSF005559:zinc finger protein ZFP-36	13	3.54	4.58	2.22E-05	0.025
SP_PIR_KEYWORDS	disulfide bond	78	21.25	1.59	2.44E-05	0.033
UP_SEQ_FEATURE	zinc finger region:C2H2-type 20	7	1.91	11.56	2.64E-05	0.042
SP_PIR_KEYWORDS	blood	5	1.36	25.59	2.89E-05	0.039
SP_PIR_KEYWORDS	cell membrane	63	17.17	1.70	3.07E-05	0.041

Moore *et al. Genome Biology* (2015) 16:110
DOI 10.1186/s13059-015-0661-x

Genome Biology

RESEARCH Open Access

Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization

Benjamin L Moore, Stuart Aitken and Colin A Semple*

Abstract

Background: Interphase chromosomes adopt a hierarchical structure, and recent data have characterized their chromatin organization at very different scales, from sub-genic regions associated with DNA-binding proteins at the order of tens or hundreds of bases, through larger regions with active or repressed chromatin states, up to multi-megabase-scale domains associated with nuclear positioning, replication timing and other qualities. However, we have lacked detailed, quantitative models to understand the interactions between these different strata.

Results: Here we collate large collections of matched locus-level chromatin features and Hi-C interaction data, representing higher-order organization, across three human cell types. We use quantitative modeling approaches to assess whether locus-level features are sufficient to explain higher-order structure, and identify the most influential underlying features. We identify structurally variable domains between cell types and examine the underlying features to discover a general association with cell-type-specific enhancer activity. We also identify the most prominent features marking the boundaries of two types of higher-order domains at different scales: topologically associating domains and nuclear compartments. We find parallel enrichments of particular chromatin features for both types, including features associated with active promoters and the architectural proteins CTCF and YY1.

Conclusions: We show that integrative modeling of large chromatin dataset collections using random forests can generate useful insights into chromosome structure. The models produced recapitulate known biological features of the cell types involved, allow exploration of the antecedents of higher-order structures and generate testable hypotheses for further experimental studies.

Background

The chromatin structure of human interphase chromosomes plays critical roles in a wide range of cellular functions and consists of many hierarchically arranged but interconnected layers of structure. These range from the three-dimensional arrangement of multi-megabase-scale domains within the nucleus down to the chemical modifications carried by individual nucleosomes and nucleotides at particular loci. A recurring question has been how these many different levels of chromatin structure are related to one another [1]. In the wake of recent efforts to comprehensively map the epigenomic landscape in human cells, integrative approaches have suggested classifications of chromatin into distinct, functional states. The number of chromatin states identified in these pioneering studies has varied widely, from as few as 6 to as many as 51, using a variety of locus-level features such as DNA methylation, histone modifications and transcription factor binding patterns [2-5]. These states usually encompass small, sub-genic regions and have provided intriguing insights into chromatin-mediated variation in promoter and enhancer activity. At the same time technological developments such as the Hi-C method have provided datasets describing the overall spatial organization of the human genome [6], but the relationships between such datasets and the wide spectrum of locus-level features are not well understood. A recent study examining seven such features and their relationships to the spatial organization of the mouse genome in embryonic stem cells (ESCs) concluded that chromosome architecture is largely determined by the

*Correspondence: colin.semple@igmm.ed.ac.uk
MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Road, Edinburgh, EH4 2XU, UK

binding patterns of particular transcription factors, and that these cells have a unique higher-order chromatin structure as a result [7]. Thus it is unclear whether such results are relevant to other cell types and species, or whether the inclusion of a broader range of features would provide additional insights.

Many aspects of higher-order chromatin remain broadly invariant between cell types, and genome-wide datasets as diverse as replication timing domains, lamin association domains and Hi-C interaction matrix eigenvectors show strong correlations across many different human cell lines [8]. Indeed, most measurable aspects of higher-order structure have been conserved during evolution across the majority of the mammalian genome [8-10]. However, a minority (perhaps 20% to 30%) of the genome is within more labile structures, such that the behaviors of many replication timing domains and lamin association domains change significantly upon cellular differentiation from ESCs, altering the transcriptional output of many resident genes [10,11]. A large literature surrounds the dynamics of locus-level chromatin during differentiation and reprogramming, emphasizing the critical importance of genomic patterns of DNA binding proteins, particular histone modifications and DNA methylation (for example, [12]). Yet we still lack an integrated view of chromatin dynamics that details the dependencies between these locus-level phenomena, the remodeling of large domains and changes in nuclear organization. The extent to which higher-order chromatin dynamics depends upon the spectra of features occurring at these lower levels has not been studied quantitatively.

Given the existence of neighboring chromatin domains with distinct structures and activities, the boundaries defining such domains have been a focus of particular interest. The topological domains (TADs) described by Dixon et al. [9] were reported to be separated by boundary regions showing pronounced peaks of the insulator binding protein CTCF, although depletion of CTCF appears to have little effect on TAD boundaries [13]. Similarly, deletion of a TAD boundary on the mouse X chromosome resulted in many altered interactions, but did not cause the two TADs separated by this boundary to completely merge [14]. Thus there is much left to learn about the basis of TAD boundaries. The scale of TAD organization (median length 880 kb) is below that of the multi-megabase chromatin domains delineating occupancy of A and B nuclear compartments [15]. These compartments constitute domains of transcriptionally active, relatively centrally positioned chromatin, and relatively inactive, peripheral chromatin respectively; consequently compartment boundaries often mark a profound divergence in functional state. It is not known whether TAD boundaries coincide with compartment boundaries, and the similarities or differences in the features underlying these two boundary classes also remain unstudied.

Here we exploit the unprecedented volumes of data produced recently [4] to provide an integrated and rigorously quantitative view of locus-level chromatin features, higher-order chromatin structure and nuclear organization across three cell types. We use integrative modeling approaches to directly study the contribution of 35 locus-level chromatin features to chromosome architecture across three human cell types as measured by Hi-C. These data are relevant to the quantitative, molecular basis of higher-order chromatin, the dominant determinants of chromatin dynamics, and prominent features conferring the structure of domain boundaries.

Results

Higher-order chromatin organization is largely concordant and predictable across cell types

In common with previous studies of higher-order chromatin structure [8-11], there was evidence for good concordance of Hi-C data between different cell types. Hi-C eigenvectors were calculated for three human cell types (GM12878, H1 hESC and K562 cell lines) using the same analysis protocols, and were found to be strongly and significantly correlated (Figure 1; Additional file 1: Figure S1). Most 1-Mb regions appear to be constitutively present (that is, across cell types) in either the A or B compartments, corresponding to relatively centrally positioned, transcriptionally active or more peripheral repressive chromatin, respectively [15]. Strong correspondence across cell types was also observed for TAD boundaries, and for the positioning of compartment boundaries, separating A and B compartments (Additional file 1: Figure S2).

