Unravelling higher order chromatin organisation through statistical analysis

Benjamin L. Moore

July 28, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

MRC Medical Research Council

THE UNIVERSITY OF EDINBURGH

CANCER RESEARCH UK

DECLARATION

This thesis presents my own work, wherever the contributions of others were involved this is clearly indicated.

—Benjamin L. Moore (2015)

ACKNOWLEDGEMENTS

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 Hi-C, 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 multi-megabase 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 lamin-associated domains, as well as exploring the concept of super-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.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3C	Chromsome 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	Flurorescent *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
MAD	Median absolute deviation
MSE	Mean squared error
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
TAD	Topologically-associating domains
TSS	Transcription start site

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 5.4 are used in a submitted manuscript:

> 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 reorganization underlies transcriptional changes in cellular differentiation. *Submitted*

1 INTRODUCTION

1.1 GENOME ORGANISATION

1.1.1 C-methods and Hi-C

Classical studies of chromosome conformation relied on microscopy techniques to visualise nuclear architecture, most commonly fluorescence *in situ* hybridisation (FISH). These techniques led to the discovery of "chromosome territories", regions of the nucleus wherein distinct chromosomes were thought to occupy, and more broadly identified the non-random arrangement of loci in three-dimensional space.[1,2] 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.[1]

With the advent DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[3] was the first sequencing-based method of measuring chromosome conformation. The method uses formaldehyde to cross-link nuclear proteins in place, trapping genomic regions that were physically co-located through bound proteins, then to apply a frequent restriction enzyme to shear the sample into fragments. Next, under dilute conditions, DNA fragments are ligated together. The dilute conditions favour ligations between fixed fragments, with the aim of generating hybrid fragments form two genomic regions which were close together in the original preparation. Cross-linking can then be reversed and, in the case of the original 3C method, measured by quantitative PCR using pre-designed primers for your fragments of interest. The end result is a relative measure of interaction frequency between any two regions of interest, in theory directly proportional to their distance in three-dimensional space.

The rapid advancement of sequencing, allowed the original 3C method to be further developed, first through microarray technology, then using high-throughput sequencing. Two protocols were proposed for a 3C-inspired one-to-many assay[4,5] (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.[6]

The final step was an all-versus-all assay, capable of assaying pairwise interaction frequencies between all restriction fragments of a genome. This assay was published by Lieberman Aiden *et al.*[7] and named Hi-C (Fig. 1). The Hi-C method added biotin tagging to pull-down only ligated fragments for sequencing. At the time of publication, resolution of Hi-C data for analysis was limited by sequencing depth, given the huge number of restriction fragments produced by a 6-cutter enzyme (HindIII and NcoI were used in[7]) but the falling costs of sequencing and proven utility of the assay meant subsequent Hi-C papers incrementally increased their sequencing depth, to a point where analysis could be performed at the level of individual restriction fragments, genome-wide.[8–11]

1.1.2 Hi-C variants

The interaction maps produced by Hi-C were found to exhibit several inherent biases. Fragment properties, such as their length, GC content and mappability, were confounding interaction frequency estimates and there-

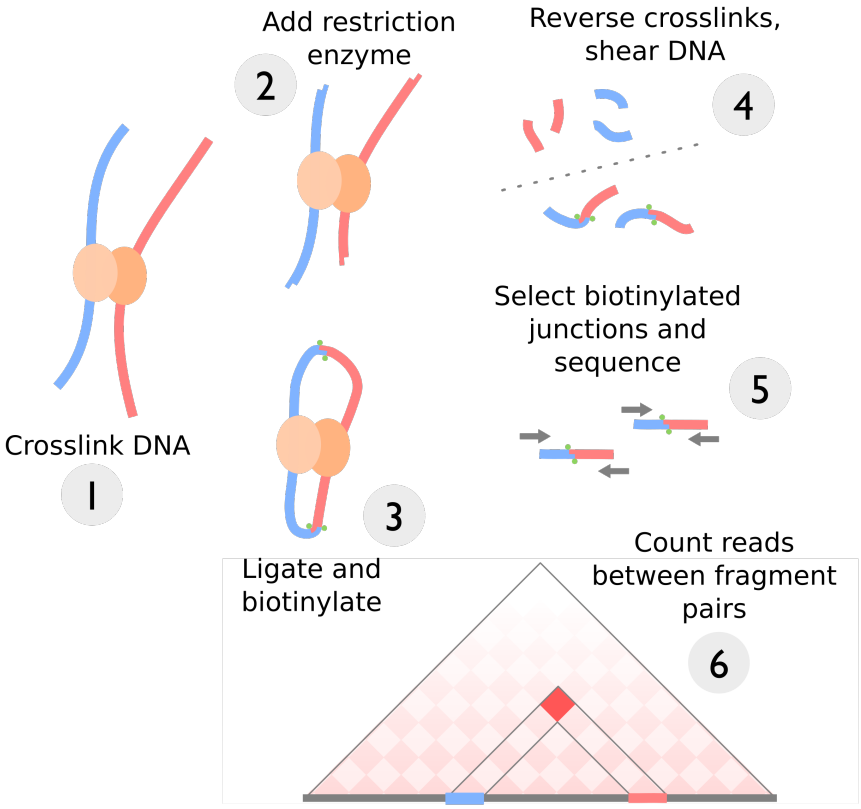

Figure 1: Steps in the Hi-C assay. Schematic of the Hi-C experimental procedure as described in Lieberman Aiden *et al.*[7]

fore needed to be normalised-away before subsequent analysis.[12,13] A range of statistical techniques were developed to correct for these latent variables,[14–17] while experimentalists instead looked to improve on the experimental procedure itself.

Tethered chromosome capture (TCC)[18] 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.*[18] reported a large decrease in observed interchromosomal contacts in their tethered library, suggesting many of those originally observed were caused by spurious ligation of non-crosslinked fragments.

Hi-C is a population-level assay, as the retrieved interaction counts are from a huge number of different cells. As well as building population-averaged models of genome structure, it is also of interest to probe cell-to-cell variability through single-cell approaches. For instance, it's been estimated that long-range contacts identified with C-methods may occur in as few as 10% of cells at any one time.[2]

In the first single-cell Hi-C study, Nagano *et al.*[19] aimed to explore this cell-to-cell variability by performing the Hi-C assay on single, hand-selected nuclei. An obvious limitation this Hi-C variant is that a single restriction fragment can ligate to at most one other fragment, meaning even if 100% yield were to be achieved, any $n \times n$ restriction fragment interaction matrix could at most populate $\frac{n}{2}$ cells; in practice, the realised yield of this first single cell Hi-C experiment was just 2.5%.[19] Nevertheless, single-cell Hi-C was able to reproduce findings from population-based (or "ensemble") Hi-C, such as preferential interactions between active domains, but also was able to dissect *trans* interactions, suggesting high cell-to-cell variability leads to their relatively uniform appearance in normal Hi-C interaction maps.[19] Combined with observations from TCC which gave evidence that

interchromosomal contacts were disproportionately the result of spurious ligation,[18] the functional significance of these *trans* interactions seems at best unclear in the general case.

Capture-C is another recent Hi-C derivative which attempts to address resolution problems associated with the genome-wide pairwise assay by enriching for promoter-enhancer interactions using *a priori* selection.[20] It could be said that Capture-C is to Hi-C as exome-capture sequencing is to a whole-genome approach. Indeed, a suggestion in the original Hi-C paper was that resolution could be improved by either increased sequencing or using hybrid capture.[7]

Use of a cell population also averages away cell-cycle effects, with the vast majority of results coming from cells during interphase (around 97%[21]). Naumova *et al.*[21] looked to assay chromosome conformation specifically over different cell cycle stages, to better understand chromosome compaction during mitosis.

In-site Hi-C was a recent refinement of the Hi-C method, from the published of the original method.[11] The principle difference is that fixation and ligation now happen in place, within intact cell nuclei.

1.1.3 Chromosome compartments

In the paper describing the Hi-C technique,[7] Lieberman-Aiden *et al.* described low-resolution structures they name "A" and "B" nuclear compartments. These are regions with a median size of around 5 megabases which showed properties typical of euchromatin and heterochromatin, respectively. A compartments were observed through 3D-FISH to be centrally-positioned in the nucleus and ChIP-seq data showed several hallmarks of transcriptional activity. B compartments, conversely, were heterochromatic and lamina-associated regions, with little transcription and repressive histone modifications such as H3k9me3.[1,7] As expected from positioning data, the co-location of compartment types is also visible in their contact maps.

These compartments were identified through a continuous eigenvector profile, derived from a normalised Hi-C contact matrix.[7] Importantly, this measure holds more information than a simple two-state classification, rather the continuous values can be interpreted as relative levels of compaction or activity.[14,15]

1.1.4 Topological domains

The falling cost of high-throughput sequencing enabled increasingly deep sequencing of Hi-C experiments. Sequencing is the main resolution-limiting resource for this assay, as to increase the analysis resolution and maintain the level of coverage requires an exponential increase in the total amount of sequencing required.[7,22]

In experiments totalling around two billion total sequencing reads, Dixon *et al.*[8] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. The authors noticed smaller domains they designated "topological associative domains" (or TADs) which were observable as self-interacting, off-diagonal blocks of higher-than-expected self-interaction frequency. They defined a domain calling algorithm based on the directional bias of a genomic region's contacts, and used a Hidden Markov Model to infer blocks of strongly up- or downstream-biased, reasoning that domain boundaries are present when a strongly upstream biased region is adjacent to a region of opposite bias (Fig. 3). These boundaries themselves were investigated and were found to display suggestive functional enrichments for DNA binding proteins including CTCF, long thought to act as an insulator of chromatin state.

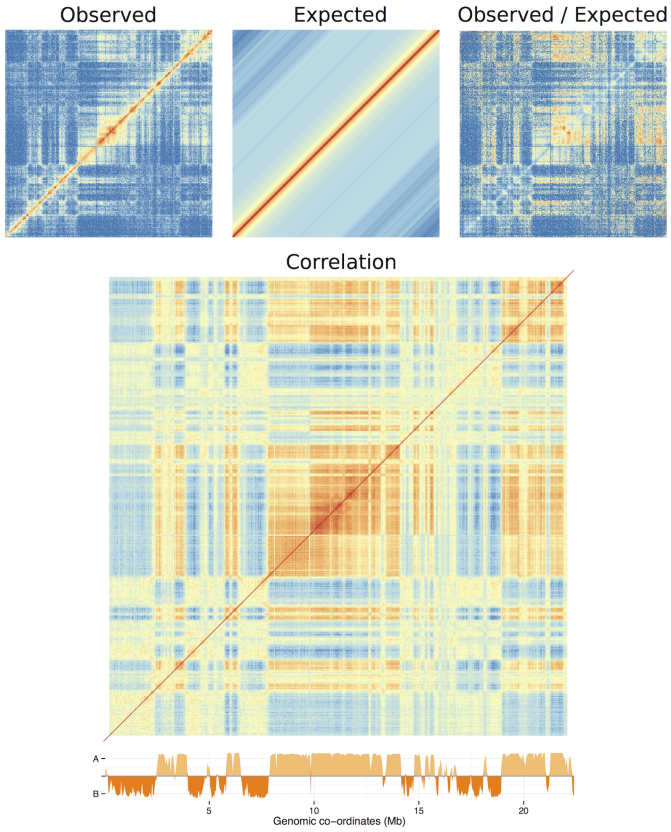

Figure 2: Derivation of A/B compartment profile from Hi-C data. Observed interaction frequencies (O) are averaged along super-diagonals to give a distance-normalised expected matrix (E). The Pearson correlation of the O/E matrix then can undergo eigenvector expansion; in most cases eigenvector $\mathbf{v}$ with the largest eigenvalue, λ, then reflects A/B compartmentalisation.[7]

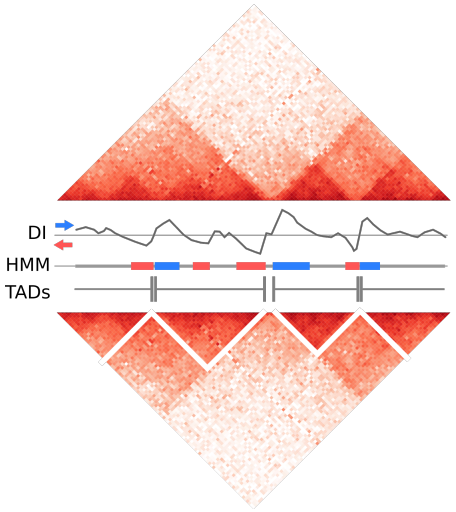

Figure 3: Dixon *et al.* pipeline for calling topological associating domains (TADs). First a directionality index (DI) is calculated for each bin based on the ratio of upstream:downstream contacts. Secondly a Hidden Markov Model (HMM) is used to infer the most likely state sequence that emitted the DI variable. Finally a simple rule is applied whereby a run of high-confidence upstream-biased state calls marks the end of a domain. New domains begin with any subsequent downstream-biased state. Gaps between TAD calls can be observed, and as labelled border regions up to a size threshold of 400 kb, whereafter those regions are unclassified.[8]

Dixon *et al.*[8] also performed some comparative analysis, reporting large and significant overlap of domain boundary positions both within species and between human and mouse cell lines.

1.1.5 Other proposed structures

Filippova *et al.*[23] developed a tuneable algorithm which identifies "alternative topological domains".

A study of *Drosophila* embryonic chromosomes found a similarly hierarchical organisation of physical domains, and also was able to relate these to "epigenomics domains" showing specific sets of enrichment signatures representing active, null, polycomb-associated and telomeric regions.[24]

Recent high-resolution studies have been able to resolve ever-smaller levels of sub-structure. Rao *et al.*[11] 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.[11] The authors also suggest that previously-observed large TADs may be the result of insufficient sequencing; that is, not all boundaries could be detected using 40 kb binned contact maps thus multiple contact domains were unintentionally combined into large domains.

1.2 MODELS OF CHROMATIN FOLDING

Theoretical mechanistic models of chromatin folding such as the "strings and binders switch" model[25] and the "fractal globule" model[7,26,27] have both produced simulated data that reflects empirical C-method observations and potentially describe the polymer dynamics of chromatin folding.

1.2.1 Fractal globule

Lieberman Aiden *et al.*[7] tested a number of theoretical models of genome folding to see which best explained the observed power-law scaling between distance and observed contact frequency ($IF = 1/dist^{-\alpha}$ where $\alpha \approx 1.08$). The authors sought to distinguish two previously-described models of genome organisation: the "fractal globule" and "equilibrium globule". The authors found that a theoretical fractal globule, embodying scale-independent self-similar aggregate folding, better fit the observed data than an equilibrium globule null model where simulated polymer folding was allowed to proceed unchecked.

The fractal globule model was noted for its appealing functional properties. Under this model, for example, the polymer folds are knot-free hence could facilitate local dynamics of repression and activation without wider disruption. Despite this appeal, the authors were careful to state that while their simulations show good agreement with observed data, this does not preclude other organisational models from having similar or greater explanatory power.[7]

1.2.2 Strings and binders switch

Subsequent modelling techniques integrated known biological phenomena as well as polymer models. This formed the basis of Barbieri *et al.*'s[25] "strings and binders switch" (SBS) model, where the authors simulated polymer folding in the presence of DNA binding factors, such as the known genome organiser CCCTC-binding factor (CTCF).[28] This organisational model was developed in an attempt to consolidate global Hi-C measures of contact scaling with C-based experiments on smaller regions and FISH studies,

which found a range of scaling parameters. The authors also explore the different values of α between cell lines and even chromosomes, and find that their mechanistic model can explain each case using variable concentrations of binders which causes phase-switching between open and compacted chromatin, with fractal globule existing at the phase transition boundary.

This model offers broad explanatory power for a range of observed power law coefficients (α) and from simple underpinnings, but critics point out that simulations were performed on a polymer composed of just 500 monomers.

1.2.3 Looping

1.2.4 Cell cycle changes

Chromosome structure has been assayed both through mitosis[21] and Studies have also focused on the edge-case of chromatin structures on X-chromosomes.

1.3 CRITICISMS OF C-METHODS

The resolution of a Hi-C experiment has a hard-limit imposed by the choice of restriction enzyme. For example, the commonly-used HindIII enzyme is a six-cutter that recognises the motif AAGCTT and cuts approximately every 4 kb, on average.[1] More recent studies have switched to a four-cutter restriction enzyme, for example MboI,[11] which increases this upper-bound on resolution to the order of hundreds of basepairs (i.e. naively, $4^4 = 256$bp fragments, on average). A downside of using more frequent restriction enzymes is the potential side-effect of promoting more non-specific ligations by increasing the concentration of fragments in solution.[11]

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",[29] which is then assumed to be a proxy for physical distance within the nucleus. This is a marked difference from aforementioned FISH methods, where the physical distance is observed directly, albeit through the addition of non-native probes. So strong is this assumption, that methods have been developed that use a known FISH distance to then calibrate genome-wide Hi-C distances,[30] yet it remains unclear to what extent these two methods are compatible.

When interpreting C-methods data it should also be kept in mind that even verifiable contacts are by no-means functional. To elaborate, C-methods may find two regions to be strongly co-localised, but an understanding of the region may explain their co-localisation to be caused by mutual interaction with a nuclear lamina or nucleolus, for example, rather than any specific functional relationship between the two loci.[15] In addition, a functional enhancer–promoter interaction will necessarily constrain the contacts of other nearby regions, potentially causing highly-reproducible "bystander interactions"[15] that are nevertheless uninteresting form a functional perspective.

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,[31] 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.[32] Problems in the field such as gene prediction and inferring regulatory networks can be approached by employing a learning algorithm, either in a supervised

way based on a known truth set, or through unsupervised methods aimed at pattern detection or clustering. If a successful predictive model can be built, it can then be dissected to explore statistical rules which may impart novel biological insight. As a toy example, learning a highly-accurate model of enhancer prediction could itself identify novel epigenetic marks indicative of enhancers, generating testable hypotheses about how enhancers are activated.

The link between epigenomic features and local chromatin state has been analysed computationally in a number of publications, notably in developing the Hidden Markov Model-based ChromHMM[33] algorithm which predicts states such as active promoters and enhancers, using a range of histone marks and other underlying features.[34] Similarly a Random Forest-based algorithm was developed to predict enhancers from histone modification data.[35] However few studies have spanned all of these levels of chromatin structure and nuclear organisation, and it is not yet known how locus-level chromatin features may be related to higher order genome organisation.

1.4.1 ENCODE

The recent comprehensive ChIP-seq datasets produced by the ENCODE consortium[36] combined with Hi-C genome-wide contact maps in a number of human cell types[7,8,18] present a remarkable opportunity to investigate the relationships between local chromatin features and higher order structure. In this work, a machine-learning approach was employed to model the compartmental characteristics of large genomic regions based on their aggregate levels of various histone marks and DNA binding proteins. Dissection of the resulting models was then used as a means of gleaning biological insights into the basis of higher order structure and of highlighting important differences between cell types.

1.5 AIMS

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

2 METHODS

2.1 HI-C DATA

2.1.1 Mapping

Raw Hi-C reads were downloaded from published datasets (Table 1) through the Gene Expression Omnibus (GEO)[37] or the Short Read Archive (SRA)[38] with identifiers: GSE35156 (H1 hESC), GSE18199 (K562) and SRX030113 (GM12878). These paired reads were mapped independently to a reference genome: hg19/GRCh37 for human data, and mm10/GRCm38 for mouse.

Mapping was performed using the `hiclib` software package[14] and `bowtie2`[39] with the `--very-sensitive` flag. An iterative mapping approach was used to maximise the number of aligning fragments.[14] 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.

2.1.2 Filtering

After mapping, interactions are first aggregated into restriction fragments then by regular binning of various resolutions (particularly 40 kb, 100 kb and 1 Mb). Several filters were applied at this stage, with the following cases removed:[14]

- 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
- Extremely over-represented fragments (top .05%) which may throw-off eigenvector derivation

2.1.3 Correction

Iterative correction and eigenvector expansion (ICE) is an approach to normalisation and processing Hi-C data, implemented as software library written in `python`.[14] 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 $\mathbf{A}$ 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[12] but without explicitly modelling

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

Cell line	Total reads	Accession	Citation
Gm12878	31×10^6	SRX030113	18
H1 hESC	331×10^6	GSE35156	8
K562	36×10^6	GSE18199	7
Cortex	373×10^6	GSE35156	8
mESC	476×10^6	GSE35156	8
IMR90	355×10^6	GSE35156	8

these latent variables. This procedure is thus converting 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 1). Despite differences in the levels of sequencing, otherwise the experiment 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 (that is, we did not consider data from Hi-C variants such as TCC[18] and *in-situ* Hi-C[11]).

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.*,[7] as opposed to the related eigenvectors calculated in Imakaev *et al.*[14] from the corrected maps alone. The details of this procedure are described in section 2.5.2. 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.

Importantly, 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.[7,14] Eigenvector expansion was performed on both 1 Mb and 100 kb matrices, below these resolutions results became less stable, and it has been shown that eigenvectors at

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[36,40] 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[41] to produce fold-change relative to input chromatin. GC content was also calculated and used in the featureset to give 35 total inputs (Table 2).