Although it is often assumed that higher-order chromatin domain organization (at the megabase scale) across the genome is to some degree dependent upon lower-level features (at the scale of tens or hundreds of base pairs), the identity and independent contributions of these features are unknown. Beyond this it has also been unclear whether there are strong enough dependencies to allow accurate prediction of higher-order structure. For each of the three Hi-C eigenvector datasets corresponding to the Tier 1 ENCODE cell lines (GM12878, H1 hESC and K562) we assembled datasets of 35 matched locus-level chromatin features, including sites bound by 21 DNA binding proteins, and 11 histone modifications/variants and DNase hypersensitive sites (see Materials and methods). The GC content of each 1-Mb region, which is known to be correlated with higher-order structure (for example, [8]), was also included as an additional feature in each model for comparison with chromatin features. Importantly, each Hi-C dataset was re-analyzed to provide comparable identically processed data, which

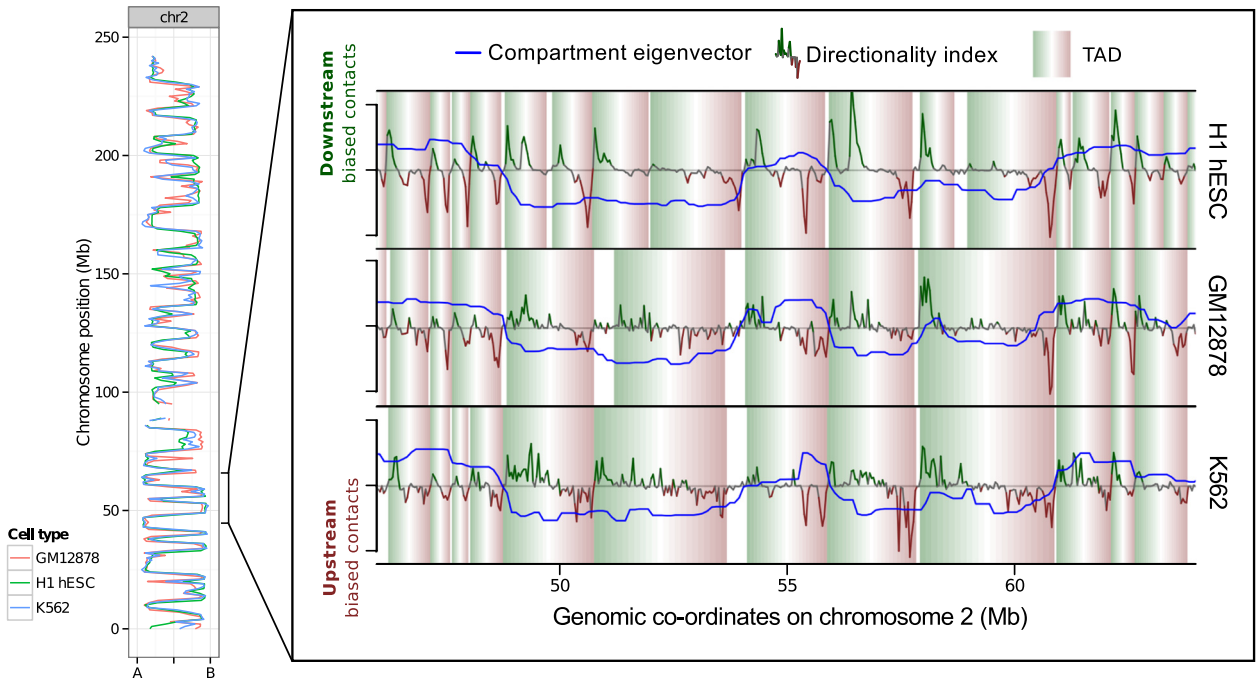

Figure 1 Concordance of chromatin structure at multiple scales over three human cell types. The eigenvector compartment profile is shown for chromosome 2 for three human cell types (left). Genome-wide Pearson correlation coefficients between eigenvectors at 1-Mb resolution are in the range 0.75 to 0.80 (Additional file 1: Figure S2B). At higher resolution, the zoomed region illustrates conservation of topological domains (TADs) over 20 Mb of the same chromosome. Genome-wide, 33% of all H1 TAD boundaries have a matching boundary in GM12878 in the same or an adjacent 40-kb bin (K562: 31%, null model: 18%; Kolmogorov–Smirnov test: $D = 0.26$, $p \approx 0$). Similarly, for H1 compartment boundaries, 37% have a matching boundary in the same or an adjacent 100-kb bin in GM12878 (K562: 35%, null model: 7%; Kolmogorov–Smirnov test: $D = 0.47$, $p \approx 0$, Additional file 1: Figure S2A). Mb, megabases; TAD, topological domain.

was complementary to the identically processed, locus-level ENCODE data. It was possible to construct random forest models with good predictive accuracy, and strong and significant correlations were seen between predicted and empirically measured eigenvector values for each cell type (Figure 2). The models show high predictive power, particularly for GM12878 where the model achieved a Pearson correlation coefficient (PCC) of 0.805 between predicted and measured values. These levels of accuracy are similar to those reported (median PCC = 0.83 over seven cell types) for strikingly successful models of the transcriptional output of promoters using locus-level chromatin features [16]. Other evaluation metrics also suggested successful models, such as the ability to correctly assign 1-Mb regions to compartments A and B (see area under the receiver operating characteristic data in Figure 2). It would be feasible to construct similar, but more comprehensive models using all ENCODE chromatin features for a given cell type, although the resulting models would not be comparable between cell types. However, the high accuracy of the current models suggests there is limited potential for improvement by adding further features. Also, even the most comprehensive models that could be constructed, using all currently available data, inevitably represent a minority of the features actually present in chromatin [1].

While 1-Mb compartment eigenvectors are low resolution relative to that typically employed for chromatin immunoprecipitation sequencing (ChIP-seq) data, megabase bins are a suitable choice for analyzing large chromosomal compartments [15,17]. To confirm our modeling accuracy is not sensitive to resolution, we applied models trained with 1 Mb to 100 kb resolution datasets and saw similarly high levels of accuracy (88% to 95%, as accurate as 1-Mb models in terms of predicted and empirical PCC, Additional file 1: Figure S3).

Influential features underlying higher-order structure differ between cell types

Given the correlations seen between Hi-C eigenvectors from different cell types (Figure 1) and the similar predictive power of cell-type-specific models (Figure 2A), one might assume that a similar combination of informative variables appears in each of the models. The broad trends in relative variable importance (see Materials and methods) do indeed suggest that many features have a similar influence in each of the three models (Additional file 1: Figure S4A). For example the genomic distributions