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

2.2.1 Clustering input features

To quantify collinearity of input features, correlation matrices built from genome-wide vectors of input feature measures were build and hierarchicaly

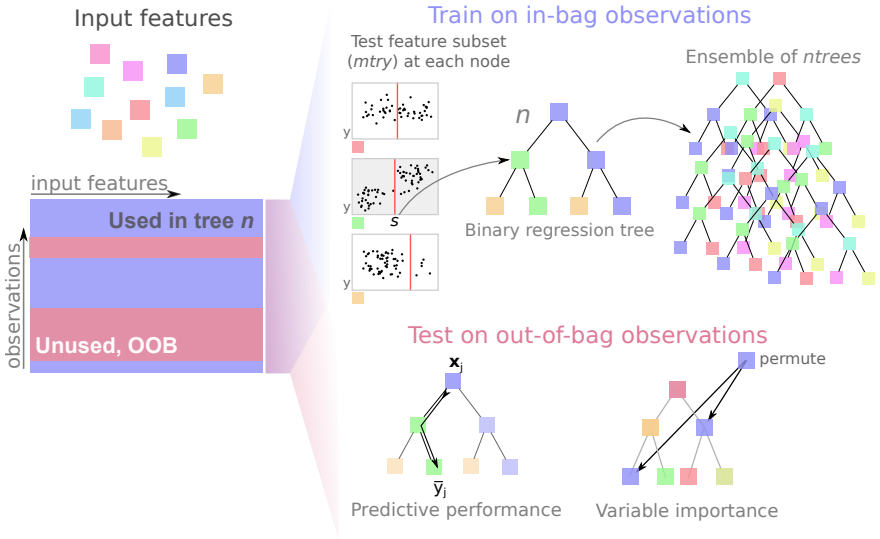

Figure 4: Random Forests overview. Random Forests are an ensemble of bagged, de-correlated classification or regression trees first described by Breiman.[43]

clustered. The "significance" of observed clustering was assessed using sub- and super-sampled bootstrapping, with stable clusters deemed significant, as implemented in the `pvclust` R package.[42]

2.3 MODELLING

2.3.1 Random Forest

Random Forest (RF) regression,[43] was used as implemented in the R package `randomForest`.[44] The RF algorithm (Fig. 4) 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 a specific 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 value across the forest. RF regression was used as it is known to be one of the most powerful regression methods developed to date,[45,46] typically providing low bias and low variance predictions without the need for variable selection.[47,48]

Additionally the RF method represents an example of "algorithmic modelling"[49] in that it makes no assumptions about the underlying data model. Parameters of $mtry = \frac{n}{3}$ (where n is the number of input features) and $ntrees = 200$ were assumed as they are known to be largely insensitive;[48,50] this was verified with the dataset used in this work (Fig. 5).

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 unpermuted versions of a given variable (Fig. 4), in units of mean squared error (MSE).[46,48]

2.3.2 Model performance

The effectiveness of the modelling approach 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,

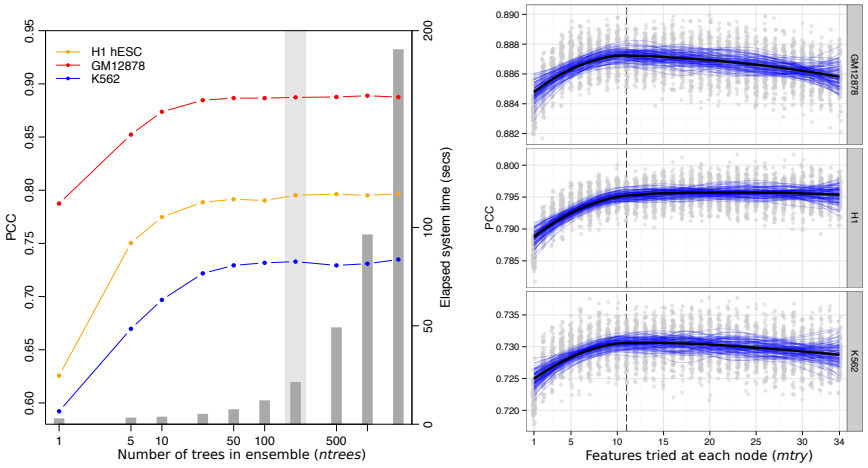

Figure 5: Random Forest parameters are largely insensitive. Two user-facing Random Forest paramters are known to be insensitive over a broad range.[50] Optimisations for *ntrees* (the number of trees in the forests) and *mtry* (the number of features tested at each node) are shown for three different models, with typical values of 200 trees and $\frac{1}{3}$ of input variables highlighted.

when predictions where 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 all bins from each of the other two cell types.

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.

2.3.3 Other modelling approaches

Linear regression was used as a baseline for comparison with more complicated approaches such as Random Forest. If the same modelling accuracy could be achieved with simple multiple linear regression, this would be a faster and more interpretable modelling framework.

Partial least squares (PLS) regression was also used to model compartment profiles. 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.[51] 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, K562 and GM12878. 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.

"Flipped" regions are those 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 a "flipped" compartment (to open).

2.4.2 Enhancer enrichment

Chromatin state annotations used in this work were retrieved from the ChromHMM[34] and SegWay[52] combined annotations.[53] 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, Hoffman *et al.* caution that in areas of disagreement, each algorithm may highlight differing biological phenomena so should also be considered separately.[53]

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

Variable regions (section 2.4.1) were tested for functional enrichments using Gene Ontology (GO) annotations.[54] The DAVID tool[55] was used to compare GO terms for genes located in variable compartments with a background set of genes within all annotated compartments.

2.5 BOUNDARIES

2.5.1 TADs

TAD boundaries were called using the software provided in Dixon *et al.*[8] using their recommended parameters. For the generation of boundary profiles, input features were averaged into 40 kb bins spanning $\pm$450 kb from the boundary bin.

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 (5 from either side). 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.

To compare boundaries between cells, each TAD boundary called in K562 and GM12878 were 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. A Kolmogorov-Smirnov test was then used to compare the empirical cumulative distributions of these distances.

2.5.2 Compartments

Eigenvectors were calculated as described in section 2.1.4. 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, more repressive state.[7]

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. The point at which transitions occurred between states was taken as a 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, described in section 2.5.1.

2.5.3 MetaTADs

MetaTADs are a concept discovered by collaborators. Their method for calling such features involve the constrained hierarchical cluttering of neighbouring TADs with the greatest inter-TAD contacts. This results in a tree of increasing metaTAD aggregation. For boundary analysis of metaTADs, again a similar approach was used to that of TADs (section 2.5.1) but thresholded to within a given range of sizes. MetaTADs 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 power to analyse any differences. 10 Mb was chosen in an attempt to compromise minimising the overlap between TAD and metaTAD boundaries, while also retaining a large enough sample size. 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. 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.

2.6 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.[56]

To compare G-band boundaries with our compartment data, we allowed for a ± 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, ...

2.7 NUCLEAR POSITIONING

Previously published data on chromosome positioning preference within the nucleus was used to label each chromosome as "inner", "middle" or "outer".[57] Chromosomes whose DAPI hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by Boyle *et al.*[57], 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. The remaining chromosomes in our filtered dataset, 6, 14 and 15, were assigned to the "middle" group and showed no significant to either inner or outer nuclear shells ($p \geq 0.1$).[57] The significance of the difference in distribution of eigenvectors in the inner versus outer shell was determined by a one-sided Kolmogorov-Smirnov (K-S) test, with the alternative hypothesis that the empirical cumulative density function of the inner chromosome eigenvectors F_{inner} is greater-than or equal-to F_{outer}. This chromosomal positioning data was measured in lymphoblastoid cells though nuclear architecture is though to be largely conserved between cell types[58,59] and even higher primates.[60]

2.8 4C ANALYSIS

The experimental protocol used by our collaborators to generate 3C-seq data (also known as 4C) recommends the `r3Cseq` R package,[61,62] part of the bioconductor repository[63,64] for the R programming environment.[65]

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

2.8.1 Normalisation

The normalisation procedure is adapted from a previous method for normalising deepCAGE data between samples.[66] In short, the reverse-cumulative distribution of read counts per restriction fragment is fit 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 the empirical coefficient, $\alpha = -1.35$.[62]

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 reduces the impact of artefacts and errors by enforcing the expected power-law relationship of restriction fragment read counts.

2.8.2 Significance estimation

The `r3Cseq` package[62] also attempts to assign a measure of 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 expected to be significantly enriched, thus won't unduly perturb an average signal.[62] An improved method that avoids this assumption has since been developed where a background model was iteratively fitted, with outlier removal at each revision.[67]

Here a non-parametric cubic smooth spline is fit 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:

$$Z = \frac{(O - E)}{\sigma} \tag{1}$$

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[68] (as implemented in the `qvalue` R package[69]). 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.[5,70–73]).

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) terms in eqn. 1 are genome-wide averages excluding regions ± 100 kb around the viewpoint.[62] 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.

2.8.3 3–D modelling

2.9 5C ANALYSIS

2.10 SCRIPTS AND OTHER ANALYSES

Much of this work has been performed by writing custom scripts in the R programming language.[65] 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`[74] and `dplyr`[75].

The programming language python[76] was also employed to a lesser-extent, as were command–line tools such as `bedtools`[77] and `SAMtools`[78]. Additionally command-line `BigWig*` tools[79] were used, as well as the UCSC genome browser associated data tracks.[80–82].

3 REANALYSIS OF HI-C DATASETS

3.1 INTRODUCTION

Since the initial publication of the Hi-C technique in 2009,[7] there has been rapid advancement of both the technique itself and the resolution at which interaction frequencies have been analysed. From the proof-of-concept analysis at 1 megabase (Mb) and 100 kilobase (kb) resolution,[7] subsequent experiments achieved first 40 kb[8], then 10 kb[10] and most recently 1 kb,[11] 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 within and between species.

In this work I uniformly reprocessed publicly-available human Hi-C datasets, in order to address fundamental questions about the stability of higher order genome organisation within cell populations from the same species. Previously Hi-C studies have compared two samples per species, such as K562 against GM06990[7] or IMR90 against GM12878.[8] Here I 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,[36] hence have huge amounts of matched ChIP-seq and histone modification data available.

By combinatorial reanalysis of these cell-matched datasets, I can investigate the 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 using the same pipeline from raw sequencing reads (Methods 2.1). Briefly, raw sequencing reads were sourced from three different publications (Lieberman-Aiden *et al.*[7], Dixon *et al.*[8] and Kalhor *et al.*[18]). 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 2.1.1).

Next a filtering step was applied, which removed those fragment pairs that were likely artifactual or erroneous (Methods **??**). A correction step was then applied, whereby biases such as mappability and GC content were removed to give each fragment equal visibility (Methods 2.1.3). Overall these steps produced comparable maps of interaction frequency in different cell types, despite their differing origins (Fig. 6).

Figure 6 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. 6).

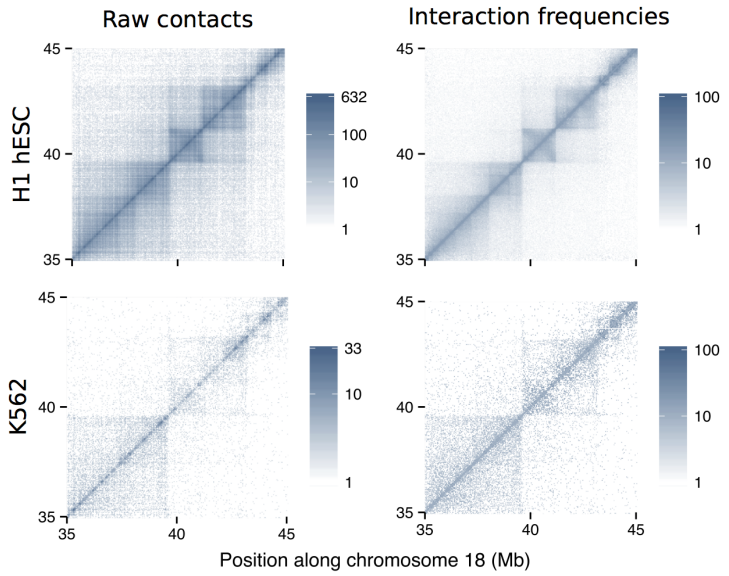

Figure 6: Iterative correction converts raw counts to normalised interaction frequencies. The sample with highest sequencing depth (H1 hESC) is shown alongside that with much lower sequencing (K562) both before and after iterative correction and normalisation procedures were applied (Methods 2.1) at 40 kb resolution for a 10 Mb section of human chromosome 18. Fill gradients are on a $\log_{10}$ scale.

3.3 COMPARTMENT PROFILES

After uniformly reprocessing each Hi-C dataset and calling compartment eigenvector profiles (see *Methods*), we can compare these between three human cell lines. Compartment profiles have a visibly high-correspondence (Fig. 7), despite the variable sources of both sample material and experimental data.

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 Pearson correlation coefficients between these independent measures are in the interval $R = [.75, .8]$ (Fig. 8).

3.4 DOMAIN CALLS

3.4.1 Compartments

The continuous compartment eigenvector can be used as-is to classify A/B compartments, using positive and negative eigenvector values after first orientating the vector with respect to, for example, PolII Chip-seq data.[18] However, given the definition of compartments as generally broad and alternating domains along a chromosome, often matching other large domains of Lamin association, an improved classification method might penalise the calls of 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. This can be modelled through a Hidden Markov Model (HMM), whereby we first parameterise models of state and their transitions, then infer the most likely state sequence to have emitted our observed data. This unobserved two-state sequence is then used for compartment calls (see Methods 2.1.4).

In practice, this acts to de-noise our compartment calls. Where single sign-changes along the series would have resulted in a single-block compartment, these may now be modelled as noisy emissions from a single unobserved

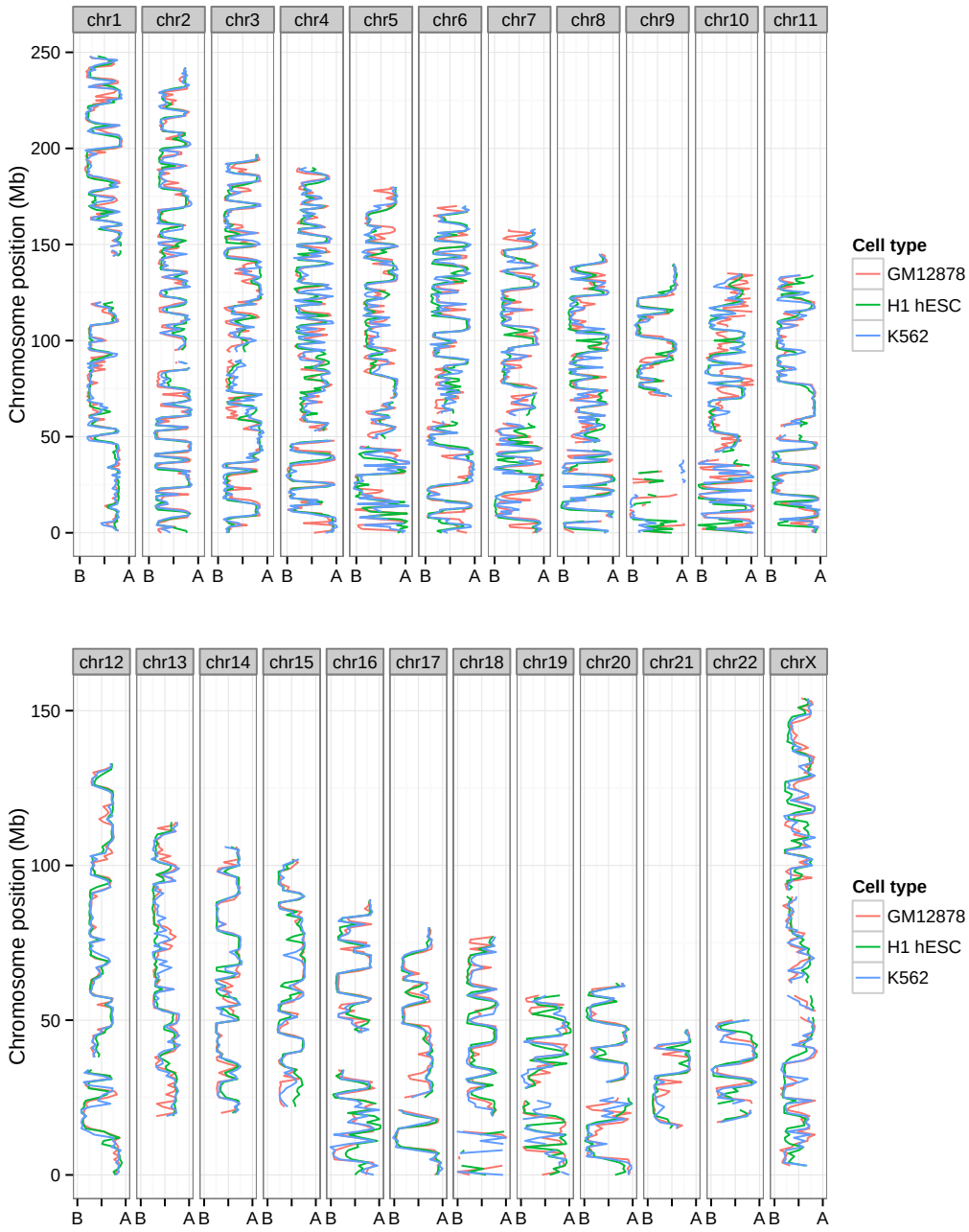

Figure 7: Compartment profiles are observably well-correlated between human cell types and across all chromosomes. Compartment eigenvectors are plotted along the lengths of each human chromosome (chrY and chrM are omitted) displaying strong concordance between three different human cell types.

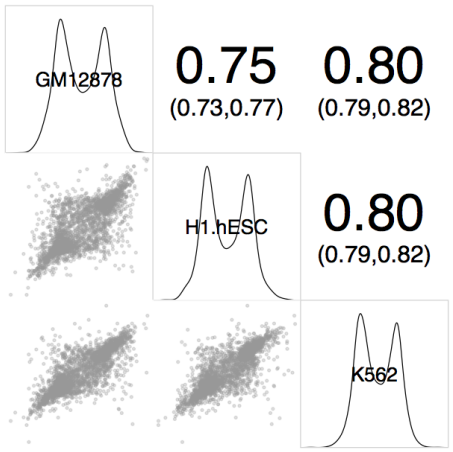

Figure 8: Compartment eigenvectors are highly correlated between human cell types. Megabase resolution compartment eigenvector values are shown in a plot matrix. *Upper triangle*: Pearson correlation coefficients between pairs, with 95% confidence intervals; *diagonal* Kernel density estimates of eigenvector values per cell type; *lower triangle*: *x*-*y* scatterplot of values.

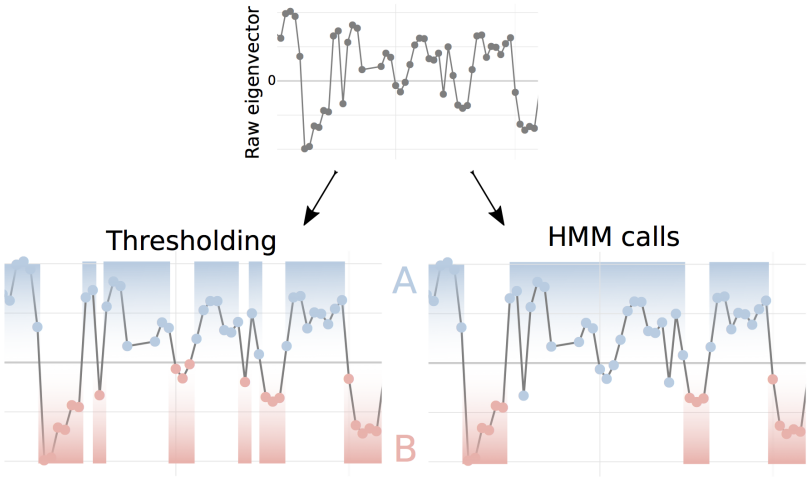

Figure 9: Compartment calls by simple thresholding method or context-aware HMMs. Chromosome compartments have previously been called through simple thresholding at 0,[7] in this work we also use an HMM-based method to call unobserved states that have emitted our noisy observed values (*right*).

state. An examplar region is showing in Fig. 9. 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 (Fig. 9).

3.4.2 TADs

Topological associating domains (TADs) are self-interacting blocks of the genome first described by Dixon *et al.*[8]. We applied the original TAD calling method without modification, which uses a measure of the directional contact bias of a fragment (Section 1.1.4 and Fig. 3).