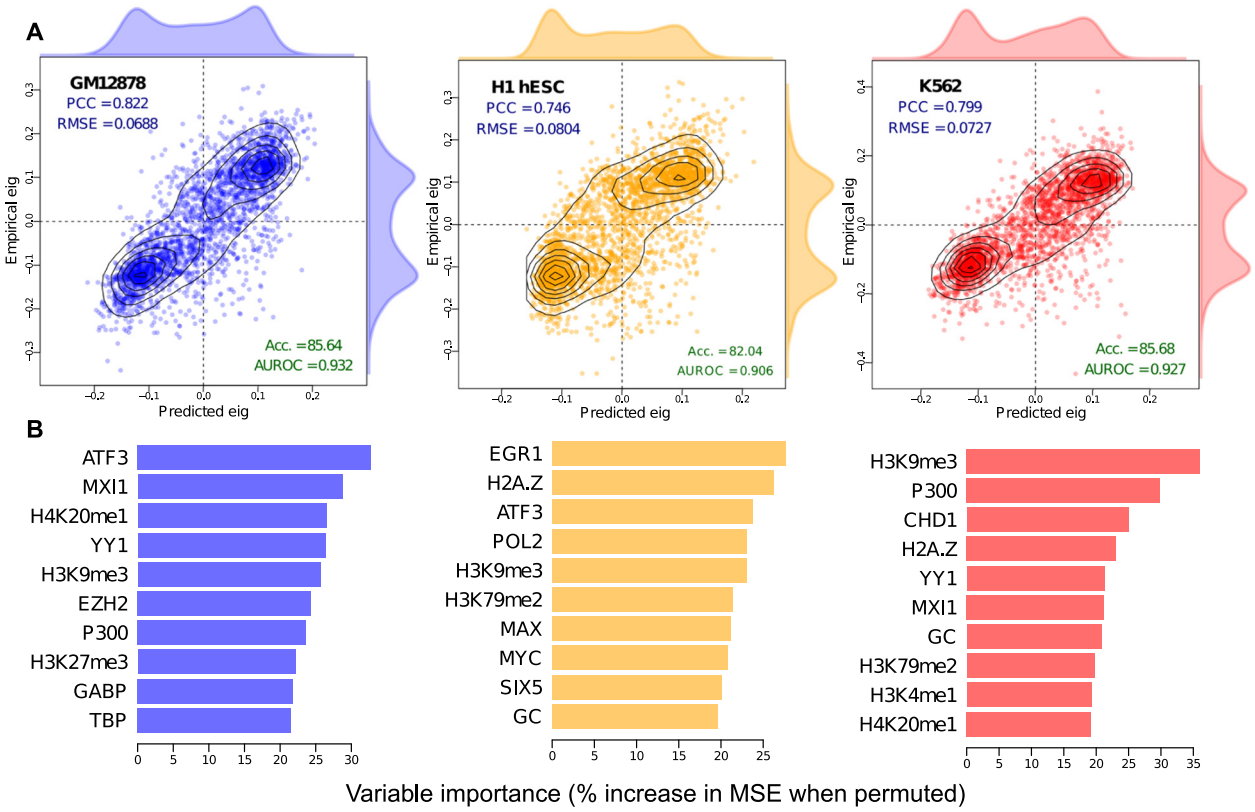

Figure 2 Accurate models of higher-order chromatin state built from locus-level features. **(A)** Model predictions (Predicted eig) are compared to observed values (Empirical eig). Various metrics are used to measure the accuracy of regression modeling – the Pearson correlation coefficient (PCC) and root mean-squared error (RMSE) – and to evaluate classification accuracy (for A $eig \geq 0$ and B $eig < 0$) – accuracy (percentage of true positives, Acc.) and area under the receiver operating characteristic curve (AUROC). **(B)** Variable importance (shown for the ten most informative features per model) is calculated as the decrease in the accuracy of predictions for the permuted variable relative to observed (in units of percentage increase in MSE), averaged over the forest (see Materials and methods). Acc., accuracy; AUROC, area under the receiver operating characteristic curve; eig, eigenvector; MSE, mean-squared error; PCC, Pearson correlation coefficient; RMSE, root mean-squared error.

of CTCF binding patterns, H3K36me3, H3K27ac and GC content maintain very similar influence across all three models, while certain variables depart from this trend and show a notably higher variable importance in a particular model. Thus substantial levels of variation between cell types are seen for the top ten most influential variables across models (Figure 2B), such that the repressive histone modification H3K9me3 is the only feature, among the ten most influential, shared between all three cell-type models. This is expected since H3k9me3 is anticorrelated or uncorrelated with most other input features (Additional file 1: Figure S5), and is therefore a relatively information-rich variable. Overall, more highly ranked features are shared between the two relatively differentiated, hematopoietic cell lines (GM12878 and K562), with the pluripotent ESC line (H1 hESC) showing more distinct characteristics. The EGR1 transcription factor plays critical roles in cellular differentiation and shows markedly higher variable importance in the H1 hESC model. While the P300 transcriptional co-activator protein, which controls the proliferation and differentiation of hematopoietic progenitor cells, ranks more highly in the two hematopoietic cell line models (Figure 2B, Additional file 1: Figure S4).

Many of the variables examined here are heavily interdependent, and for example co-occur in clusters denoting functional chromatin states [4]. Care must be taken not to over-interpret the differences in variable importance between models, given the pervasive multi-collinearity and clustering between variables in the input locus-level feature set (Additional file 1: Figure S5). For instance, MXI1 is an influential feature in both the hematopoietic models, while MYC and MAX are among the highest ranked features in the H1 hESC model. This is in keeping with recent results suggesting MYC binds open chromatin as a transcriptional amplifier in ESCs [18,19], with MAX and MXI1 long being known as antagonistic co-regulators of MYC [20]. Thus, in identifying nominally different informative variables for each model we will, to some extent, select different representatives of

the same cluster (Additional file 1: Figure S5). It follows that we would expect a large number of different feature combinations to have similar predictive power in broadly equivalent random forest models. With a broader perspective, there are general similarities across all three models, in that all derive much of their predictive power from indicators of transcriptional activity, markers of heterochromatin and the binding levels of combinations of broadly expressed transcription factors (Additional file 1: Figure S6).

Consistent with the presence of broad commonalities among the three models, cross-application of models showed that models trained in one cell type often performed well in another (Figure 3). In each instance of cross-application, predictive accuracy declined by no more than 21% relative to the model's native cell type. In reciprocal crosses between the two hematopoietic cell lines (K562 and GM12878), this loss of accuracy was between 5.9% and 7.8% (Figure 3A), but was 20.2% to 20.4% when these models were applied to H1 hESC data.

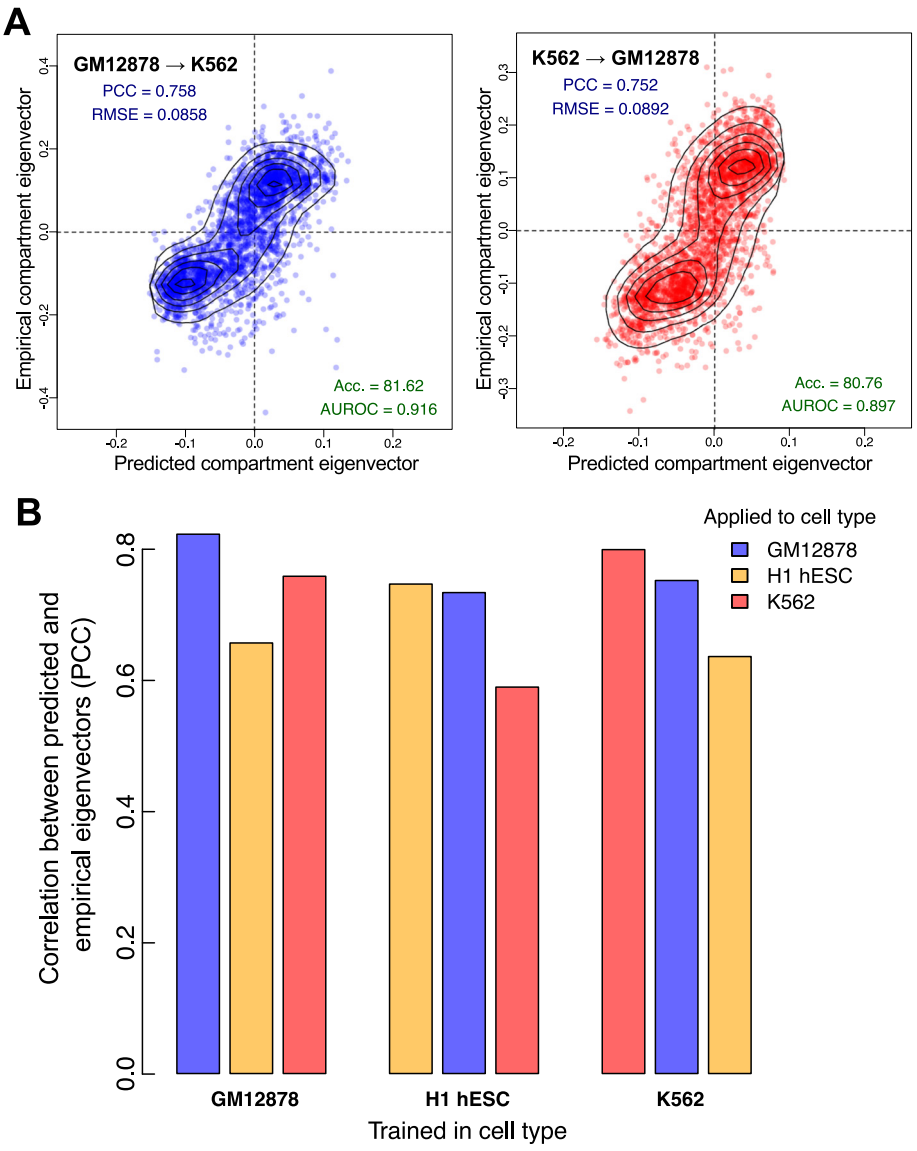

Figure 3 Models trained in one cell type can generalize to others. Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. **(A)** The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross (PCC = 0.75). **(B)** In each case, predictive accuracy decreased on cross-application but there remains significant agreement between predicted and empirical values. Acc., accuracy; AUROC, area under the receiver operating characteristic curve; PCC, Pearson correlation coefficient; RMSE, root mean-squared error.

This again highlights the relatively unusual structural features of the pluripotent state.

We compared the performance of our random forest approach with two other regression methods: simple multiple linear regression and partial least squares regression, a method particularly well suited to highly correlated inputs [21]. While cell-type-specific prediction accuracy remained high for each method, cross-application between cell types confirmed our random forest approach as that most capable of learning generalizable rules of compartment prediction (Additional file 1: Figure S7).

Regions of variable structure are enriched for cell-type-specific enhancers

Although the chromatin organization of much of the genome appears to be invariant between cell types (Figure 1), some regions are more dynamic. There is a clear relationship between modeling accuracy and structural stability between cell types such that the structures of more variable regions are more challenging to predict. This is evident even with the most liberal definitions of variability; for instance, if we calculate the median absolute deviation between eigenvectors across all three cell types and simply trisect the distribution, we found that the most structurally variable regions between cell types were significantly less accurately modeled in each case (Figure 4A). This could indicate the cell-type-specific features responsible for organizing these regions are largely missing from our training set, which undoubtedly represents a tiny minority of all the actual components of chromatin in real human cells. However, it is unclear whether structural variability defined so broadly reflects altered biological function or is dominated by stochastic variations in structure among cells [22].

A more conservative definition of structurally variable regions is that they are regions altering their compartment state (between A and B compartments) in one cell type relative to the other two. Such regions will often undergo dramatic changes between transcriptionally permissive and repressive environments and might be expected to be associated with cell-type-specific biology, such as functional chromatin states [4]. This indeed seems to be the case, with regions occupying altered compartments showing corresponding changes in enhancer activity. Regions undergoing a B to A compartment transition, to a relatively transcriptionally permissive structure, were enriched for cell-type-specific enhancers in the two derived cell types used in this study but not in the ESC line, which would not be expected to have lineage-specific enhancer contacts active in its pluripotent state (Figure 4B). The same pattern was not seen for enhancers shared between two or more of the cell types under study. We observed a similar enrichment for cell-type-specific transcription (Additional file 1: Figure S8) but not for several other chromatin states including promoter activity (Additional file 1: Figure S9).

For each cell line, we identified all regions showing cell-type-specific occupancy of the active A compartment and ranked these regions according to the density of predicted active enhancers. Close examination of these regions reveals many examples of enhancer

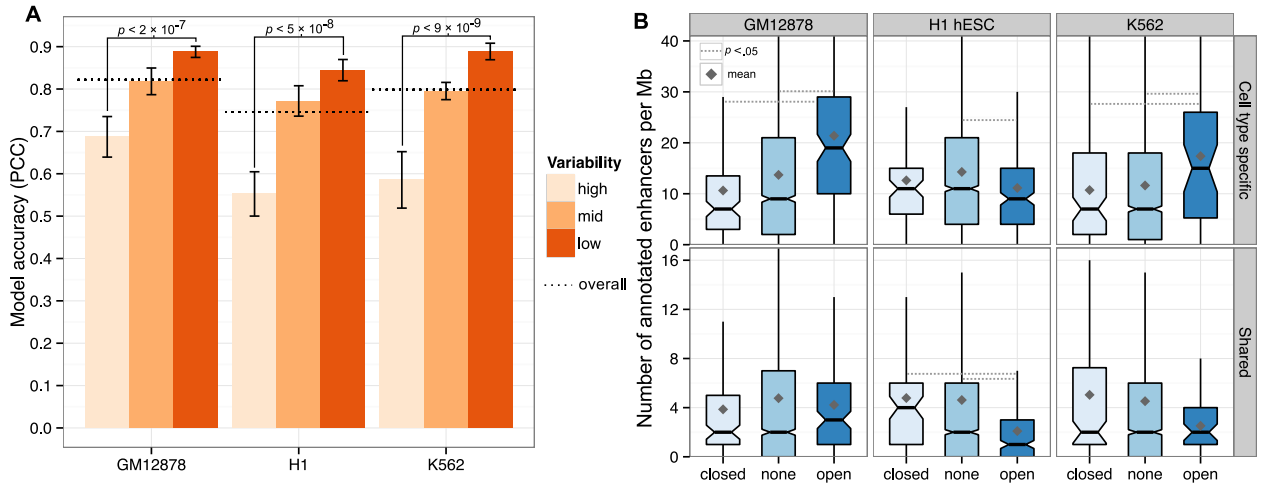

Figure 4 Regions with variable structure are less successfully modeled and are associated with cell-type-specific enhancer activity. **(A)** Model accuracy is significantly different between low- and high-variability regions, defined as, respectively, the lowest and highest thirds of the distribution of median absolute deviations between cell types. **(B)** Regions occupying altered compartments are defined as those assigned to A in one cell type but to B in the other two cell types, or vice versa. The numbers of enhancers (cell type specific or shared between two or more cell types) are shown for regions with altered (open or closed) and non-altered (none) compartments in each cell type. PCC, Pearson correlation coefficient.

activity nucleated upon genes associated with cell-type-specific biology (Figure 5A, Additional file 1: Figure S10). For the GM12878 (B-cell derived) cell line, an active region of variable structure rich in active enhancers was found to contain the EBF1 (early B-cell factor 1) gene (Figure 5A). The transcription factor encoded by this gene has been identified as essential in maintaining B-cell identity and establishing early lineage commitment [23,24]. Similarly a variable region active in H1 hESC (Additional file 1: Figure S10B.1) harbors the PAX1 regulator of patterning during embryogenesis [25], while a K562-specific active region (Additional file 1: Figure S10C.3) contains a gene encoding a regulator of hematopoiesis (ZFPM2/FOG2 [26]). Each example is concordant with the known biology of the cell type concerned, and each is illustrative of the genome-wide relationship between higher-order structural variability and cell-type-specific enhancer activity (Figure 4B). We explored the functional annotations of genes in regions of cell-type-specific structure (Additional file 2: Tables S1, S2 and S3), and although we observed some artifactual enrichments (generated by duplicated gene clusters within some of these 1-Mb regions), no significant enrichments were seen across regions.

A defining characteristic of active A compartment regions is a preferential bias in contacting other A compartment regions [15]. However, it is not clear whether cell-type-specific transitions in higher-order structure are solely compartment-level phenomena, or involve other structural strata. We therefore examined the genome-wide contact profiles of each region of variable cell-type-specific chromatin structure in detail. If these