The Dixon *et al.*[8] method of calling TADs relies on the detection of boundaries,[11] thus 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

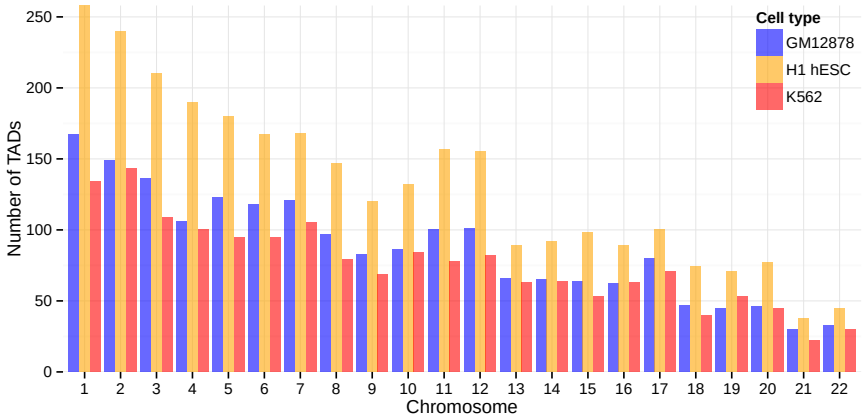

Figure 10: The number of called TADs per cell type under study. A greater number of TADs were called in H1 hESC (2,897 total) than in GM12878 (1,925) or K562 (1,677), due to the difference in sequencing depths in each experiment when matrices were binned at 40 kb resolution.

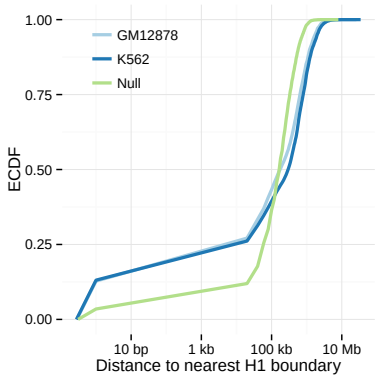

Figure 11: TAD boundaries are shared between cell types. The empirical cumulative density functions (ECDF) of distances between H1 TADs and those called in GM12878 and K562 are shown. These distances were compared with a null distribution calculated from randomly placed boundaries allocated at a matched resolution.

approximately 50% more TADs called than in the GM12878 cell type (Fig. 10). 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. 7, 8).

Despite differing numbers, there is still detectable levels of conservation of TADs between cell types (Fig. 11). 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%; K–S test: $D = 0.26$, $p \approx 0$). To illustrate this conservation with a real example, a 20 Mb region of chromosome 2 is pictured (Fig. 12), 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.

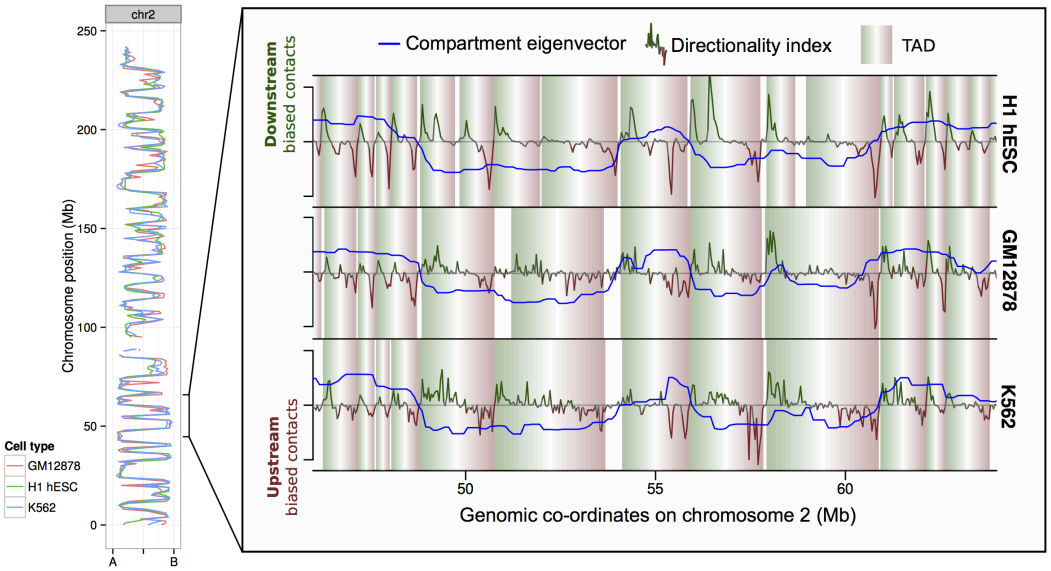

Figure 12: Concordance of chromatin structure at multiple scales over three human cell types. The eigenvector compartment profile is shown for chromosome 2 for three human cell types (*left*). At higher resolution, the zoomed region illustrates conservation of topological domains (TADs) over 20 Mb of the same chromosome.

3.5 DOMAIN EPIGENETICS

The use of well-studied human cell types allows intersection with publicly-available epigenomics datasets, such as those produce by the ENCODE consortium.[36] 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 2.2).

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. 13). This is expected, A compartments represent the actively-transcribed and accessible portions of the genome, and have previously been shown to positively correlate with many of the features shown.[7,15]

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. 13). 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. 13). Recent evidence suggests specialised roles for H2A.Z in regulating both repression and activation during embryonic stem cell differentiation, acting as a "general facilitator".[83] Additionally H2A.Z has been reported to mark histone octamers for depletion, thereby permitting gene activation during differentiation.[84] 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 TAD were not observably correlated with blocks of chromatin features (e.g. 8). Later studies have linked TADs with such enrichments, first in *Drosophila*[24] and later in human cells, where it was argued TADs are merely a low-resolution window to smaller "sub-

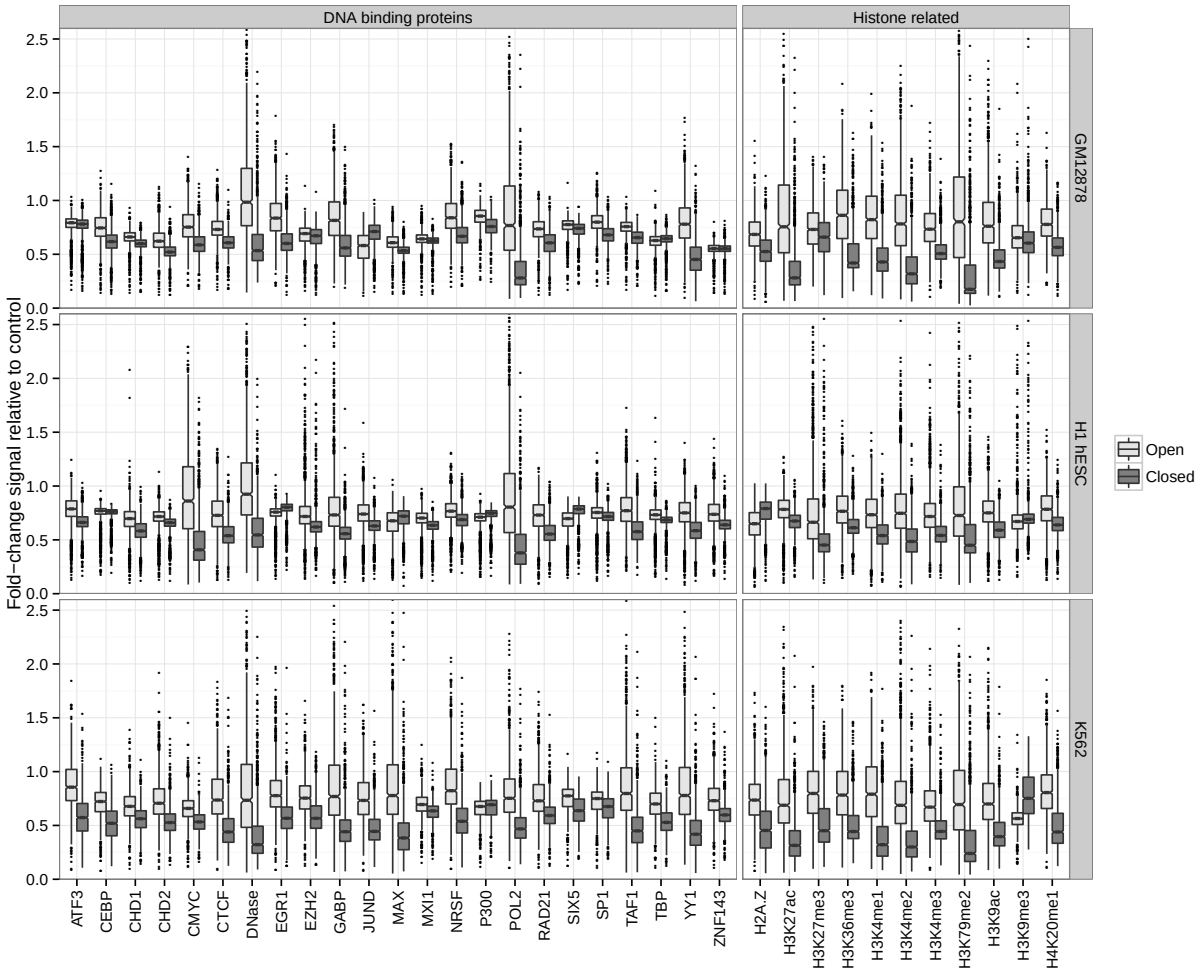

Figure 13: The chromatin signatures of A/B compartments. Notched boxplots summarise the distribution of each feature over 1 Mb bins in open (A) and closed (B) compartments genome-wide.

compartments", bearing similar active and inactive marks to their much-larger namesakes.[11]

Here we look for evidence of the Sexton *et al.*[24] characterisation of TADs called in our human cell types. 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 (Fig. 14). 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.

These findings also link with recent work that suggested TADs are windows unto "sub-compartments"[11] 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 Rao *et al.*[11] report sub-compartments only on extremely deep-sequenced samples, and at a scale below that of TADs.

3.6 VARIABLE REGIONS

Despite the vast majority of the genome being in matched chromatin compartments, there are also regions of disagreement. Reasons for observable differences include technical errors and bias, 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.

3.6.1 Chromatin state enrichment

Given our conservative definition of RVS (Section 3.6), 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[33] and SegWay[52,53], and tested for enrichment or depletion in our set of RVS (Methods XX).

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. 16). This observation can be explained as the undifferentiated cell type would not be expected to have 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. 17).

Together these state enrichments suggest the identified RVS reflect regions of cell-type specific biology, with heightened enhancer and transcriptional activity in their active cell type Fig. 17). Combined with the observed large-scale concordance of higher order chromatin organisation between cell types (Figs. 7, 12), these results reinforce a model of organisation whereby chromatin organisation is largely conserved and static across cell types,

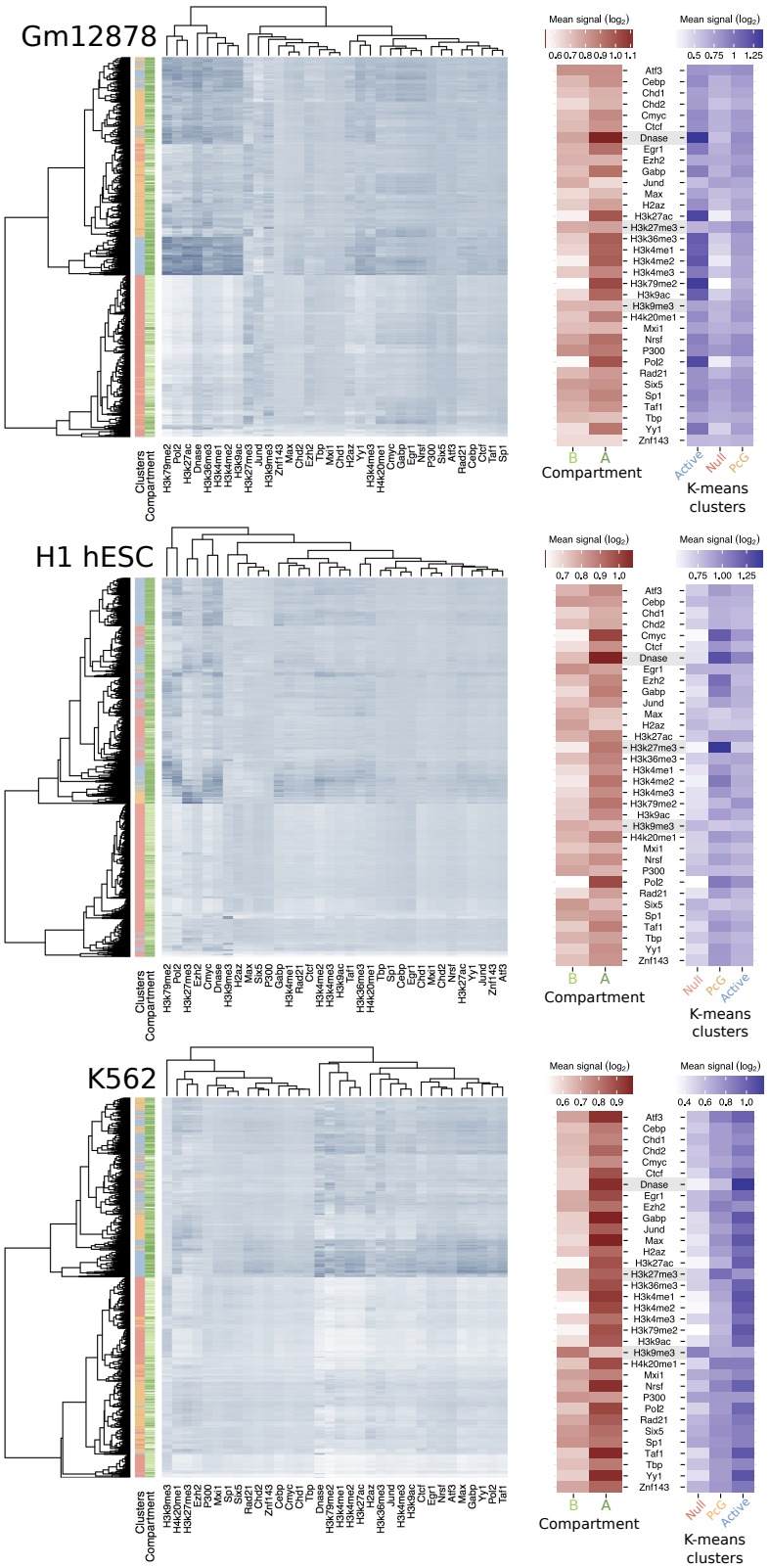

Figure 14: TADs reflect epigenetic domains. Following the Drosophila results of Sexton *et al.*[24], clustering of TAD domains by mean log2 signal of 34 ENCODE features distinguishes null, active and polycomb-associated (PcG) domains, as well as reflecting the encompassing A/B compartments.

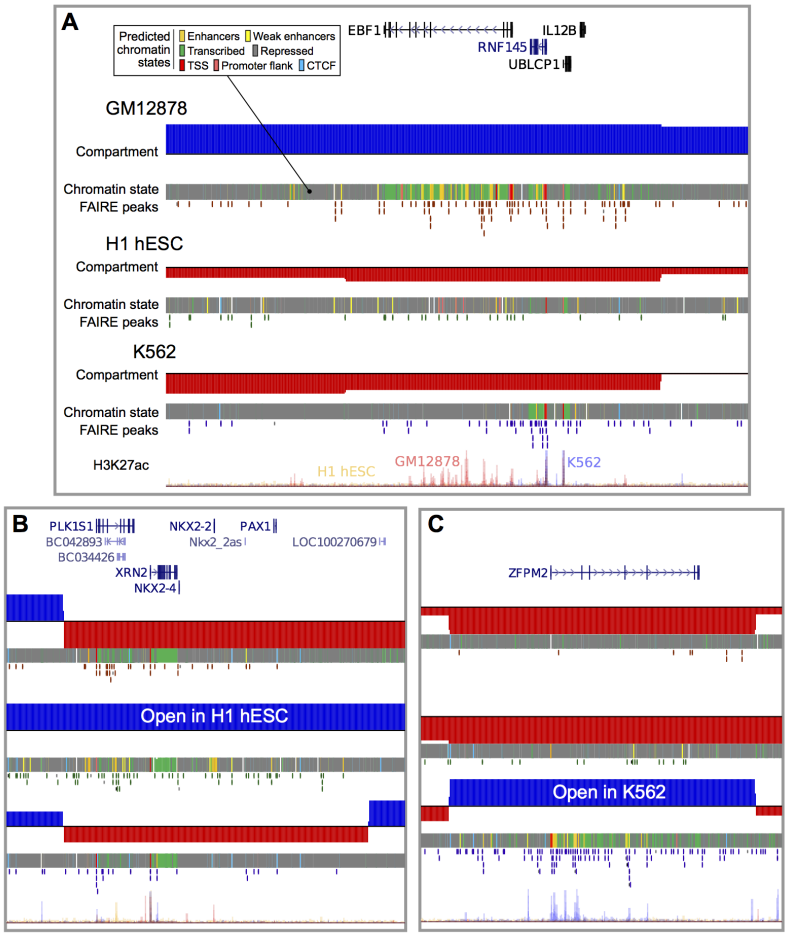

Figure 15: Structurally variable regions indicate cell type specific biology. Regions occupying the active A nuclear compartment in one cell type, but the repressive B compartment in the other two, were selected and ranked by the number of predicted active enhancers. Examples of particular interest from 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.

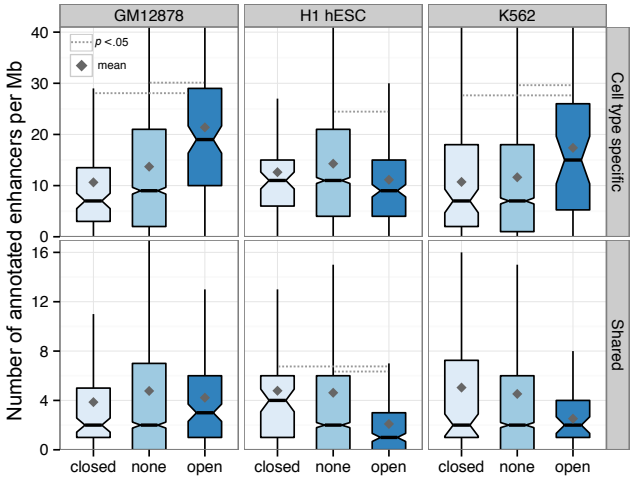

Figure 16: Regions of variable structure are enriched for cell type specific enhancers. Numbers of predicted enhancer states (cell type specific or shared between two or more cell types) are shown for regions with altered (open or closed) and non-altered (none) compartments in each cell type.

but also permits local flexibility in order to activate or repress regions of biological importance to a given cell type.

3.6.2 Gene ontology enrichment

Specific examples of RVS highlight genes of interest (Fig. 15) 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 RVS in each cell type were calculated using DAVID[55,85] and filtered by false discovery rate (FDR $< .05$). 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.[86] 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. Our RVS blocks were at least 1 Mb each so could contain more than one gene, thus enrichments were seen for those genes known to for paralogous clusters along a chromosome, such as olfactory receptors. The full results of these tests are given in the appendix (Tables A1, A2, A3).

The size of RVS could also explain we do not capture the relationship hinted at by cherry-picked examples (Fig. 15). Given 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.[7] 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 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 (18), 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.

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.[87] Kalhor *et al.*[18] 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 expectation with the eigenvector data used in this work is that active A compartments are enriched in the central

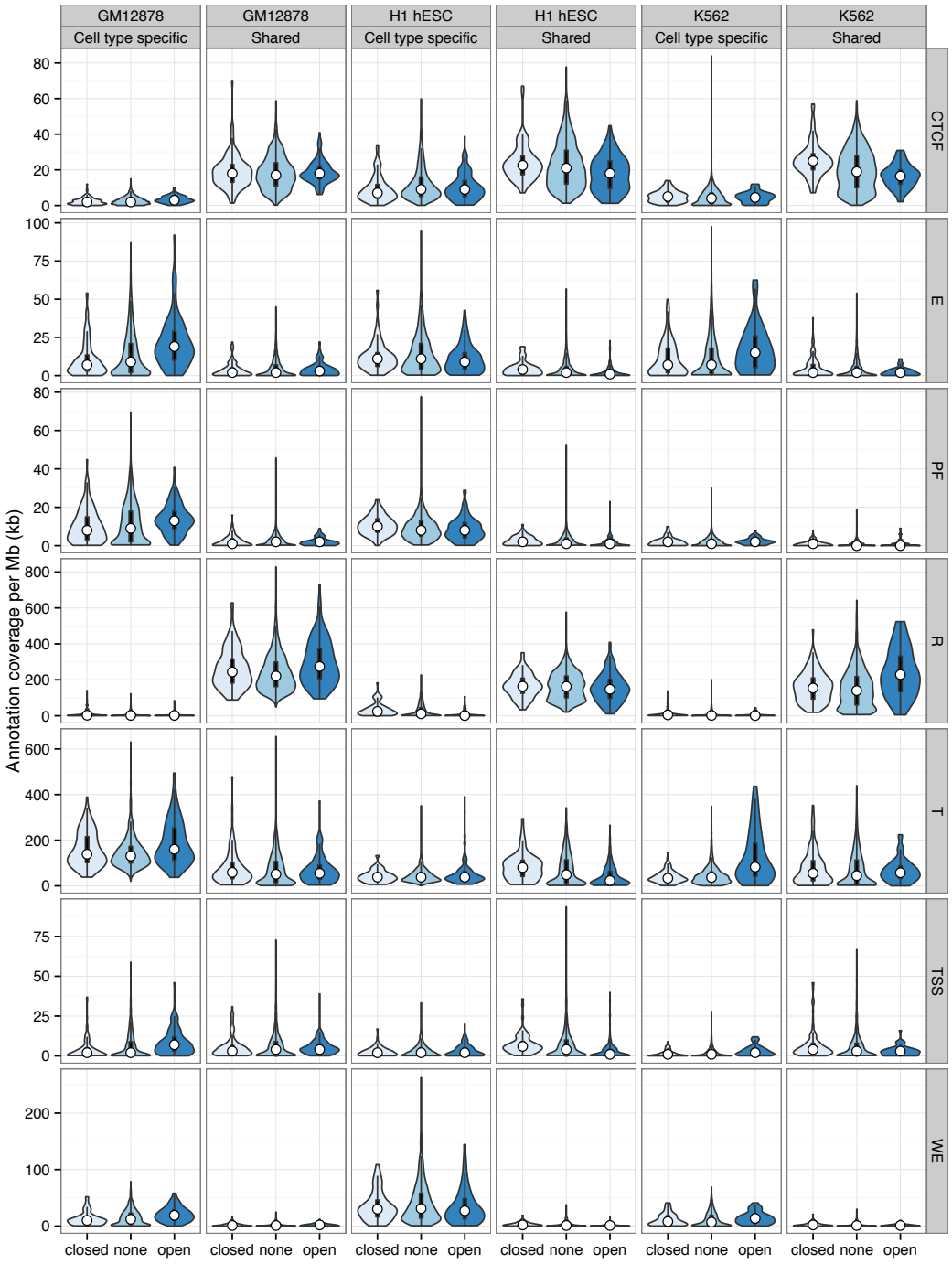

Figure 17: Distributions of features across all predicted chromatin states in regions of variable higher order structure. Distributions of the summed coverage of predicted chromatin states in each Mb per cell type are shown as bean plots. Predicted chromatin states are those from Hoffman *et al.*[53] 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.

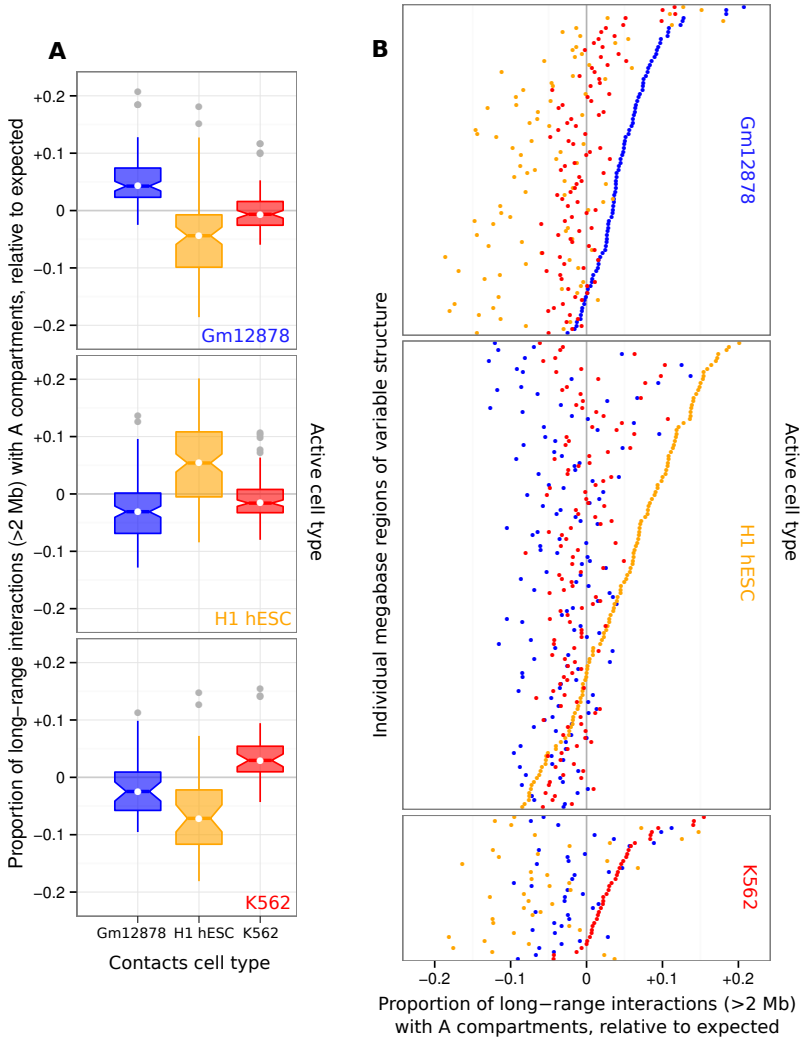

Figure 18: Regions of variable higher order structure change their genome-wide contact profiles to favour active compartments. Genome-wide normalised contacts were summed for each region of variable structure and the relative proportion of those that were with active / A compartments is shown across 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 boxplot (A) and with each individual region visualised (B).

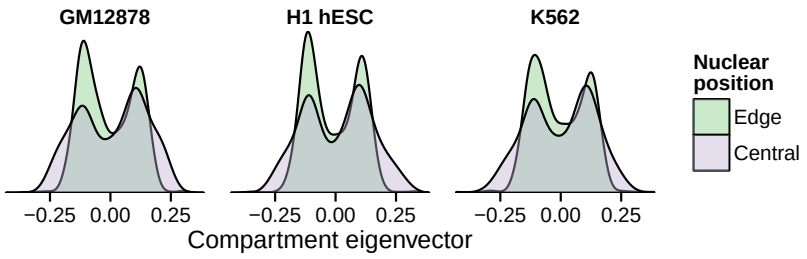

Figure 19: Nuclear positioning of chromosomes relative to compartment eigenvectors. Kernel density estimates showing peripheral chromosomes have a greater proportion of B compartments (negative eigenvectors) relative to centrally-position chromosomes. Positioning data from Boyle *et al.*[57] (Methods XX).

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".[57] Chromosomes whose DAPI hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by Boyle *et al.*[57], 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 $\alpha = 0.05$.[57]

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. 19). 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: $D = 0.10$, $p < 5 \times 10^{-3}$)

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 the ENCODE[36] and NIH Roadmap Epigenomics[88] projects. The breadth and depth of this new data offers unprecedented opportunities to further our understanding regarding the fundamental biology of the chromatin landscape. While many histone modifications can now be quantified experimentally,[34,89,90] an integrated understanding of general mechanisms underlying the cause or effect of these marks lags behind. A 2011 opinion piece asked the question "Histone modification: cause or cog?"[91] and speculated that nucleosome modifications could be by-products of transcription machinery, as opposed to the "histone code" hypothesis which suggests that histone modifications are placed to direct alterations in chromatin state. This latter hypothesis is often tacitly invoked in the chromatin literature, wherein a mark may be described as "repressive" or "activating" despite only the observation of a correlative relationship.[91] However, the recent flood of data from high throughput sequencing technologies have provided fascinating new glimpses of the ways chromatin and transcription are functionally related.

Recent studies have shown convincingly that local chromatin state measurements can accurately predict expression levels of genes on a genome-wide basis. Tippmann *et al.*[92] 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."[92] An independent study used a similar regression modelling approach to 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.[93]

A recent key study from the ENCODE consortium used chromatin (ChIP-seq) datasets to predict gene expression in a range of cell types as measured by a variety of experimental techniques.[94] 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 all cell lines and expression level technologies.[94] 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,

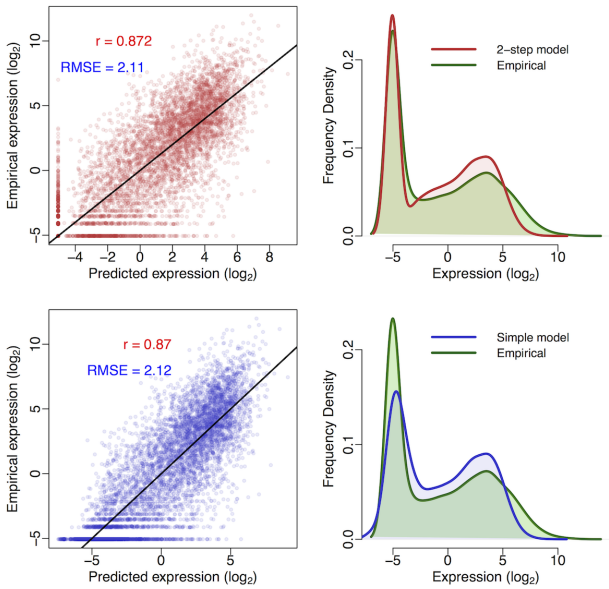

Figure 20: Comparison of a published two-step classification-regression model of transcription with a simple linear regression model. Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for the two-step model of Dong *et al.*[94] and simple multiple linear regression are shown (*left*) along with frequency distributions of predicted and observed expression levels (*right*). Scatterplots are annotated with Pearson's correlation coefficient (r) and the root mean squared error (RMSE); the black trendlines describe $y = x$. Following 10-fold cross validation, overall correlation coefficients were: linear model $0.87 \pm 1.77 \times 10^{-5}$; Two-step model $0.872 \pm 9.89 \times 10^{-5}$.

specific tags which precisely identify TSS positions as well as quantifying the abundance of a given transcript.[95,96]

These recent publications highlight the importance and relevance of advancing our understanding of chromatin biology through a model-based approach. We can extend this approach 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 REPRODUCING DONG *et al.*

Following Dong *et al.*[94], we first reimplemented the published ENCODE modelling framework to replicate their results. In doing so we were able to analyse the strengths and caveats of their approach; for example we found, surprisingly, 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. 20).

An innovative element of Dong *et al.*'s modelling approach 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 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, then the bin producing the highest correlation is designated as the 'bestbin' and that bin's normalised ChIP-seq signal intensity is taken forward for the full model. This was shown to raise the correlation (between predicted and observed expression) by 0.1 in the simple regression model, an increase in accuracy of almost 13%, relative to simply taking the average value across all bins.[94]

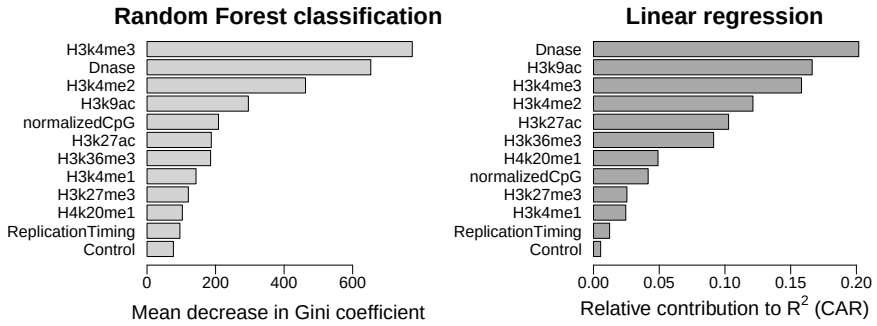

Figure 21: Relative importance metrics for variables in both stages of a reimplementation of a published model for predicting transcriptional output. Variable importance is measure by decrease in Gini coefficient for the RF classification step, and by CAR R^2 decomposition[97] for the linear regression step.

4.2.1 Model adjustments

We attempted to improve the accuracy of predicted expression values produced by Dong *et al.*[94] through increasing the number of informative regressors. While Dong *et al.*[94] 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 Dong *et al.*[94] with previously-published genome-wide replication timing ratios measured in BG02 ESCs.[98] This data is of a different origin to the transcriptional data in this case (which was record in H1 hESC) but replication timing is thought to be largely conserved between cell types.[58]

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*). The reasons for this are likely that the data were relatively low-resolution (1 Mb), from a imperfectly matched cell line and also that the existing model is already achieving such accurate results that they must already be accounting for most of the maximal explainable variance in gene expression given experimental and biological noise. With this in mind, 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. 21), implying that large-scale chromatin domains and do not have significant influence on the expression of the genes resident within them.

4.3 MODELLING FANTOM5 EXPRESSION DATA

Using unpublished FANTOM5 CAGE data and the approach established above, I 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.

The first stage of the analysis was to map each CAGE cluster to a representative TSS. FANTOM5 robust gene mapping[99] provided corresponding Entrez Gene IDs for gene-associated CAGE clusters, and we selected the most expressed cluster to represent the expression level of its mapped gene. We then compared these to Ensembl TSS annotations (v69) and discarded those tag clusters centered on a point > 50 bp from an annotated TSS associated with the mapped Entrez Gene ID, thereby removing enhancers and other non-genic transcribed regions.

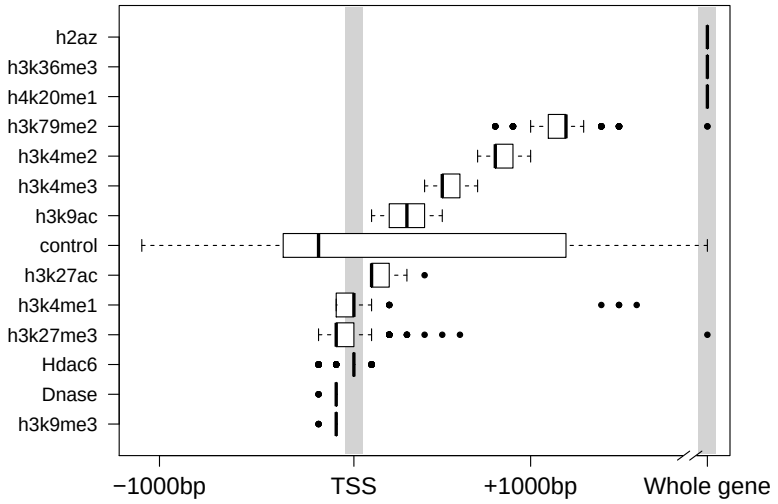

Figure 22: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, DNase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 kb flanking 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.

Next we retrieved a number of genome-wide histone modification datasets from the ENCODE and NIH Roadmap consortia which were measured in H1 hESC cells, taking these to be reflections of the chromatin state t_0. I implemented the previously-described 'bestbin' strategy[94] (Section 4.2) to objectively select the most-correlated binned signal for each chromatin H1 hESC mark. To explore 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 dataset (Fig. 22).

Figure 22 shows that bestbin selections are often consistent, indicating there are predictably informative regions relative to a TSS for each chromatin factor. Furthermore, the selected bestbins match known biological mechanisms; for example the H3K36me3 mark's bestbin is consistently the whole gene measurement and this mark is known to be enriched in actively transcribed exons.[92,100,101] The negative control variable (ChIP-seq input) shows no strong location bias, as expected (Fig. 22).

Having matched a variety of genome-wide H1 hESC chromatin datasets to the FANTOM5 timecourse expression data, we then built a regression model using a Random Forest (RF) approach.[102] This method outperforms a simple linear model in initial comparisons and is able to capture non-linear relationships as well as interactions without them being explicitly specified.[47]

Figure 23 shows the resulting predictions of a preliminary RF model against the actual recorded expression over a test set of approximately 11,000 TSS. This model was built with 15 predictors including control ChIP-seq input, though some of these could be removed without loss of accuracy. The model predictions evaluated with 10-fold cross validation show a 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 $PCC =$

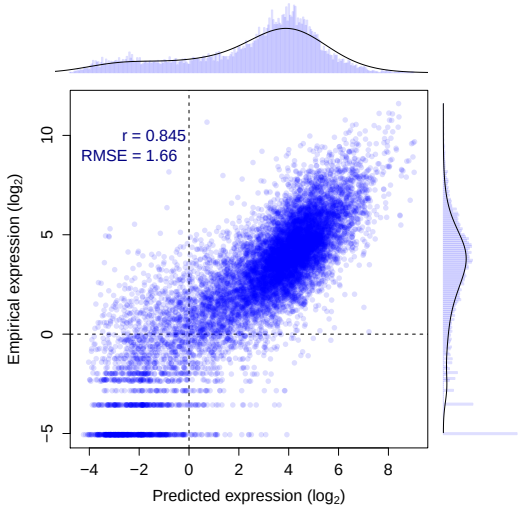

Figure 23: Random Forest predictions of FANTOM5 expression data. RF model predictions are plotted against their empirical values. The marginal distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson's correlation coefficient (r) and the root mean-squared error (RMSE) are shown (*inset*).

$0.825 \pm 3.2 \times 10^{-5}$, $p < 2 \times 10^{-15}$). This result is worse than that of Dong *et al.*[94] 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.[94] The RMSEs, when normalised by the range of observed values, compare more favourably (0.11, compared with Dong *et al.*'s: 0.14).

A possible explanation for our lower modelling accuracy relative to that of Dong *et al.*[94] is that in our case, while both chromatin data and expression timecourse were measured in H1 hESC cells, the experiments took place at different institutes and using unstandardised protocols and cell cultures. For comparison, a previous study using chromatin measurements from a number of different sources to predict expression in a matched cell-type reported a predictive correlation of 0.77.[103] The ENCODE consortium, on the other hand, went to some lengths to standardise protocols and minimise batch effects between samples.[36] Additionally, Dong *et al.*[94] implemented a pseudocount optimisation step whereby an additional count added to each binned signal intensity prior to log transformation to maximise expression correlation. In the model presented above, a fixed psuedocount of 1 was used to avoid introducing positive bias towards higher correlation. Another difference between the two approaches is our use of a single-step model; Dong *et al.* found a small increase in correlation using their classification-regression approach but with the model implemented herein (Fig. 23) this approach gave no obvious advantage (for example, $r = 0.834 \pm 0.007$, RMSE $= 1.77$ when applied to the same test and training data used in Fig. 23).

4.4 MODELLING HIGHER ORDER CHROMATIN

Accurate predictive modelling of transcription in a variety of cell types offered several novel insights into the internal between histone modifications and transcription factors with transcriptional machinery, and advanced 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 3).

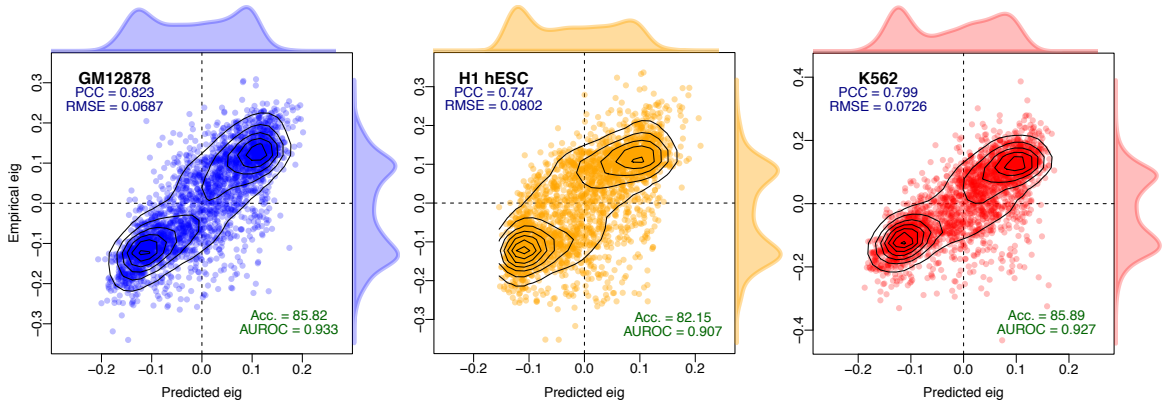

Figure 24: Compartment eigenvector model predictions are highly correlated with observed values. Pearson correlation coefficient (PCC) and root mean-squared error (RMSE) report the degree of success of the regression model, whereas accuracy (Acc.) and area under the receiver operating characteristic (AUROC) give the classification accuracy of binarized outcomes.

Previous publications have identified several correlates which track compartment eigenvector profiles to varying degrees,[7,14] yet to date these relationships have not been quantitively investigated. The above-described modelling framework offers a statistical approach to understanding the drivers of these observed correlations.

4.4.1 Predictive model

We built Random Forest regression models (Methods 2.3.1) 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 0.82–0.75 (Fig. 24), comparable to that achieved by Dong *et al.*[94] in the prediction of transcription.

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 retraining a classifier and building parallel models, instead for an estimate of classification accuracy we threshold our regression predictions (Methods 2.3.2). We found our Random Forest models achieved high classification accuracy with upwards of 82% of the all genomic bins correctly assigned in each cell type (Fig. 24).

This predictive performance underlines the strong connection between locus-level features and higher order chromatin structure previously noted by Lieberman-Aiden *et al.*[7] 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 "over-fitting". In machine-learning, over-fitting refers to the point at which parameters are being optimised to capture noise within a feature set, as well as signal, thereby giving an overoptimistic model performance which would not generalise to another featureset with different noise profiles.