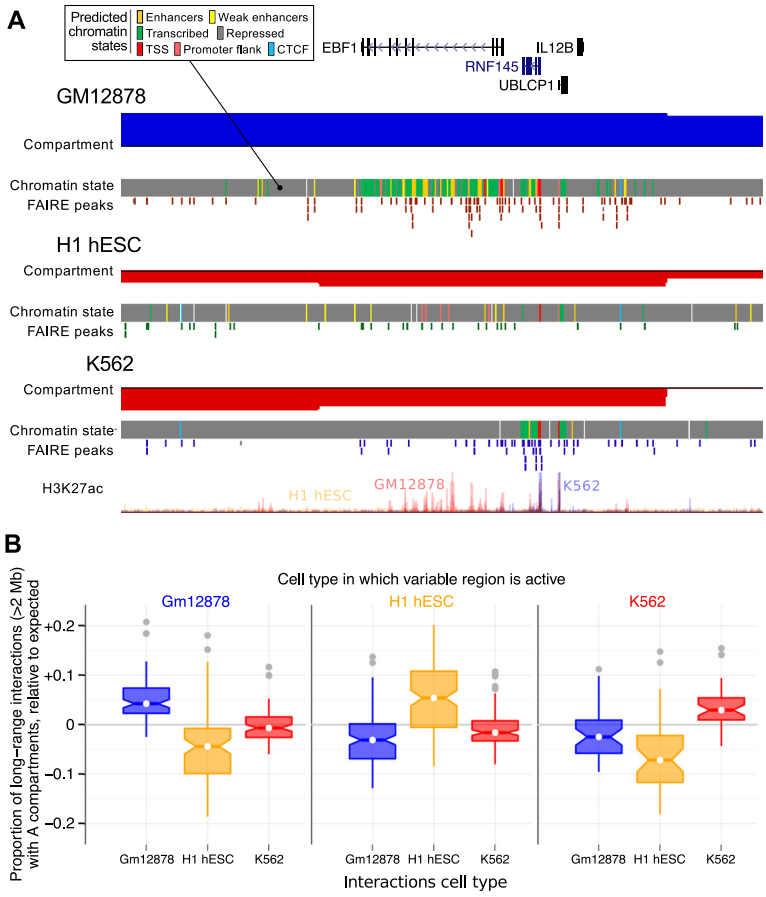

Figure 5 Structurally variable regions indicate cell-type-specific biology. Regions occupying the active A nuclear compartment in one cell type, but the repressed B compartment in the other two, were selected and ranked by the number of predicted active enhancers (Figure 4). **(A)** The region chr5:158–159 Mb, which occupies the open A compartment in GM12878 cells, is shown as an example (top five regions for each cell type are shown in Additional file 1: Figure S10). Displayed tracks are: known genes (UCSC), compartment eigenvectors, `chromHMM/Segway` combined chromatin state predictions, open chromatin FAIRE peaks, and H3K27ac signal. **(B)** Structurally variable regions show a greater than expected proportion of contacts with other active A compartments, in the cell type in which they are active relative to those same regions in the other two cell types. Box plot notches represent 95% confidence intervals of the median. Each variable region is also shown individually in Additional file 1: Figure S11. TSS, transcription start site.

cell-type-specific structures are mediated by finer-scale structural levels (such as TADs) we might expect to see predominantly short-range contacts in their underlying contact profile. Instead, we found that variable regions preferentially interact with other A compartment regions in the cell types in which they are active (Figure 5B, Additional file 1: Figure S11), but not in the other cell types in which they are inactive. This supports the idea that these cell-type-specific regions are undergoing compartment-level transitions, disproportionately mediated by the formation of long-range contacts, while also not precluding additional changes at lower levels such as TADs.

TAD boundaries and compartment boundaries possess similar features

The mammalian genome is organized into TADs, predominantly self-interacting chromatin domains, with boundary regions reportedly associated with pronounced peaks and troughs of particular features within 500 kb of the predicted boundary [9]. Exploration of this phenomenon using a set of 24 mouse ESC chromatin features (and a smaller number of human ESC features) reportedly revealed enrichment peaks of CTCF, H3K4me3 and H3K36me3, as well as a pronounced dip in H3K9me3, suggesting that high levels of transcription may contribute to boundary formation [9]. However, it was unclear whether other features show unusual patterns in TAD boundary regions, and whether the constellation of features involved changes between cell types. The features associated with boundaries separating A and B compartments calculated from Hi-C eigenvectors have not been studied to our knowledge. The datasets assembled here, consisting of 35 matched chromatin features across three cell types, allow us to conduct the first comparative study of the constituents of human TAD and compartment boundary regions.

We derived TAD boundaries according to established methods [9] for all three cell types under study. We then sought evidence for significantly enriched or depleted features at TAD boundary regions using a conservative approach (a nonparametric statistical test and Bonferroni multiple testing correction, see Materials and methods), and confirmed the previously reported peaks (CTCF and POL2) and dip (H3K9me3) in ESC data, but also revealed substantial heterogeneity between cell types. CTCF binding was found enriched at TAD boundaries across all cell types, but other features, including H3K36me3 and H3K4me3, show dramatic peaks of enrichment in H1 hESC cells that are not seen consistently in other cell types (Figure 6, Additional file 1: Figure S12). Although the dip in H3K9me3 at TAD boundaries is seen in all cell types, the extent of the depletion varies and is weakest in H1 hESC cells. Many other features show significant, though often modest, enrichments in a particular cell type. However, overall the complexity of TAD boundaries (measured as the number of strongly enriched features) is notably higher in H1 hESC than in the other two, more differentiated, cell types (Figure 6), involving large increases in the binding of sequence specific factors such as SP1 and JUND.

Across all three cell types, several features demonstrate consistent and statistically significant patterns at TAD boundaries (Figure 6, Additional file 1: Figure S12), including peaks associated with active transcription of genes (POL2 and H3K9ac) and dips in H3K9me3, as previously reported [9]. However, other novel feature peaks of interest emerge across cell types, such as peaks of H4K20me1, a modification previously implicated in chromatin compaction [27]. Significant peaks in YY1 are evident in all cell types, which is intriguing given the evidence that YY1 and CTCF cooperate to affect long-distance interactions [28]. Co-binding of CTCF with YY1 has also been shown to identify a subset of highly conserved CTCF sites [29]. Co-binding of CTCF and YY1 may also therefore be a contributing factor in the establishment of TAD boundaries, which appear to be broadly conserved across mammals [9]. To test this, we split our sets of TAD boundaries into those possessing ChiP-seq peaks (region peaks called by ENCODE [4]) for CTCF, YY1, both CTCF and YY1 (overlapping peaks) and neither. We then tested each boundary subset for genome-wide enrichments of the other features in our dataset (Additional file 1: Figure S14). Unexpectedly, we found that boundaries marked by YY1 (without overlapping CTCF peaks) were generally most strongly enriched for other features in our dataset. We also found that boundaries lacking both CTCF and YY1 peaks showed instead the strongest enrichments for RAD21 in each cell type (Additional file 1: Figure S14), reinforcing previous findings that describe the distinct influences of CTCF and cohesin in organizing chromatin structure [13,30,31]. We also observe consistent increases in GC content at TAD boundaries, at a scale that is difficult to reconcile with the presence of smaller-scale features such as repeat elements or CpG islands (Additional file 1: Figure S12).

Where neighboring genomic regions occupy contrasting A and B nuclear compartments, the disparity implies the presence of a boundary region. Putative compartment boundaries were identified by using a hidden Markov model to infer the state sequence of A/B compartments across the genome based on observed principal component eigenvectors. Analogously to the TAD boundary analysis, we then sought significant enrichments or depletions in 36 chromatin features over these compartment boundaries (Figure 6, Additional file 1: Figure S13). Compartment boundaries display similar spectra of enrichments to previously studied TAD boundaries [9] but at

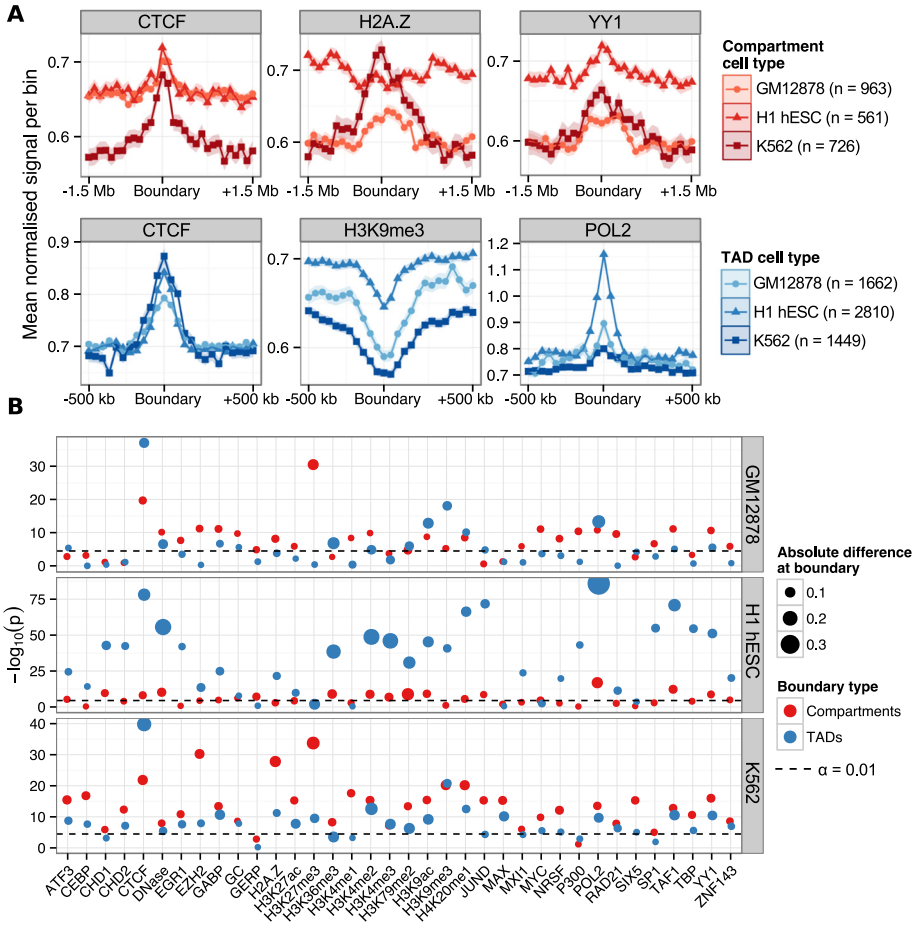

Figure 6 Chromatin features underlying TAD and compartment boundaries. **(A)** Selected profiles for locus-level features are shown for TAD boundaries (CTCF, H3K9me3 and POL2) and compartment boundaries (H2A.Z, H3K4me2 and YY1), as a mean normalized ChIP-seq signal relative to input chromatin per bin (±1 standard error). TAD boundaries were examined over 40-kb bins over the 1 Mb flanking each boundary; compartment boundaries were examined over 100-kb bins over 3 Mb. **(B)** The significance of enrichment or depletion ($-\log_{10} P$ two-tailed Mann–Whitney test) of a feature was calculated as the boundary bin relative to the ten most peripheral bins (five either side). Points are scaled by the absolute mean difference in signal over the boundary relative to the mean of peripheral bins. ChIP-seq, chromatin immunoprecipitation sequencing; TAD, topological domain.

lower resolution, reflecting the different scales of these levels of organization (Figure 6B, Additional file 1: Figure S13). Peaks associated with active promoters (POL2, TAF1 and H3K9ac) are again evident. Parallel enrichments of CTCF, YY1 and H4K20me1 are also seen at compartment boundaries, as they were for TAD boundaries, in each cell type under study. In addition, compartment boundaries show enrichments of H3K79me2, which is known to play critical roles in cellular reprogramming [32]. Remarkably, H3K79me2 has also recently been shown to mark the borders of small regions of open chromatin (hundreds of base pairs) [33]. Thus, there may be similarities in chromatin compaction boundaries at very different scales.

Certain features show intriguing contrasts between cell types. The histone variant H2A.Z lacks any trace of enrichment at H1 hESC compartment boundaries, but is significantly enriched in the other two cell types (Figure 6A), consistent with reports describing H2A.Z relocation during cellular differentiation [34]. Compartment boundaries also show enrichment for the cohesin complex subunit RAD21 in the two hematopoietic cell types (Additional file 1: Figure S12), and cohesin is another factor implicated in modulating nuclear architecture in partnership with CTCF [13]. Various other enrichments with very modest effect sizes are also evident at compartment boundaries (Figure 6B, Additional file 1: Figure S13). In contrast to TAD boundaries, the composition of compartment boundaries appears least complex in H1 hESC, relative to the other two cell types. Overall compartment and TAD boundaries are associated with overlapping spectra of chromatin features across cell

types. These involve DNA-binding proteins implicated in chromosome architecture (CTCF, YY1 and RAD21), but also implicate the initiation and repression of transcription as critical to boundary formation. However, these two boundary classes occur at different scales, with patterns of informative features typically spanning regions up to 500 kb for TAD boundaries, and patterns associated with compartment boundaries often spanning more than 1 Mb (Additional file 1: Figure S12, Additional file 1: Figure S13).

Topological domains cluster by epigenetic enrichments

Sexton et al. [35] showed that, in the *Drosophila* genome, topological structures termed physical domains could observably be clustered into distinct functional groups based on their average feature enrichments. It is of interest to repeat this experiment with our human datasets and across multiple cell types to detect finer delineation of chromatin state beyond A and B compartmentalization. We found that TADs called across the three cell types used in this work could be clustered into transcriptionally active (active), repressed heterochromatin (null) and polycomb-associated (PcG) domains, based on the patterns of DNase hypersensitivity, H3k9me3 and H3k27me3, respectively (Additional file 1: Figure S15). This analysis reveals that active compartments typically cover both active and PcG-associated TADs, while B compartments appear more homogeneous and are composed mostly of H3k9me3-enriched heterochromatin even when considering fine-grained TAD structures rather than megabase-sized genomic blocks.

Discussion

The recent abundance of epigenomic data for model cell types has enabled accurate modeling of the transcriptional output of human promoters, and a rigorously quantitative assessment of the most influential chromatin features underlying gene expression [16]. We have shown that it is possible to construct comparable models describing the features underlying higher-order chromatin structure, and that their predictive accuracy can be high. Our analysis exploits Hi-C datasets that have been re-analyzed, from the initial sequence read mapping onwards, identically for three different cell types. These data were collated with 35 locus-level ENCODE chromatin datasets, also processed identically, and matched across the same cell types. In common with previous studies [8,9], we observed good concordance of higher-order chromatin structure, reflected in Hi-C data, between different cell types. Random forest models summarized the important relationships among these many variables, providing insights into the quantitative contributions of locus-level chromatin features to higher-order structures. Although certain features were notably more influential in a particular cell type, the models shared overlapping constellations of informative features, allowing the cross-application of models between cell types.

Integrative analyses of locus-level chromatin data have allowed the prediction of functional chromatin states [2-5] but these states typically encompass small regions such as the enhancers examined here. The prediction of higher-order chromatin domains has received much less attention, and it was not clear until now that sufficient data existed to allow accurate predictions. Our data show that accurate predictions of Hi-C-derived eigenvector values, and the nuclear compartment domains based upon them, are entirely feasible. Strong and significant correlations are seen between cell types for a variety of human higher-order domains, delineating variation in replication timing, lamin association and nuclear compartments derived from Hi-C eigenvectors [8]. The data presented here therefore suggest that a variety of such domains could be successfully modeled. Given that the binding patterns of most human chromatin components have not yet been mapped, the models presented here are remarkably successful, though will undoubtedly improve with further data and algorithm development. These models also allowed us to probe the features underlying regions with variable higher-order structure between cell types, revealing enrichments of cell-type-specific enhancer activity, and suggesting links between functional chromatin states and higher-order domain dynamics. It is not possible to distinguish cause and effect using the current data, but it seems likely that the alterations in domain organization occur prior to enhancer activity.