To test if over-fitting was causing our high observed accuracy, we cross-applied models learnt in one cell type to unseen input data from each of the other 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 be overfitted to their respective cell types. Conversely, it could be the case that

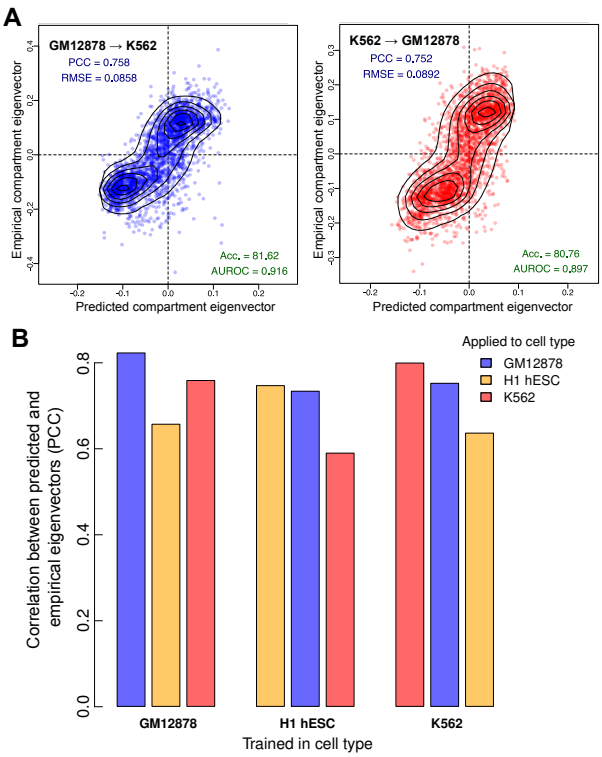

Figure 25: Models of higher order chromatin structure learned in one cell type can be cross-applied to others. Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. (A) The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross (PCC = 0.75). (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.

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 with similarly-high levels of accuracy (Fig. 25). This gives good evidence not only that are models are not overfitting to cell-type specific noise, but also that there exist broad rules linking chromatin conformation and locus-level feature aggregation. The cross-application suggests there exists enough commonalities for compartment profile predictions to transcend the cell-type specific biology inherent to an embryonic stem cell or differentiated lymphoblast.

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. 8), a testable hypothesis is that these conserved regions are drivers of cross-applicability between cell types. Conversely, genomic regions which vary most across the cell types in our dataset should be more difficult to predict.

Indeed we found the most variable regions across cell types were then most difficult to predict through our Random Forest modelling framework (Fig. 26). 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 that cell type, and conversely the most variable third shown significantly depleted predictability (Fig. 26). This latter result suggests these variable regions could either be those which are noisiest, where the

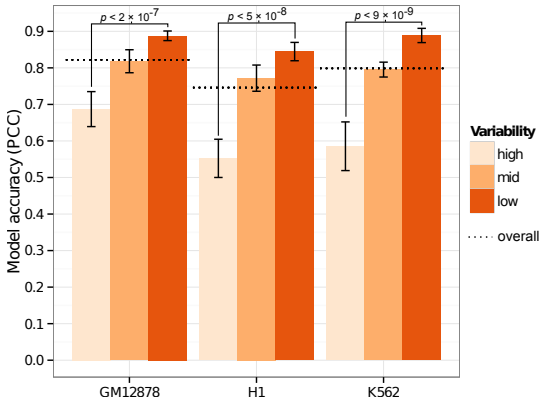

Figure 26: Genomic regions that vary across cell types are modelled less successfully than static regions. Genome-wide compartment eigenvectors were partitioned into thirds according to their median absolute deviation (MAD) across the three cell types under study. Models were fit independently to each third, and the modelling accuracy is compared.

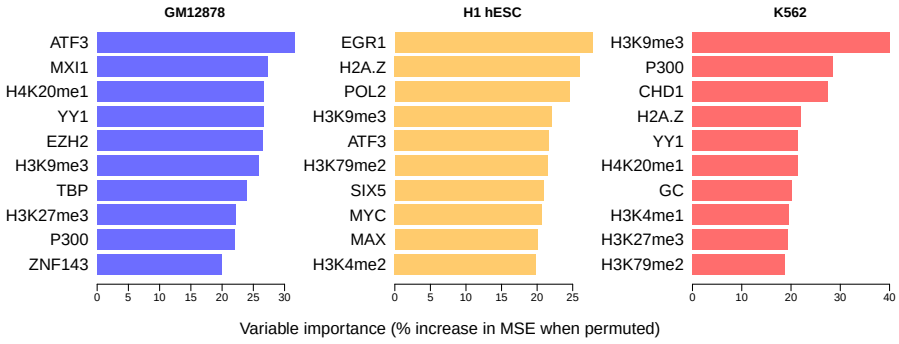

Figure 27: Variable importance per cell type specific model. Variable importance for each Random Forest model was measured in terms of percentage increase in mean squared error when permuted (Methods XX) and the top 10 ranking variables are shown for each model.

eigenvector is least capturing compartment structure, or where cell-type specific biology is influencing compartment structure in each case, in ways not captured by out input feature set and low resolution modelling pipeline.

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 on the top 10 most highly-ranked features in cell type specific models shows some agreement but also substantial differences between cell types (Fig. 27)

Only one input feature, H3k9me3, is present in the top 10 most important variables of each model (Fig. 28). H3k9me3 is one of the few features to be negatively correlated with compartment eigenvectors (e.g. Fig. 30). Of those shared between two cell type models, H3k27me3 is also a repressive mark and deposited by polycomb repressive complex 2 (PRC2)[104] while H2A.Z is a histone variant again linked to polycomb-regulated genes and essential for embryonic development.[105] Furthermore EZH2, the catalytic subunit of PRC2,[106] is also included in the feature set but only highly ranked in the GM12878 cell type model. As another example, MYC and MAX are found in the top 10 influential variables in H1 hESC, while MXI1 is found to be an informative variable in GM12878. This is in keeping with recent results suggesting MYC binds open chromatin as a transcriptional amplifier

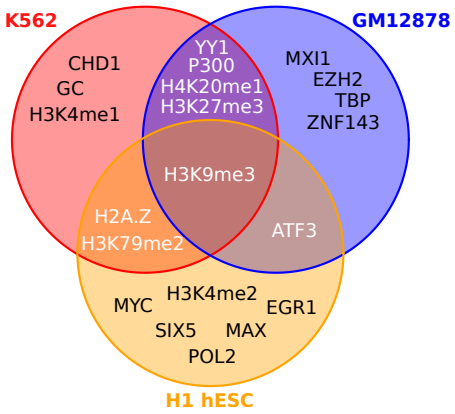

Figure 28: Intersections of the top 10 ranked variables in the cell type specific models. Venn diagram illustrating intersections between sets of ten most influential variables per cell type specific Random Forest regression model of compartment eigenvector (Fig. 27).

in embryonic stem cells,[107,108] with MAX and MXI1 long being known as antagonistic co-regulators of MYC.[109] These biological relationships between variables may help explain the observed differences between models: different representatives of correlated clusters of input variables may be being selected in each model (see Section 4.4.5).

To assess the significance of observed intersections (Fig. 28), 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 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 ten 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 Random Forest models.

In addition to rankings, raw variable importance metrics can be compared between cell-type specific models (Fig. 29). This shows 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 hematopoietic cell line models, potentially reflecting its activity as a histone acetyl transferase that regulates hematopoiesis[110] and a noted involvement with CTCF in chromatin looping.[111]

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, GC content). To explore these relationships, we performed unsupervised clustering of our feature sets in each cell type (Fig. 31).

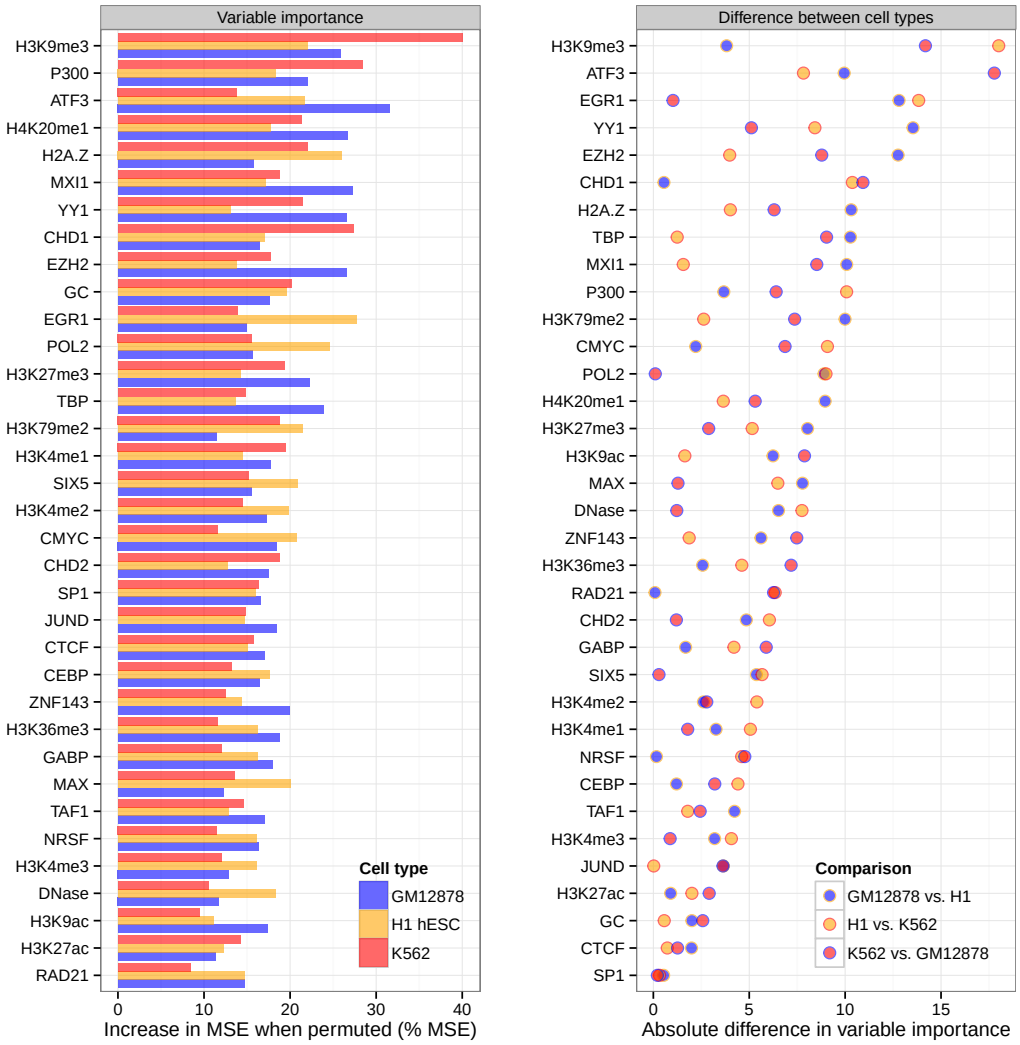

Figure 29: Comparison of variable importance between three cell type specific Random Forest models. Variable importance for each Random Forest model was measured in terms of percentage increase in mean squared error when permuted (Methods XX). Results are shown sorted by mean variable importance (*left*) and by largest absolute difference in pairwise comparisons (*right*).

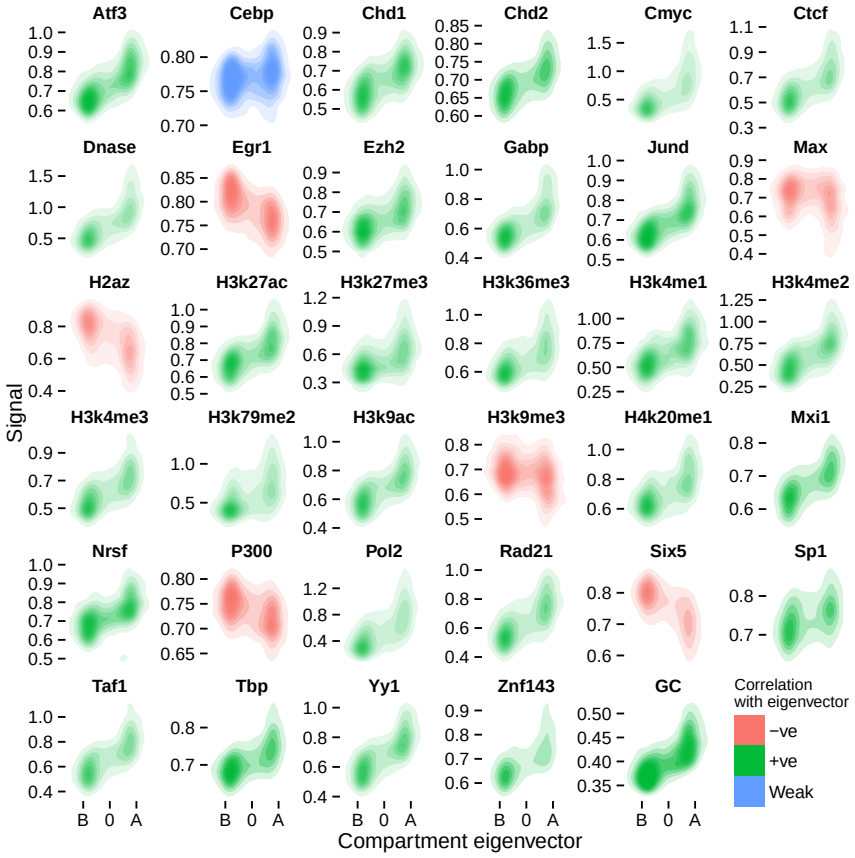

Figure 30: Correlations of individual features with compartment eigenvector in the H1 hESC cell type. Two-dimensional kernel density estimates show the density of points in a scatterplot of compartment eigenvector (x-axis) against each input feature individually (y-axes). Features with a PCC against eigenvector of above or below 0.1 are coloured as positive or negative, respectively.

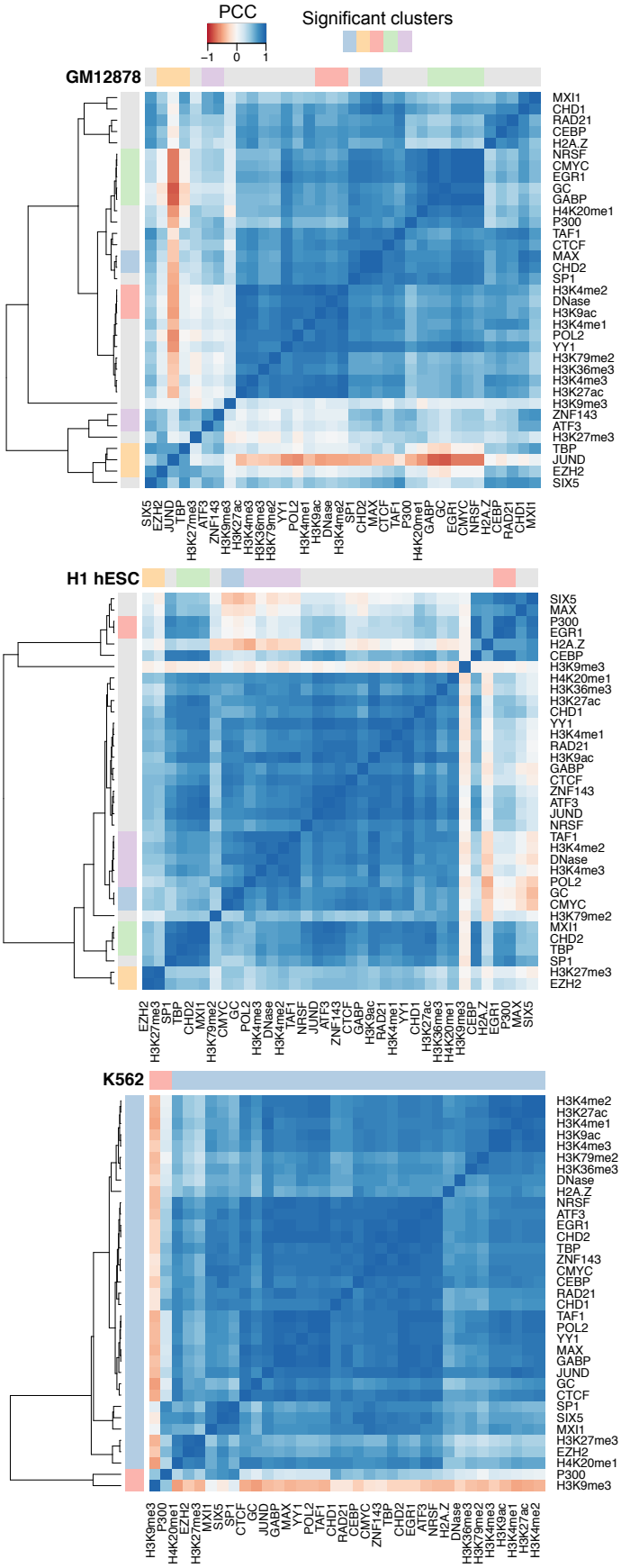

Figure 31: Correlation heatmaps of the 35 features used to model compartment eigenvectors. The Pearson correlation coefficient (PCC) of genome-wide 1 Mb bins of each feature were pairwise correlated with each other. The features were also clustered using hierarchal clustering. The significance of these clusters was determined through multi-scale bootstrap resampling, with those clusters that were stable across different sizes of resampling deemed significant, as implemented in the `pvclust` R package.[42]

We found as expected 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. 31).

Outliers are also present. H3K9me3, noted for high variable importance in each model (Fig. 27) and the only feature ranked within the top 10 in each model (Fig. 28) 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. 31). 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 leaned 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. 32).

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. Other methods could have been used and should be compared. Here we compare our RF approach with two other options: multiple linear regression and partial least squares regression.

Our results confirm RF as a suitable and powerful approach for modelling our relationships of interest in this work (Fig. 33), 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 $= 0.787$), 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 $= 0.139$) indicating a 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 (ref XX).

Partial least squares regression is another technique which used dimensionality reduction to engineer a lower-dimension orthogonal feature set. Hence this method is well-suited to multi collinear inputs, such as our feature set. As expected, PLS regression provides highly accurate cell type specific predictions (mean PCC $= 0.750$) and during cross-application (mean PCC $= 0.641$), though in both cases produces slightly inferior results to RF models (Fig. 33).

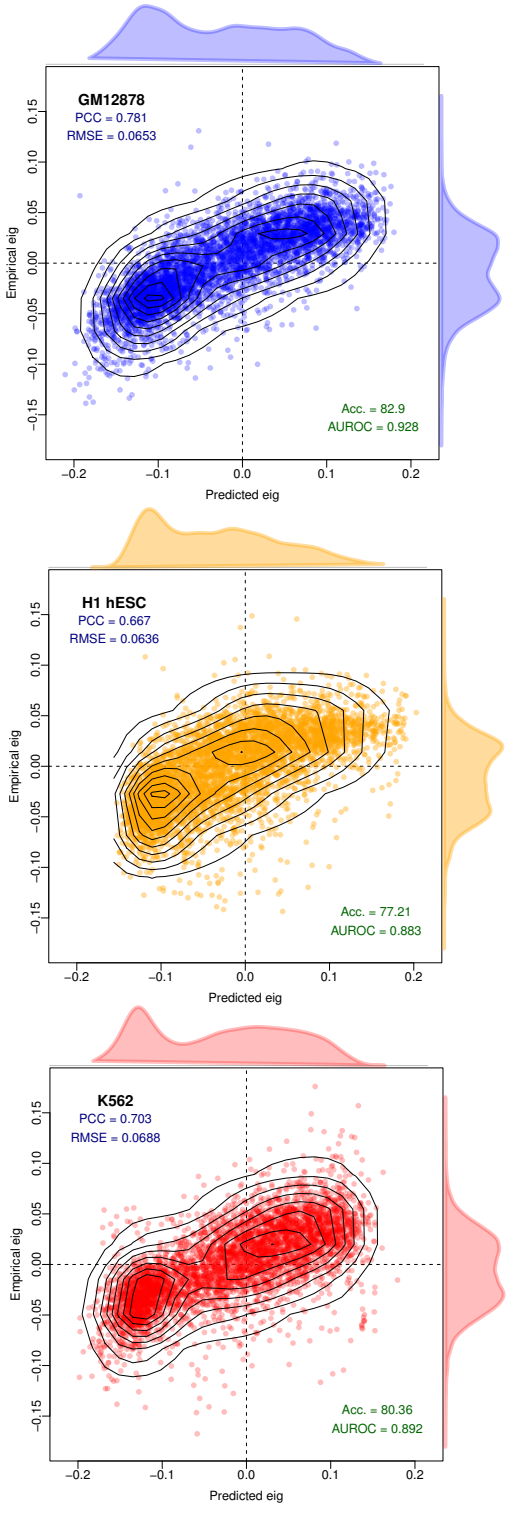

Figure 32: Models learned at 1 Mb resolution can be applied to higher resolution datasets. Despite having been trained on low resolution training sets, the Random Forest models generated can successfully predict compartment eigenvectors at higher resolution (100 kb, a 10× zoom). Eigenvectors at a higher resolution than this do not necessarily reflect A/B compartmentalisation.