The current data suggest that the contributions of certain locus-level chromatin features to higher-order structures vary between cell types. Striking examples include the strong influence of H3K9me3 in K562 leukemia cells, and EGR1 binding in H1 hESC. EGR1 is a pivotal regulator of cell fate and mitogenesis with critical roles in development and cancer [36]. The patterns of repressive H3K9me3 accumulation have been a focus in the cancer literature and have been proposed as a diagnostic marker in leukemia [37]. Similarly, the model for GM12878 (Epstein–Barr virus transformed lymphoblastoid) cells shows a disproportionate influence of ATF3 binding patterns, and ATF3 induction is a known consequence of virus-transformed cells [38]. Thus, the most cell-type-specific features in these models may be important indicators of cell-type-specific functions. These cell-type-specific features present a paradox, in view of the strong correlations in organization genome-wide across different cell types [8,9], and the demonstration that models trained in one cell type often perform well with data from other cell types. These contradictory observations are reconciled by the presence of inter-correlated clusters of features underlying A and B compartments. The

shifting membership of these clusters evidently retains enough similarity between cell types to enable the cross-application of models.

Chromatin boundaries, separating TADs and nuclear compartments at different scales, also showed cell-type-specific enrichments of various locus-level chromatin features. Across cell types, the complexity of boundary composition varies considerably so that only a few features were seen consistently enriched or depleted at boundaries. Peaks associated with active promoters were notable for both TAD and compartment boundaries in all cell types. Among the most influential variables for the random forest models constructed for the two hematopoietic cell lines was the ubiquitous transcription factor YY1, which reappeared in the analysis of chromatin boundary regions. Significant enrichments of YY1 were seen at TAD and nuclear compartment boundaries in all three cell types. Thus, the same protein was implicated at the level of broad genomic binding patterns (over 1-Mb intervals) and at the level of locally enriched peaks at boundary regions (spanning 100 to 500 kb). This is intriguing as YY1 has recently been shown to co-localize with the architectural protein CTCF [39] and suggests that these proteins cooperate in the establishment of domain boundaries. The identification of such features, significantly enriched at boundary regions, provides potential targets for deletion in experimental studies further exploring the structure and function of domains (for example, [14]). Both cell-type-specific and general constituents of boundaries may have utility in the biomedical interpretation of genomic variation in noncoding regions of the genome.

Conclusions

It has become commonplace to discuss the multi-layered, hierarchical organization of interphase chromosomes across strata ranging from nuclear compartments, down to the spectra of histone modifications and bound proteins at individual sub-genic regions. However, we lack a detailed understanding of how these strata interact. We have shown that our perspectives of features occurring at different strata can be bridged by modeling approaches, and the models produced can be used to explore the interrelationships between these different features quantitatively.

We constructed cell-type-specific models of nuclear organization, as reflected in Hi-C-derived eigenvector profiles, to discover the most influential features underlying higher-order structures. We found open and closed compartments to be well correlated with combinatorial patterns of histone modifications and DNA binding proteins, enabling accurate predictive models. These models could be cross-applied successfully between cell types highlighting constellations of common structural features associated with different nuclear compartments as expected. Dissection of the most influential variables also revealed important differences between models, consistent with the known biological contrasts among these cell types, such as the prominence of EGR1 in ESCs and H3K9me3 in the leukemia cell line. Investigation of regions showing variable nuclear organization across the three cell types under study, revealed enrichments for cell-type-specific enhancer activity, often nucleated at genes with known roles in cell-type-specific functions. Finally we used model predictions to examine boundary composition between higher-order domains across cell types. Among enrichments of a large number of factors observed at different boundaries in different cell types, CTCF and YY1 were found consistently and may cooperate to establish domain boundaries. In summary, we show that integrative modeling of large chromatin dataset collections using random forests can generate useful insights into chromosome structure and seed testable hypotheses for further experimental studies.

Materials and methods

Hi-C data and locus-level chromatin features

Hi-C datasets for human cell types H1 hESC [9], K562 [15] and GM12878 [40] were retrieved (Gene Expression Omnibus accession numbers: [GEO:GSE35156], [GEO:GSE18199] and [GEO:SRX030113]) and mapped to the genome (hg19/GRCh37). Iterative mapping was performed using the `hiclib` software package [41] and `bowtie2` [42] with the `very-sensitive` flag. Mapped reads were then binned into contact maps and iteratively corrected [41]. The `hiclib` software was also used for eigenvector expansion of each intrachromosomal contact map, performed independently for each chromosome arm.

Genome-wide ChIP-seq datasets for 22 DNA binding proteins (ATF3, CEBPB, CHD1, CHD2, CMYC, CTCF, EGR1, EZH2, GABP, JUND, MAX, MXI1, NRSF, POL2, P300, RAD21, SIX5, SP1, TAF1, TBP, YY1 and ZNF143) and ten histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3 and H4K20me1) were produced by ENCODE (July 2012 data freeze, used in [43,44]), in addition to DNase I hypersensitivity data and H2A.Z occupancy (Additional file 1: Figure S5), for each of the Tier 1 ENCODE cell lines used in this work: H1 hESC, K562 and GM12878 [4]. These data were processed using MACSv2 [45] to produce a fold-change signal relative to input chromatin and the data are available from [43]. Regional GC content was also calculated for each 1-Mb region and used in the feature modeling set (Additional file 3).

Structural modeling and variability

Random forest regression [46] was used as implemented in the R package `randomForest` [47]. Parameters of

$mtry = n/3 = 12$ and $ntrees = 200$ were assumed as the algorithm is known to be largely insensitive [48]. Variable importance within random forest regression models was measured using the mean decrease in accuracy in the out-of-bag sample. This represents the average difference (over the forest) between the accuracy of a tree with permuted and unpermuted versions of a given variable in units of percentage mean-squared error [49]. The effectiveness of the modeling approach was measured by four different metrics. Prediction accuracy was assessed by the PCC between the predicted and observed eigenvectors (out-of-bag estimate), and the root mean-squared error of the same data. Classification error, when predictions were thresholded into $A \geq 0$ and $B < 0$, was also calculated using accuracy (percentage correct classifications or true positives) and the area under the receiver operating characteristic (AUROC) curve. Together these give a comprehensive overview of model performance, both in terms of regression accuracy of the continuous eigenvector, and in how that same model could be used to label discrete chromatin compartments.

For cross-application of cell-type-specific models, a single random forest regression model was learned from all 1-Mb bins for a given cell type. This was then used to predict all bins from each of the other two cell types. The median absolute deviation was chosen as a robust measure of the variability in a given 1-Mb block between the three cell types. Blocks were ranked by this measure and the distribution was split into thirds that represented low variability (the third of blocks with the lowest median absolute deviation), and mid and high variability. Each subgroup was then independently modeled using the random forest approach described above. For each cell type we identified 1-Mb regions whose compartment state was altered relative to the other two. For example, if a 1-Mb bin was classified as occupying compartment A in H1 hESC and B in both K562 and GM12878, it is said to occupy an altered open compartment in H1 hESC. Chromatin state annotations were calculated from ENCODE `ChromHMM/SegWay` combined annotations for each cell type [5]. Annotated features were considered shared if there was an overlapping annotation in either of the two other cell types, and labeled as specific to a cell type otherwise.

Chromatin boundaries

TAD boundaries were called using software provided by Dixon et al. [9] with recommended parameters. For the generation of locus-level feature profiles over TAD boundaries, input features were averaged into 40-kb bins spanning ±500 kb from the boundary center. For compartment boundaries, a two-state hidden Markov model was trained on the compartment eigenvector data and the Viterbi algorithm was used to infer the most likely underlying state sequence that generated the observed compartment eigenvectors. Compartment boundaries were then defined as the point of transition between different compartment types. To generate boundary profiles, locus-level features were averaged into 100-kb windows extending ±1.5 Mb either side of the boundary center.

To test for the enrichment or depletion of a chromatin feature over a given boundary, a two-tailed Mann–Whitney test was used to compare the boundary bin with the ten outermost bins of the window (five from either side). The significance level at $\alpha = 0.01$ was then Bonferroni-adjusted for multiple testing correction, and results with P values exceeding this threshold were deemed significantly enriched or depleted at a given boundary.