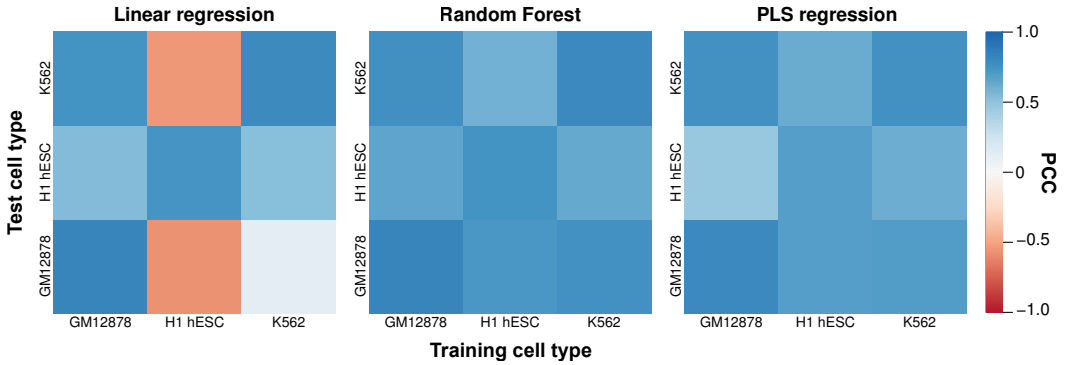

Figure 33: Comparison of Random Forest performance with other modelling approaches. Heatmaps show the Pearson correlation coefficient between predicted and observed compartment eigenvectors genome-wide for three regression techniques: multiple linear regression (LM), Random Forest (RF) and partial least squares (PLS). Results are summarised in Table 3.

Table 3: 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. 33.

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

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 revelling insight beyond simple correlations (e.g. Fig. 31). Figure 34 shows a "circle of correlations", where features are plotted onto polar co-ordinates against the first two PLS components. Interpretation of this figure is that nearby variables in the scatterplot are positively correlated, and the vector length from the circle centre is proportional to said variable's representation in the model. Negatively correlated variables point in opposite directions while uncorrelated variables are orthogonal to each other.[112] 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. 34).

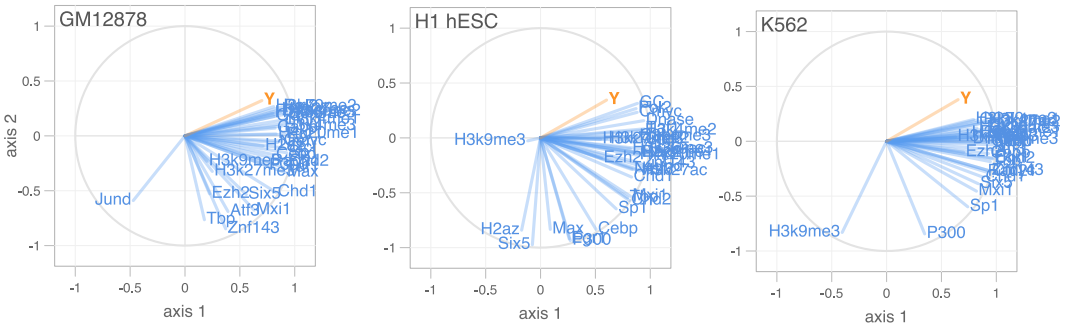

Figure 34: Circle of correlations of variables compared with PLS axes. Model variables are plotted against the first two axes used in PLS regression models per cell type. Y represents our compartment eigenvector.

4.5.3 Non-independence

As recognised through our use of Hidden Markov Models (Methods XX), 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 any region's relationships with their 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 in 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.

Conversely, 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[36,40] in each cell type which have thus far gone unused. Here we'll 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.[113] A detailed explanation of this feature selection procedure can be found in Methods XX. It should be noted that despite its continued widespread usage, several statistical issues have been identified with the stepwise procedure for model selection.[114]

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 4), however it must be said that differences in model performance across all comparisons are modest. These results do show that even expanded feature sets of up to 187 input

Table 4: Performance comparison of full and optimised RF and ML 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

features add little explanatory power beyond that of much less complex models with 20 or fewer input variables (Table 4).

4.6.2 LASSO (ℓ_1) 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 absolute values and sums, rather than squared values as in ℓ_2 regularisation, coefficients can be shrunk to 0 thereby removing terms from the model. Thus LASSO combines coefficient shrinkage of techniques like Ridge regression with a type of feature selection as seen in stepwise regression.[50,115] A detailed explanation of this method can be found in Methods XX.

Again we can perform a simplistic comparison of model performance using LASSO regression and other techniques (Table 4). 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.

The LASSO is a tunable algorithm, thus we can gain additional insights of coefficient traces over varying the regularisation parameter, λ.

4.6.3 Regularised Random Forest

Random Forest (RF) comparisons are included for comparison in Table 4 where RF models were built using model-selection procedures based on linear regression. The result of this is 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.[116,117]

While RF is a much younger technique than linear models, a framework for Regularised Random Forests has recently been described[118] and implemented in the R package `RRF`.[119] The 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 XX).

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 4).

5 CHROMATIN DOMAIN BOUNDARIES

5.1 INTRODUCTION

Multiple studies have defined chromatin domains of different types, for example: chromosome compartments;[7] topological associating domains (TADs);[8] contact and loop domains;[11] physical domains;[24,120] and others.[23] The existence of these domains necessitates "boundary regions" either between consecutive domains or bookending more sparsely-positioned domains, however the functional relevance of said boundary regions is still open to debate.

In their study of topological domains, Dixon *et al.* identified average enrichments over TAD boundary regions in both human and mouse for various features including CTCF and PolII.[8] 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,[121] however an alternative theory could be that looping between enhancer elements and promoters results in an observable boundary through C-method experiments.[11] Another non-exclusive explanation is that if chromatin domains represent co-regulatory regions as is widely thought,[121–123] 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.[124] This study found that while histone modifications within the body of a TAD could be removed without affecting the structure, deletion of a boundary did have an effect and lead to increased intradomain contacts.[124] Surpsingingly however, this effect was not total and some observable barrier remained, lending evidence that TADs may be centrally constrained, rather than by their borders.[124]

A second experiment used CRISPR genome editing to link TAD boundary changes with limb development disorders,[125] indicating that boundary changes could provide an underlying explanation for pathogenic non-coding structural variants.[126] Similarly, domain boundaries on X-chromosomes were found to be weakened following the disruption of condensation binding sites.[127] Together these studies suggest a complex scenario whereby TAD boundaries are an important structural feature, yet do not fully explain domain partitioning.

Computational analysis of boundaries has emerged during the time this work was completed. Border "strength", here defined by the ratio of total intra:inter-domain contacts, was found to correlate with increased occupancy of a combination of bound architectural proteins.[128]

Many questions remain about chromatin boundaries. For example, are the observed enrichments 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 TAD AND COMPARTMENT BOUNDARIES

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.[8] 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.[8] 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 (see Methods XX) 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 XX).

Our findings 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, Figure S12), including peaks associated with active transcription of genes (POL2, H3K9ac) and dips in H3K9me3, as previously reported.[8] However other novel feature peaks of interest emerge across cell types, such as peaks of H4K20me1, a modification previously implicated in chromatin compaction.[129] 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 neighbouring 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 an HMM 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, Figure S13). Compartment boundaries display similar spectra of enrichments to previously studied TAD boundaries[8] but at lower resolution, reflecting the different scales of these levels of organization (Figure 6B, Figure S13). Peaks associated with active promoters (POL2, TAF1, 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 reprogram-

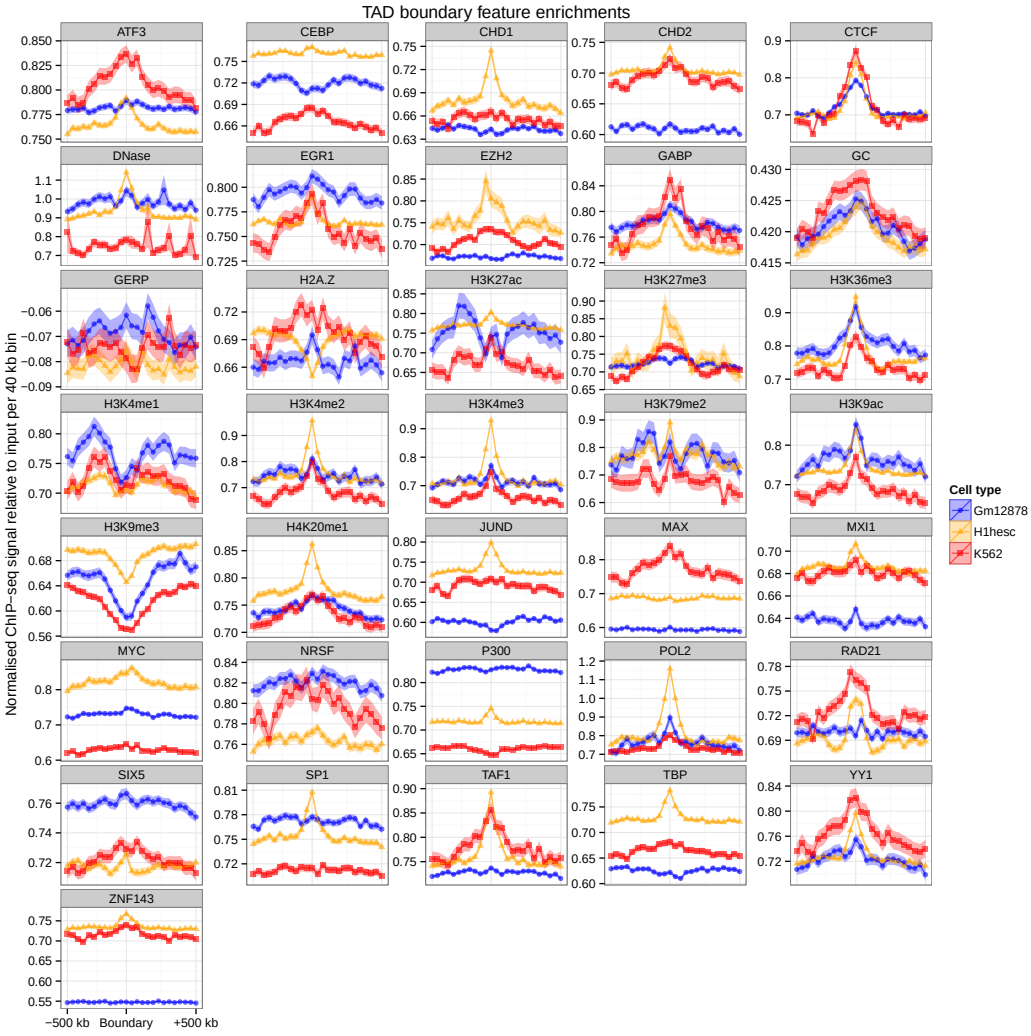

Figure 35: TAD boundary enrichments and depletions. 36 features were averaged over 1 Mb windows centred on TAD boundaries genome-wide (25 × 40 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

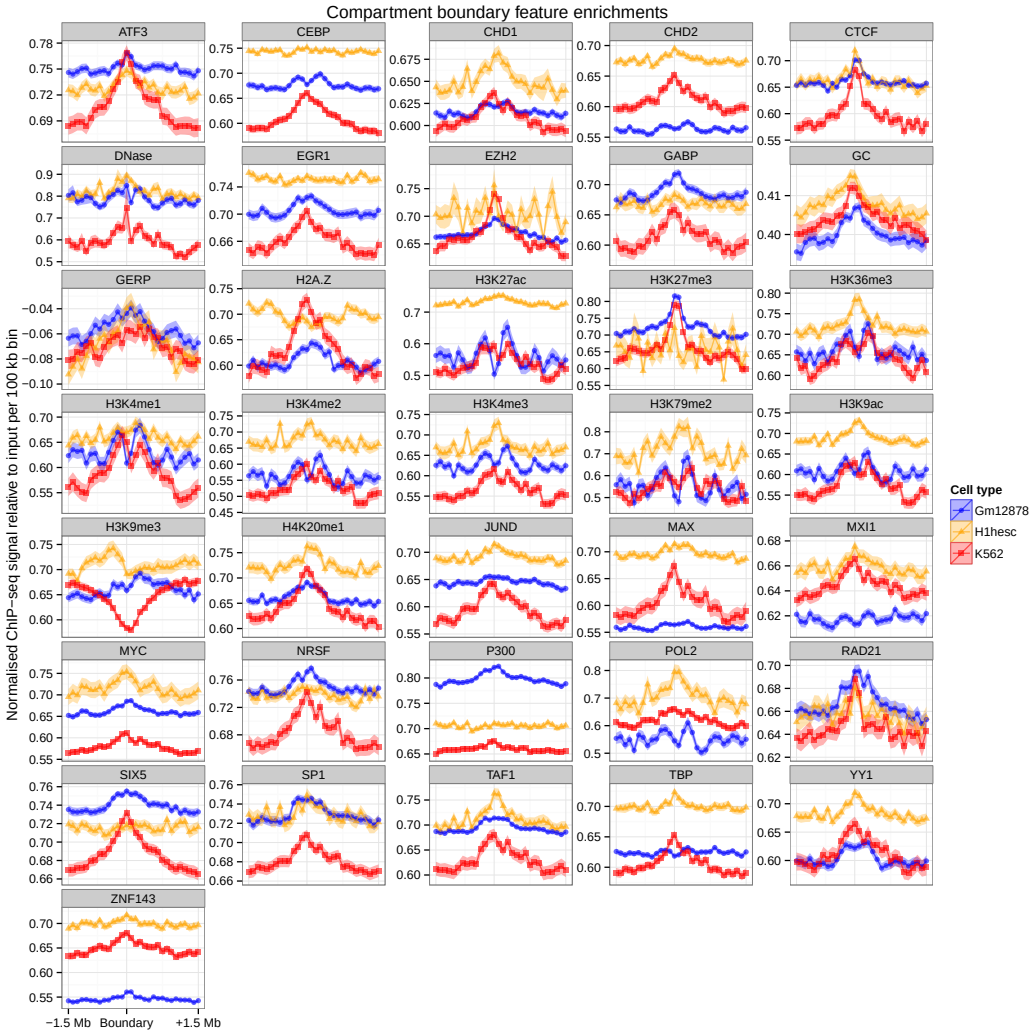

Figure 36: Compartment boundary enrichments and depletions. 36 features were averaged over 3 Mb windows centred on compartment boundaries genome-wide (30 × 100 kb bins). Ribbons represent 95% confidence intervals of the mean at each position.

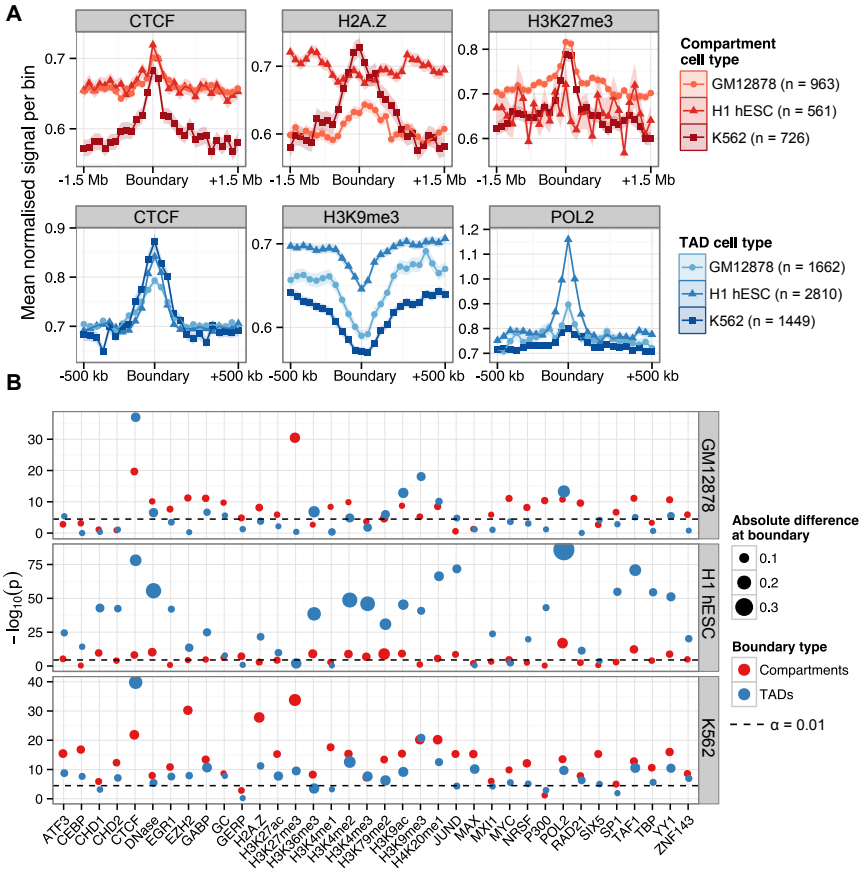

Figure 37: Compartment and TAD boundary enrichment summary in three human cell types. (A) Selected profiles for locus-level features are shown for TAD boundaries (CTCF, H3K9me3 and POL2) and compartment boundaries (H2A.Z, H3K4me2 and YY1), as a mean normalized ChIP-seq signal relative to input chromatin 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.

ming.[130] Remarkably, H3K79me2 has also recently been shown to mark the borders of small (hundreds of bp) regions of open chromatin.[131] 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.[132] 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.[133] Various other enrichments with very modest effect sizes are also evident at compartment boundaries (Figure 6B, 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, 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.

5.2.1 CTCF and YY1

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.[134] Co-binding of CTCF with YY1 has also been shown to identify a subset of highly conserved CTCF sites.[135] 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.[8] To test this, we split our sets of TAD boundaries into those possessing ChiP-seq peaks (region peaks called by ENCODE[36]) 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 (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 (Figure S14), reinforcing previous findings that describe the distinct influences of CTCF and cohesin in organizing chromatin structure.[133,136,137]

5.2.2 Repeats

Dixon *et al.*'s study of TAD boundaries identified short interspersed element (SINE) repeats as being enriched over domain boundaries and suggested roles for these repeats in altering genome organisation, in line with prior evidence.[8,138] Interestingly, SINE elements are thought to be responsible for spreading CTCF binding sites through mammalian genomes[139] (thought not in primates[135]). Analysis of recent high-resolution Hi-C data again reported a SINE B2 link with CTCF loops in mice.[11] Together these results suggest repeats could be a key component in the makeup of domain boundaries.

To investigate this, we used the `RepeatMasker`[140] software package to call repeat classes and families in the `hg19` and `mm10` genome assemblies. Counts for each annotated feature were then average over boundaries as described previously (Methods XX).

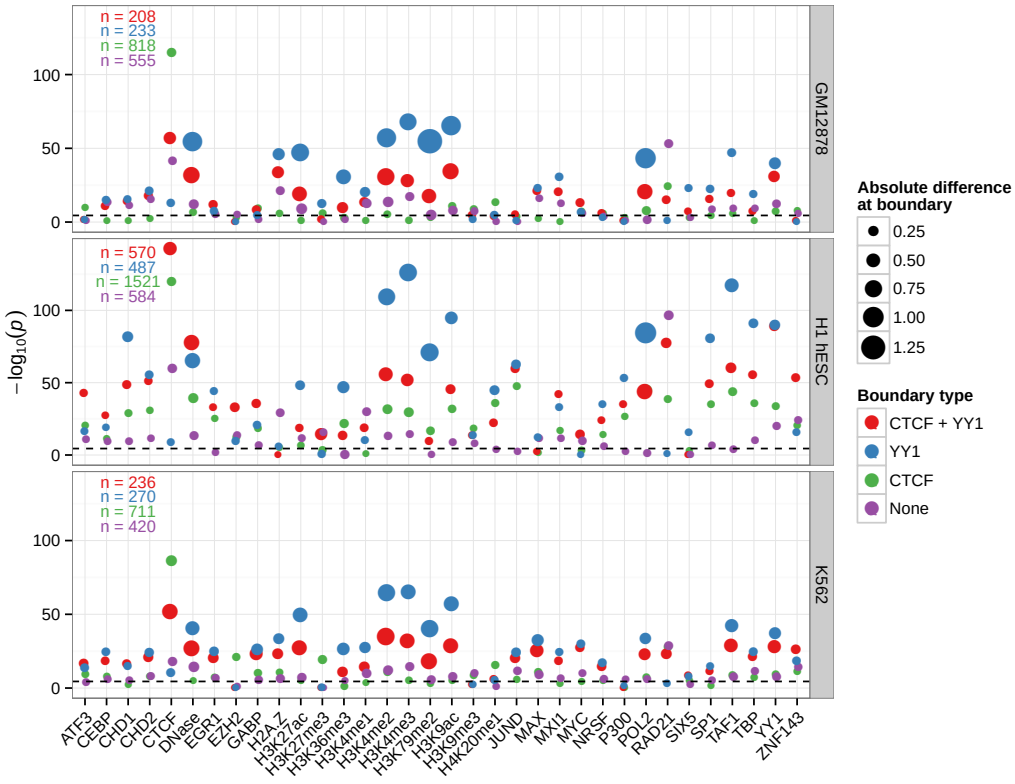

Figure 38: Distinct enrichments of CTCF and YY1 boundaries. TAD boundary feature enrichments are shown (as in Fig. 37) for boundaries split into classes based on specific enrichments: CTCF and YY1 groups are those boundaries with at least one ENCODE region peak[36] for their respective features, while CTCF + YY1 is the group of boundaries which had one or more overlapping peaks for these two factors. Boundaries in the none group has neither a CTCF or YY1 region peak called (but can still be enriched for their respective features in terms of raw signal).

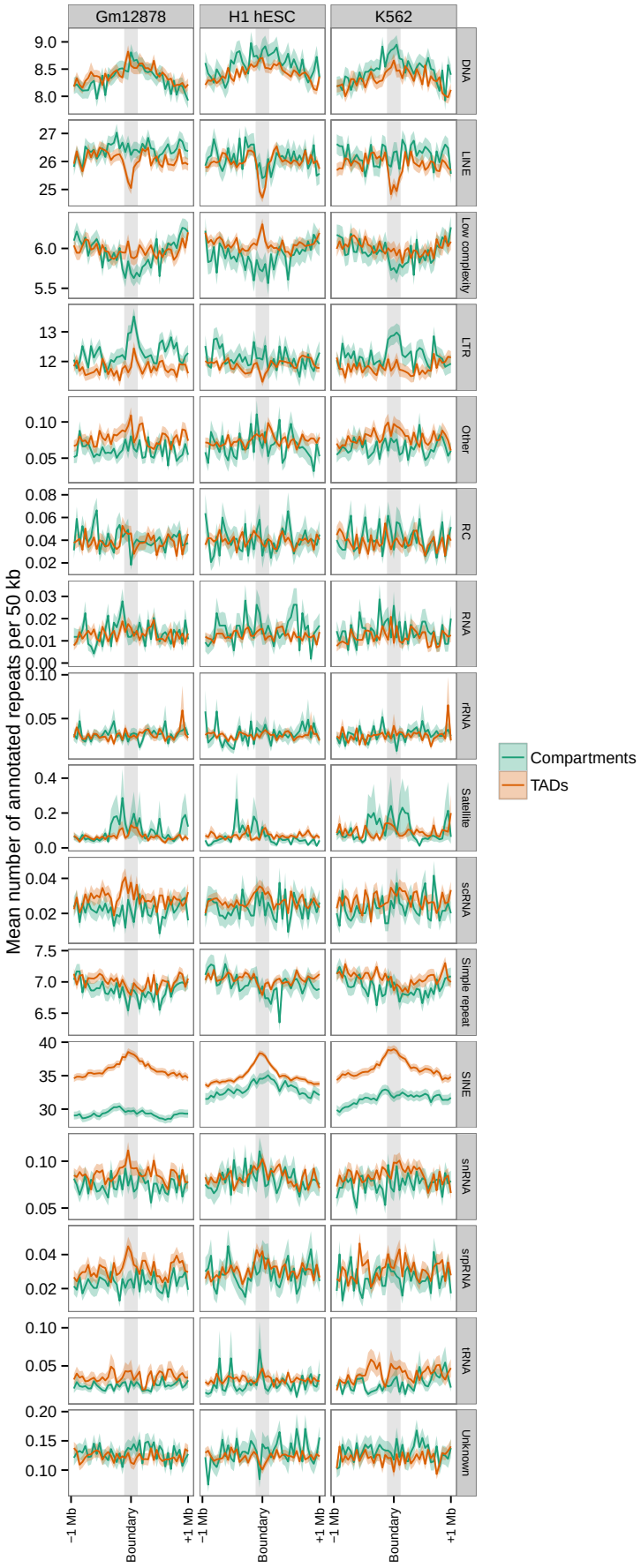

Figure 39: Repeat class average-o-grams over all TAD and compartment boundaries. RepeatMasker repeat annotations are counted per 50 kb for 1 Mb either side of each TAD and compartment boundaries. The mean count genome-wide is plotted with ±95% confidence intervals.

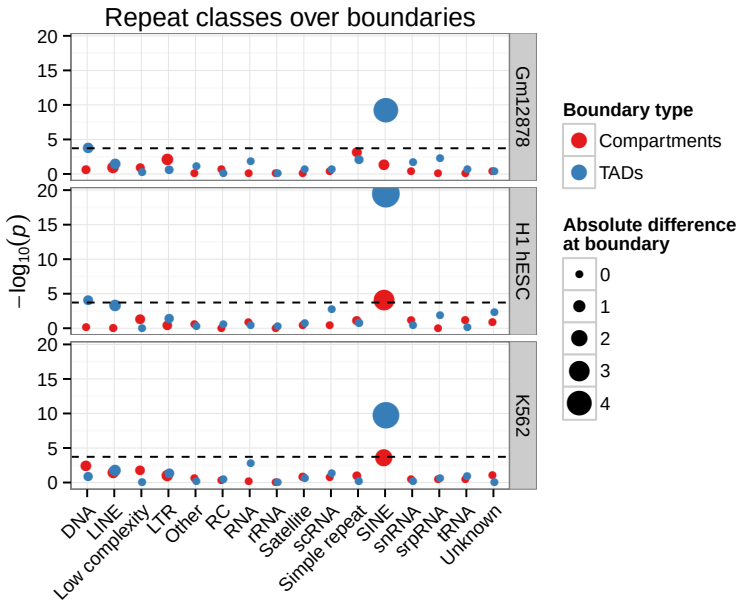

Figure 40: Significance and effect sizes of repeat class enrichments/depletions over boundaries. Boundary profiles (Fig. 39) were tested for enrichment or depletion of each factor at the boundary bin relative to peripheral non-boundary bins (see Methods XX). Bubble area is proportional to the raw effect size of an enrichment or depletion. The Bonferroni-corrected significance threshold is highlighted with a dashed line.

At the level of repeat class, we corroborate the findings of Dixon *et al.*[8] 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. 39). 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. 40).

Repeat class profiles also suggest LINEs may be depleted over TAD boundaries and DNA repeats may be enriched at both boundary types (Fig. 39), however statistically these observations do not surpass our pre-defined significance threshold ($\alpha = 0.05$) after multiple testing correction (Fig. 40).

Repeat classes can be broken into smaller repeat families. Dixon *et al.*[8] reported that the Alu (or B1 in mouse) repeat families are enriched over TAD boundaries. Again we can reproduce this finding and extend it to compartment boundaries (Fig. 41).

5.3 DE NOVO BOUNDARY PREDICTION

We have shown TAD and compartment domain boundaries to be well-marked by a variety of features. Compartment boundaries are successfully predicted as a side-effect of modelling the continuous compartment profile eigenvector (Section XX) however a related measure of activity and repression does not exist for TADs.

We attempted to model TAD boundaries in a variety of ways: firstly a using a class-balanced classification framework and secondly through indirect models of directionality index and the downstream domain-caller HMM state.[8]

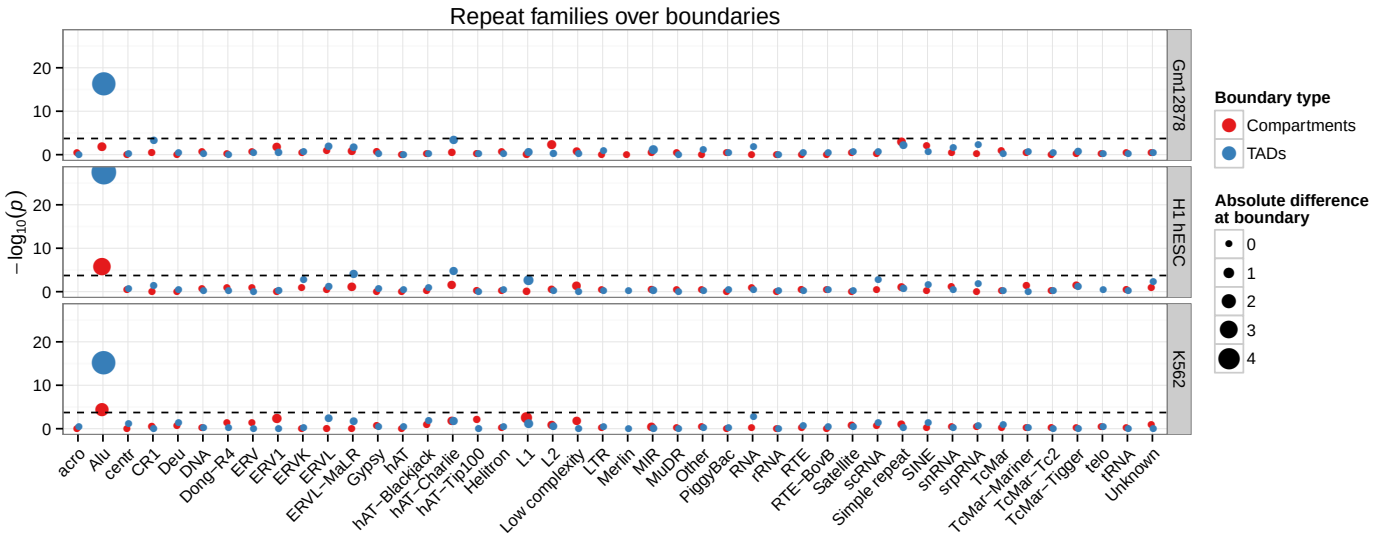

Figure 41: Significance and effect sizes of repeat family enrichments/depletions over boundaries. As per Fig. 40 but for a higher resolution repeat classification.

5.4 METATAD BOUNDARIES

Our collaborators uncovered the concept of "metaTADs": sequential aggregations of adjacent and strongly-interacting TADs to form a hierarchy of domain organisation covering each chromosome.

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

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 5.2). A hypothesis to test could be that boundaries of larger metaTAD structures could display greater enrichments for boundary-defining features.

5.4.1 Lamin associated domains

5.4.2 Boundaries over a time series

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 are effectively abolished in metaphase.[21] This raises the question of how the structural organization 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.[56] These G-band patterns are constant over human cell types at metaphase, but all traces of

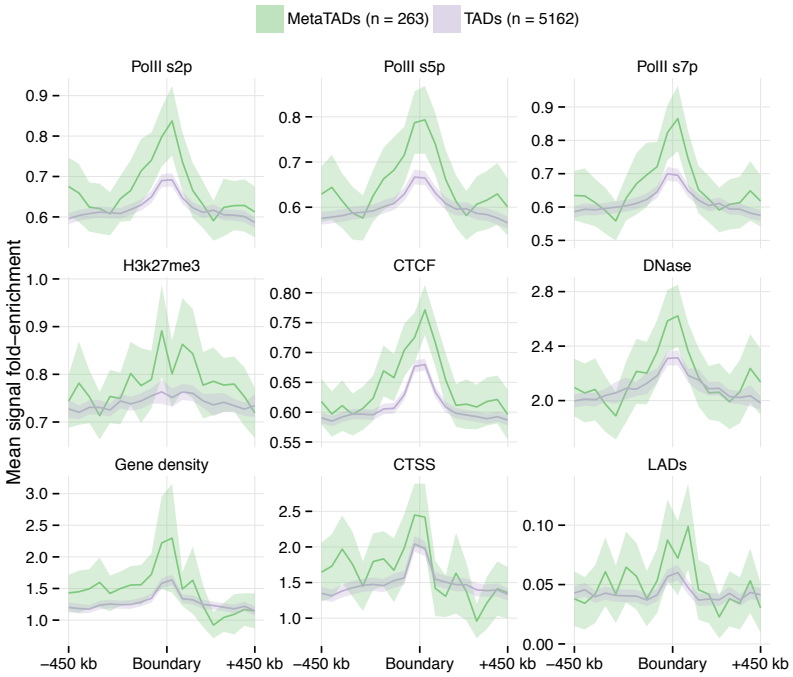

Figure 42: Large metaTADs show greater enrichments for an array of boundary features. Genome-wide profiles of epigenomic features and gene densities averaged over all TAD and metaTAD (10 – 40 Mb) boundaries (ribbons show 95% confidence intervals of the mean).

interphase higher order structure were reported to be absent at metaphase.[21] We would therefore expect no agreement between metaphase G-bands and the patterns of interphase TADs and A/B nuclear compartments defined here, over all three cell types.

We examined the genome wide similarity of all interphase domain structure boundaries to metaphase G-band boundaries, relative to an expected distribution derived by permutation (see Methods) (Figure S9). There is a significant, though extremely modest, excess of compartment boundaries within close proximity of G-band boundaries, such that 13.90% of compartment boundaries are within 500 kb of a G-band boundary (expectation = 10.50%, K-S test: $D = 0.076$, $p < 3 \times 10^{-12}$). This is seen for compartment boundaries calculated for all three cell types independently. The genome wide overlap of compartment A and B regions with particular G-band classes is nonrandom, and suggests much greater correspondence. 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. Approximately 40% of the genome jointly occupies interphase compartment A as well as gneg/gpos25 metaphase G-bands, or occupies the interphase B compartment as well as gpos75/gpos100 at metaphase. Again, the same trends are seen significantly across all three cell types. This agreement is not unexpected given the broad differences in G-negative and G-positive bands, with contrasting gene density, GC content and replication timing[56] that is strongly reminiscent of the contrasts between interphase A and B compartments,[7] but to our knowledge has not been directly studied before. These data suggest that across the genome most fine structure, reflected in domain boundaries, is not well preserved between interphase and metaphase. However there is evidence for conservation of broader structural categories across a substantial fraction of the genome, which may reflect broad similarities in the degree of compaction seen at many regions across the cell cycle.

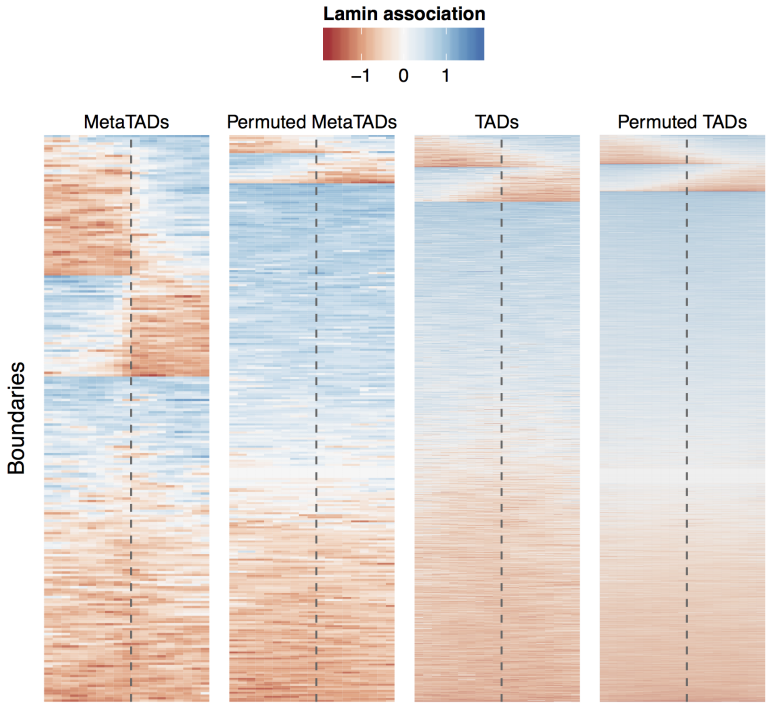

Figure 43: MetaTADs align with lain associated domains. Heatmaps of LaminB1 association microarray probe intensity values over MetaTAD boundaries (from domains of size 10 – 40 Mb) and TAD boundaries, are displayed alongside examples of circularly-permuted boundaries. 42.6% of MetaTAD boundaries (10 – 40 Mb) had an absolute linear regression coefficient $> .05$ of lamin association intensities, indicating a boundary transition (versus 15.8% expectation from 1000 circular permutations, $p < 1 \times 10^{-4}$). TAD boundaries were also significantly more associated with lamin association transitions (Observed: 11.8%, Expected: 9.5%; empirical p-value: $p < 1 \times 10^{-4}$). Profiles are shown $\pm$450 kb from each boundary.

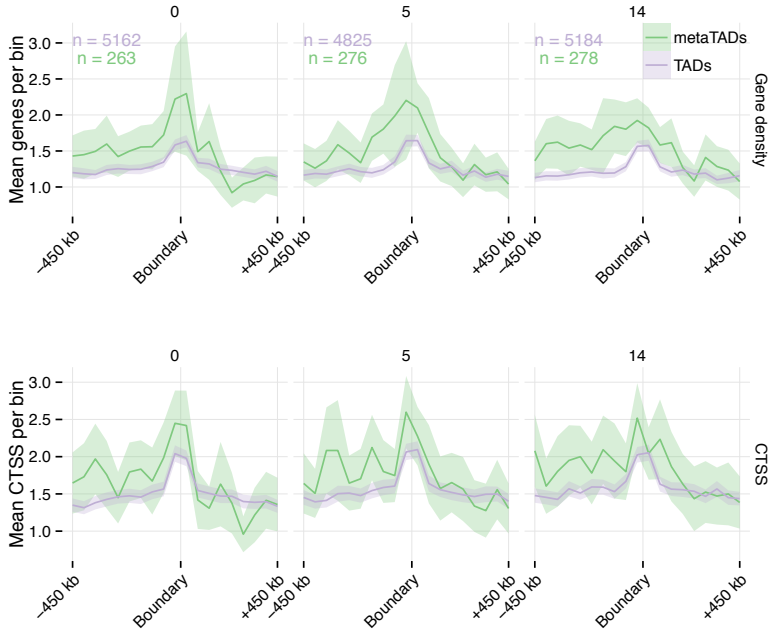

Figure 44: Observed enrichments persist over a time series. CAGE-defined active TSS (CTSS) were counted per 50 kb bin across each TAD and MetaTAD (10 – 40 Mb) boundary and averaged (ribbons show 95% confidence intervals of the mean). Gene densities refer to mean counts of annotated genes per bin, with an overlap of at least 250 bp. Peak heights suggest modestly stronger enrichments at MetaTAD boundaries relative to TAD boundaries.

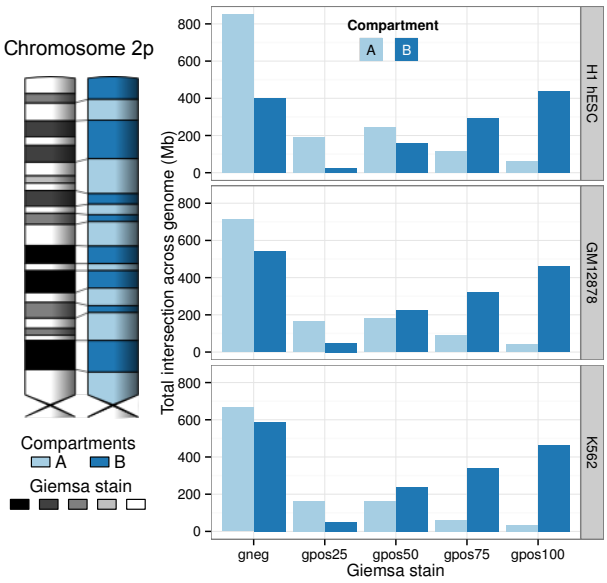

Figure 45: Giemsa–stain bands correspond to A/B compartments. Placeholder

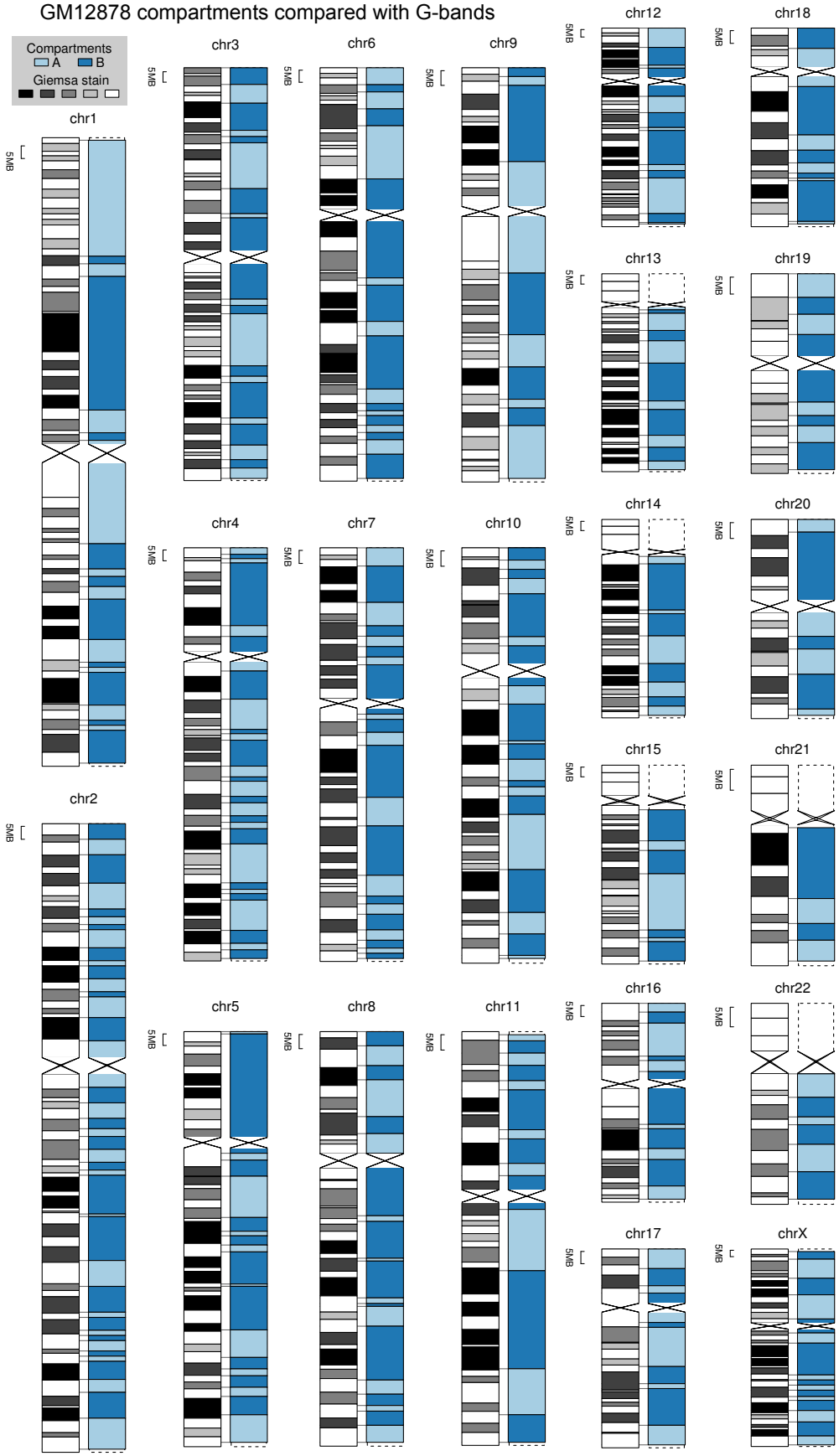

Figure 46: Genome-wide agreement between Giemsa bands and A/B compartments in the Gm12878 cell type. Placeholder

5.5.2 Superboundaries

Thus far compartment and TAD boundaries have been considered separately, however it is of interest to consider how these boundary regions interact across scales. Open questions remain about the co-occurence of these two boundary regions, and whether

6 LOCAL CHROMATIN CONFORMATION

6.1 INTRODUCTION

The Hi-C assay provides a genome-wide overview of chromatin conformation, however this broad scope imposes resolution limits inherent to an all-vs-all assay. 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 two collaborative projects involving the use of 4C and 5C data to "zoom in" on two well-studied regions related to limb development: the ZRS enhancer and HoxD gene cluster.