Scripts to reproduce the analyses and generate manuscripts figures are available at [50].

Additional files

Additional file 1: Figures S1 to S15. Collection of supplementary figures (S1 to S15) with captions.

Additional file 2: Tables S1 to S3. Functional enrichments of genes located within structurally variable regions in each cell type.

Additional file 3: cellTypeFeatureSets. Archive containing comma-separated value (CSV) files of binned input features and compartment eigenvectors used for modeling, for each of the three cell types used in this study.

Abbreviations

AUROC: Area under the receiver operating characteristic curve; ChIP-seq: Chromatin immunoprecipitation sequencing; ESC: Embryonic stem cell; kb: kilobases; Mb: megabases; PCC: Pearson correlation coefficient; PcG: polycomb-associated; TAD: Topological domain.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BLM carried out the analysis and helped draft the manuscript. CAS and SA conceived of the study, participated in its design and coordination and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are indebted to the ENCODE Consortium for timely and comprehensive access to its data. We are grateful to Anshul Kundaje, Stanford University, for advice on using these data. We thank the UK Medical Research Council for financial support.

Received: 9 September 2014 Accepted: 24 April 2015
Published online: 27 May 2015

References

1. Bickmore Wa, van Steensel B. Genome architecture: domain organization of interphase chromosomes. Cell. 2013;152:1270–84. doi:10.1016/j.cell.2013.02.001.
2. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–16. doi:10.1038/nmeth.1906.
3. Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, et al. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell. 2011;147:1628–39. doi:10.1016/j.cell.2011.09.057.

4. ENCODE. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74. doi:10.1038/nature11247.
5. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41:827–41. doi:10.1093/nar/gks1284.
6. Dekker J, Marti-Renom Ma, Mirny La. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14:390–403. doi:10.1038/nrg3454.
7. de Wit E, Bouwman BA, Zhu Y, Klous P, Splinter E, Verstegen MJ, et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature. 2013;501:227–31. doi:10.1038/nature12420.
8. Chambers EV, Bickmore WA, Semple CA. Divergence of mammalian higher order chromatin structure is associated with developmental loci. PLoS Comput Biol. 2013;9:1003017. doi:10.1371/journal.pcbi.1003017.
9. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80. doi:10.1038/nature11082.
10. Meuleman W, Peric-Hupkes D, Kind J, Beaudry JB, Pagie L, Kellis M, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 2013;23:270–80. doi:10.1101/gr.141028.112.
11. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, et al. Genome-wide dynamics of replication timing revealed by *in vitro* models of mouse embryogenesis. Genome Res. 2010;20:155–69. doi:10.1101/gr.099796.109.
12. Liang G, Zhang Y. Embryonic stem cell and induced pluripotent stem cell: an epigenetic perspective. Cell Res. 2013;23:49–69. doi:10.1038/cr.2012.175.
13. Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RWW, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci USA. 2014;111:996–1001. doi:10.1073/pnas.1317788111.
14. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485:381–5. doi:10.1038/nature11049.
15. 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;326:289–93. doi:10.1126/science.1181369.
16. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13:53. doi:10.1186/gb-2012-13-9-r53.
17. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's Guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75. doi:10.1016/j.ymeth.2014.10.031.
18. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell. 2012;151:68–79. doi:10.1016/j.cell.2012.08.033.
19. Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell. 2013;155:1507–20. doi:10.1016/j.cell.2013.11.039.
20. Zervos AS, Gyuris J, Brent R. Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. Cell. 1993;72:223–32. doi:10.1016/0092-8674(93)90662-A.
21. Wold S, Ruhe A, Wold H, Dunn III WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput. 1984;5:735–43. doi:10.1137/0905052.
22. 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.
23. Nechanitzky R, Akbas D, Scherer S, Györ I, Hoyler T, Ramamoorthy S, et al. Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. Nat Immunol. 2013;14:867–75. doi:10.1038/ni.2641.
24. Mansson R, Welinder E, Åhsberg J, Lin YC, Benner C, Glass CK, et al. Positive intergenic feedback circuitry, involving EBF1 and FOXO1, orchestrates B-cell fate. Proc Natl Acad Sci USA. 2012;109:21028–33. doi:10.1073/pnas.1211427109.
25. Pohl E, Aykut A, Beleggia F, Karaca E, Durmaz B, Keupp K, et al. A hypofunctional PAX1 mutation causes autosomal recessively inherited otofaciocervical syndrome. Hum Genet. 2013;132:1311–20. doi:10.1007/s00439-013-1337-9.
26. Svensson EC, Tufts RL, Polk CE, Leiden JM. Molecular cloning of FOG-2: a modulator of transcription factor GATA-4 in cardiomyocytes. Proc Natl Acad Sci USA. 1999;96:956–61.
27. Evertts AG, Manning AL, Wang X, Dyson NJ, Garcia BA, Coller HA, et al. H4K20 methylation regulates quiescence and chromatin compaction. Mol Biol Cell. 2013;24:3025–7. doi:10.1091/mbc.E12-07-0529.
28. Atchison ML. Function of YY1 in long-distance DNA interactions. Front Immunol. 2014;5:45. doi:10.3389/fimmu.2014.00045.
29. Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, et al. Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. Genome Biol. 2013;14:148. doi:10.1186/gb-2013-14-12-r148.
30. Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, et al. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. 2013;23:2066–77. doi:10.1101/gr.161620.113.
31. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153:1281–95. doi:10.1016/j.cell.2013.04.053.
32. Onder TT, Kara N, Cherry A, Sinha AU, Zhu N, Bernt KM, et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature. 2012;483:598–602. doi:10.1038/nature10953.
33. Chai X, Nagarajan S, Kim K, Lee K, Choi JK. Regulation of the boundaries of accessible chromatin. PLoS Genet. 2013;9:1003778. doi:10.1371/journal.pgen.1003778.
34. Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H, et al. H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. Genome Biol. 2012;13:85. doi:10.1186/gb-2012-13-10-r85.
35. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell. 2012;148:458–72. doi:10.1016/j.cell.2012.01.010.
36. Zwang Y, Oren M, Yarden Y. Consistency test of the cell cycle: roles for p53 and EGR1. Cancer Res. 2012;72:1051–4. doi:10.1158/0008-5472.CAN-11-3382.
37. Müller-Tidow C, Klein HU, Hascher A, Isken F, Tickenbrock L, Thoennissen N, et al. Profiling of histone H3 lysine 9 trimethylation levels predicts transcription factor activity and survival in acute myeloid leukemia. Blood. 2010;116:3564–71. doi:10.1182/blood-2009-09-240978.
38. Hagmeyer BM, Duyndam MC, Angel P, de Groot RP, Verlaan M, Elfferich P, et al. Altered AP-1/ATF complexes in adenovirus-E1-transformed cells due to EIA-dependent induction of ATF3. Oncogene. 1996;12:1025–32.
39. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014;15:234–46. doi:10.1038/nrg3663.
40. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2012;30:90–8. doi:10.1038/nbt.2057.
41. 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.
42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. doi:10.1038/nmeth.1923.
43. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, et al. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512:453–6. doi:10.1038/nature13668. https://www.encodeproject.org/comparative/regulation/#Humanset9.
44. Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, et al. Comparative analysis of metazoan chromatin organization. Nature. 2014;512:449–52. doi:10.1038/nature13415.
45. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:137. doi:10.1186/gb-2008-9-9-r137.
46. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
47. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18–22.

48. Hastie T. Kernel smoothing methods. In: Elements of Statistical Learning. 2nd. Springer-Verlag; 2009. doi:10.1007/b94608_6.
49. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology. 2007;88:2783–92.
50. Moore BL. 3dgenome (release v0.1.0). Github. https://github.com/blmoore/3dgenome.

REFERENCES