6.2 CHROMATIN CONFORMATION AT THE SHH LOCUS

Anterior-posterior patterning in the developing limb is regulated in mammals by *Sonic hedgehog* (SHH).[142] Specifically, the SHH gene is expressed within a confined region named the "zone of polarising activity". Its expression within this region is known to be regulated by a well-studied enhancer, the "zone of polarising activity regulatory sequence" or ZRS.[143] ZRS is located almost 1 Mb downstream of its target SHH promoter in humans, and is located in intronic regions of another gene, LMBR1, and is conserved across mammals and fish (Fig. 47).[143,144] Single point mutations and short insertions within this enhancer have been linked to various limb deformities, including pre- and post-axial polydactyly.[142,144,145] 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.[145]

Collaborators have developed a model system which allows inducible SHH expression in a non-expressing 14fp cell line derived from the developing limb bud. Application of trichostatin A (TSA) then leads to detectable SHH expression, and increased levels 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 and SHH expression is blocked through other means. Analysis of the region through FISH implies similar levels of compaction in SHH expressing and non-expressing cells (*data not shown*), suggesting the latter explanation.

My part in this collaboration was to analyse 3C-seq (also known as 4C) data recorded by our collaborators for the SHH–ZRS region in mouse. Additionally, the 4C procedure[61] was adapted for specific in-house sequencing instruments (Ion Torrent Ion Proton™ sequenced as opposed to Illumina™ technology) and as such required diagnostics to confirm the experimental data was accurate.

6.2.1 Analysis of ZRS interactions

4C experiments were performed by collaborators using the ZRS region as a bait sequence, or "viewpoint", such that it contacts were measured with all other HindIII restriction fragments genome-wide. 4C was performed in both

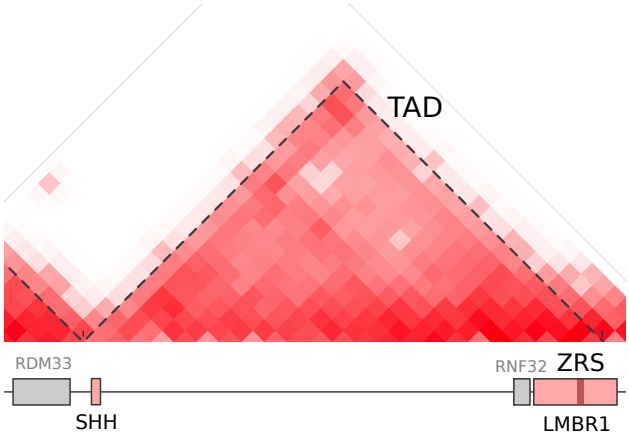

Figure 47: SHH–ZRS contacts occur within a stable TAD. An approximately 1 Mb region of the mouse genome is shown below a Hi-C contact map (derived from previously published data[8]). A clear TAD can be identified spanning from SHH to ZRS, dashed lines show TAD boundaries called by Dixon *et al.*[8]. This figure was generated for Anderson *et al.*[141].

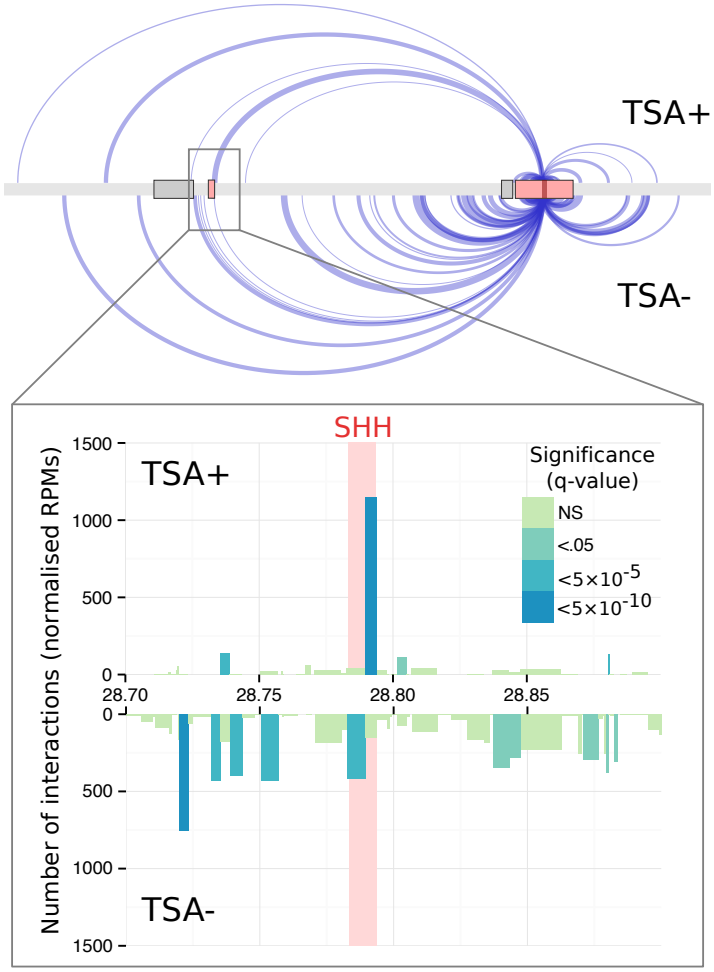

Figure 48: TSA treatment induces a strong ZRS–SHH interaction. 4C interactions are shown as edges from source node (ZRS enhancer bait fragment) to targets along an approximately 2 Mb region of chromosome 5. Edge width is proportional to the number of interactions, only highly significant interactions are shown (FDR q-value $< 5 \times 10^{-5}$). Zoomed region shows the number of interactions of the bait region with SHH in both treated and untreated samples. Each rectangle is a restriction fragment, coloured by FDR q-value indicating the significance of the interaction above expected levels.

untreated and non-SHH expressing cells (*TSA*–) and in cells treated with TSA, thereby causing SHH expression (*TSA*+).

The first stage in analysing these contacts is to convert observed raw sequencing reads to normalised frequencies (Methods 2.8.1), these normalised values are then assigned significance scores in the form of q-values, with the aim of finding those significantly over-represented relative to expectation (Methods 2.8.2).

6.2.2 4C / Hi-C comparison

Hi-C data in mouse cells has been previously published,[8] so can be compared with this novel 4C data to give broader contextual information about chromatin conformation in the region under study.

6.2.3 Assay diagnostics

The 4C protocol used by our collaborators in this work was that of Stadhouders *et al.*[61]. In it, the authors advise some statistical tests to ensure the quality of the experiment results. Among these were:[61]

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.

6.2.4 3D modelling with 5C data

All-vs-all contacts measured either genome-wide in the case of Hi-C, or over a defined region with 5C, can be used to infer the trajectory of chromatin fibres in three-dimensions through a variety of methods (e.g.[146–150]). 5C data was generated over this same SHH–ZRS region (Fig. 47) 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 a particular three-dimensional inference program (`AutoChrom3D`[150]) in an attempt to compare polymer trajectories in TSA treated and untreated 14fp mouse cells.

6.3 5C IN THE HOXD REGION

HoxD is another well-studied genetic system involved in limb development and under the control of known enhancers. In this experiment, our collaborators were interested in the chromatin conformations of HoxD13 loci in both the anterior and posterior developing limb bud, particularly how and where the two differed. To this end, our collaborators performed 5C for two biological replicates in anterior and posterior limb bud cell lines, and my contribution was to call differential contacts between the two conditions.

6.3.1 Differential contacts

6.3.2 5C / Hi-C comparison

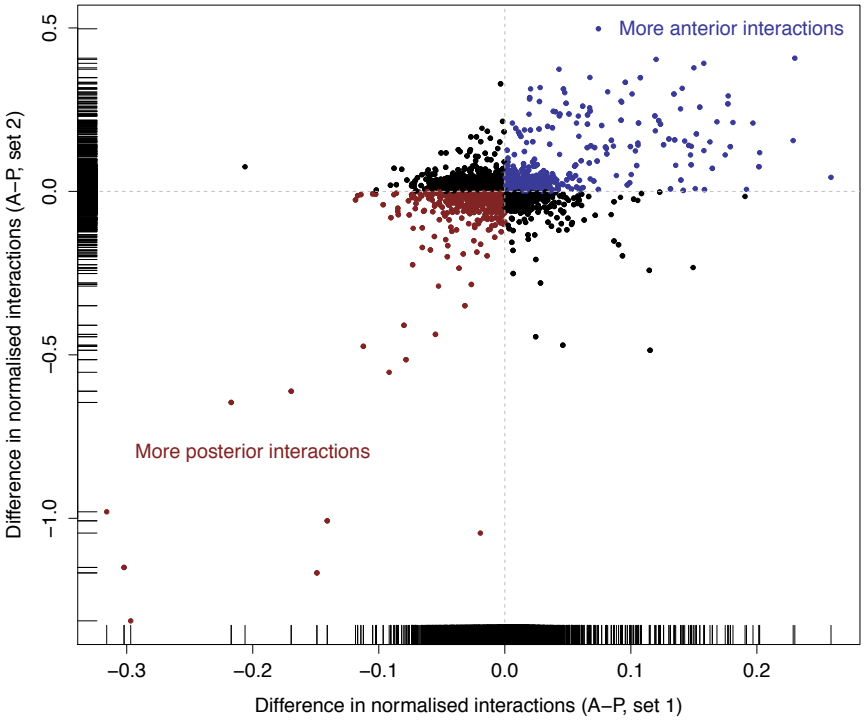

Figure 49: Raw differences between anterior and posterior 5C interactions. Placeholder

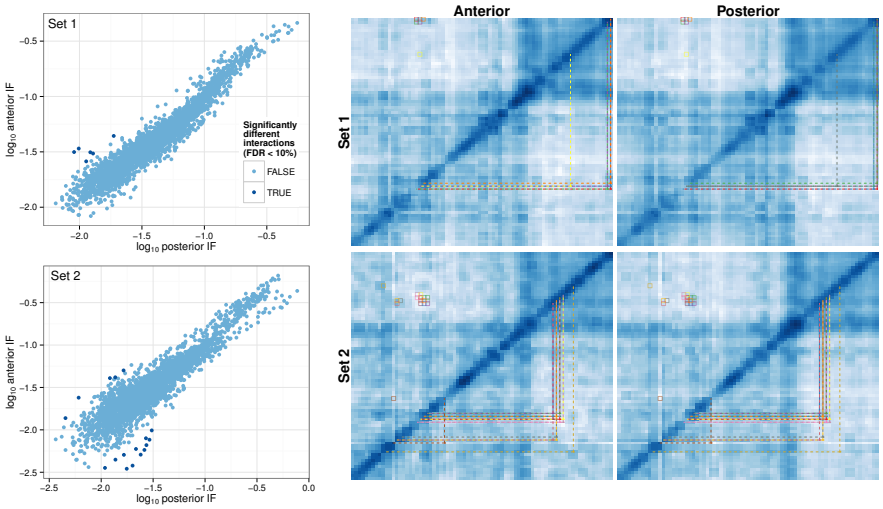

Figure 50: Will we use this stuff? Placeholder

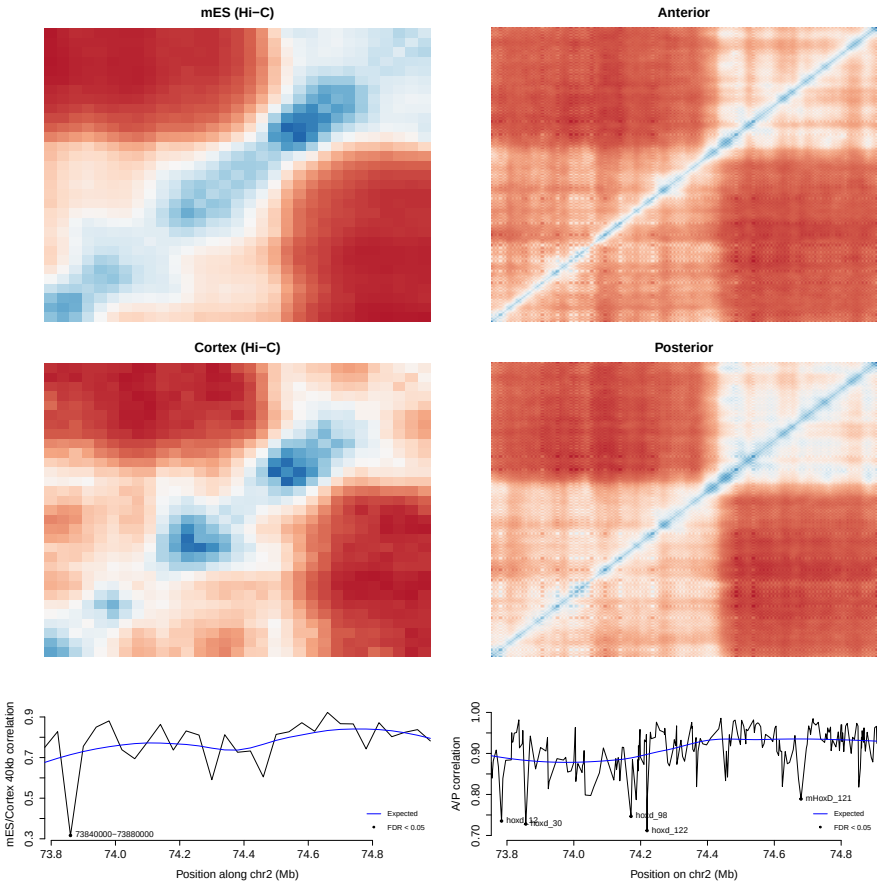

Figure 51: Will we use this stuff? Placeholder

7 DISCUSSION

The recent abundance of epigenomic data in model cell types has enabled accurate modelling of the transcriptional output of human promoters, and a rigorously quantitative assessment of the most influential chromatin features underlying gene expression[94]. 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-analysed, from the intitial 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[8,58], we observed good concordance of higher order chromatin structure, relected in Hi-C data, between different cell types. Random forest models summarised 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[33,36,53,151] but these states typically emcompass 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, deliniating variation in replication timing, lamin association and nuclear compartments derived from Hi-C eigenvectors[58]. The data presented here therefore suggest that a variety of such domains could be successfully modelled. Given the fact 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[152]. While the patterns of repressive H3K9me3 accumulation have been a focus in the cancer literature and have been proposed as a diagnostic marker in leukemia[153]. Similarly, the model for GM12878 (Epstein-Barr virus transformed lymphoblasoid) cells shows a disproportionate influence of ATF3 binding patterns, and ATF3 induction is a known consequence of virus transformed cells[154]. 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[8,58], 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 influentual variables for the random forest models constructed for the two hematopoietic cell lines was the ubiquitous transcription factor YY1, which re-appeared 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 100-500 Kb). This is intriguing as YY1 has recently been shown to co-localise with the architectural protein CTCF[155] 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 (e.g.[124]). 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.

7.1 CONCLUSION

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 modelling approaches, and the models produced can 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 embryonic stem cells and H3K9me3 in the leukaemia 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 modelling of large chromatin dataset collections using random forests

can generate useful insights into chromosome structure and seed testable hypotheses for further experimental studies.

7.2 FUTURE RESEARCH

APPENDICES

Table A1: Gm12878 functional enrichments in regions of variable structure.

Category	Term	Count	%	Fold Enrichment	*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.

Category	Term	Count	%	Fold Enrichment	*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.

Category	Term	Count	%	Fold Enrichment	*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

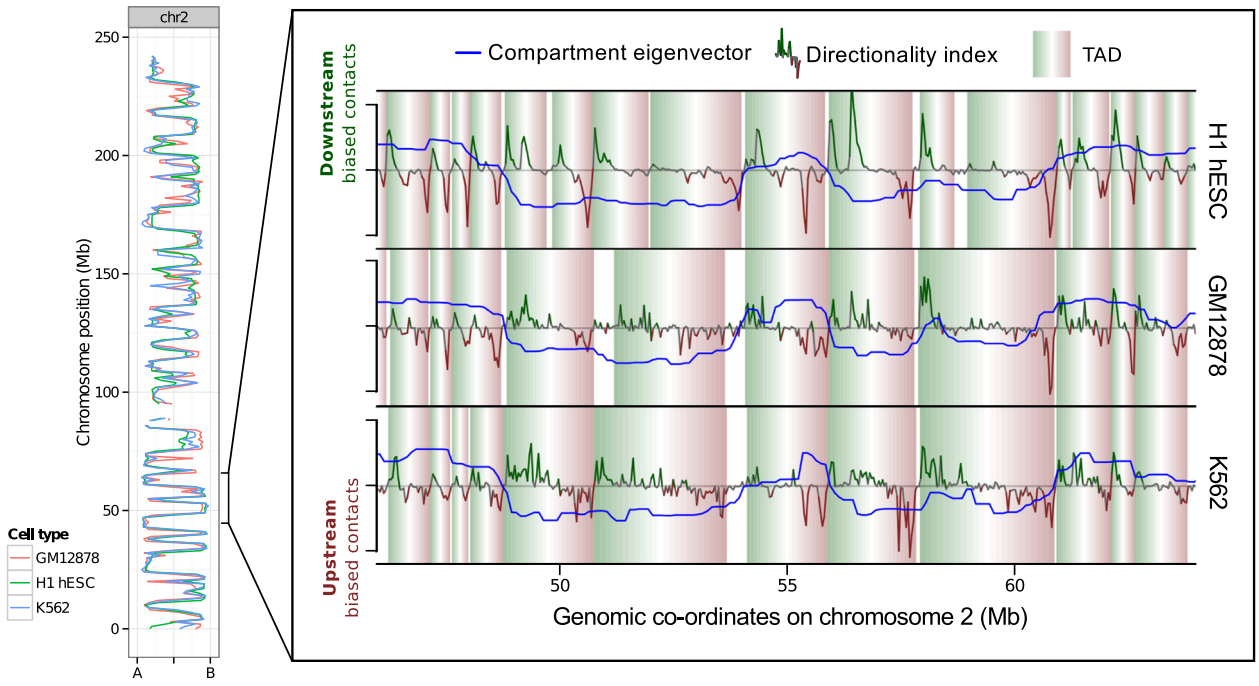

Figure 1 Concordance of chromatin structure at multiple scales over three human cell types. The eigenvector compartment profile is shown for chromosome 2 for three human cell types (left). Genome-wide Pearson correlation coefficients between eigenvectors at 1-Mb resolution are in the range 0.75 to 0.80 (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

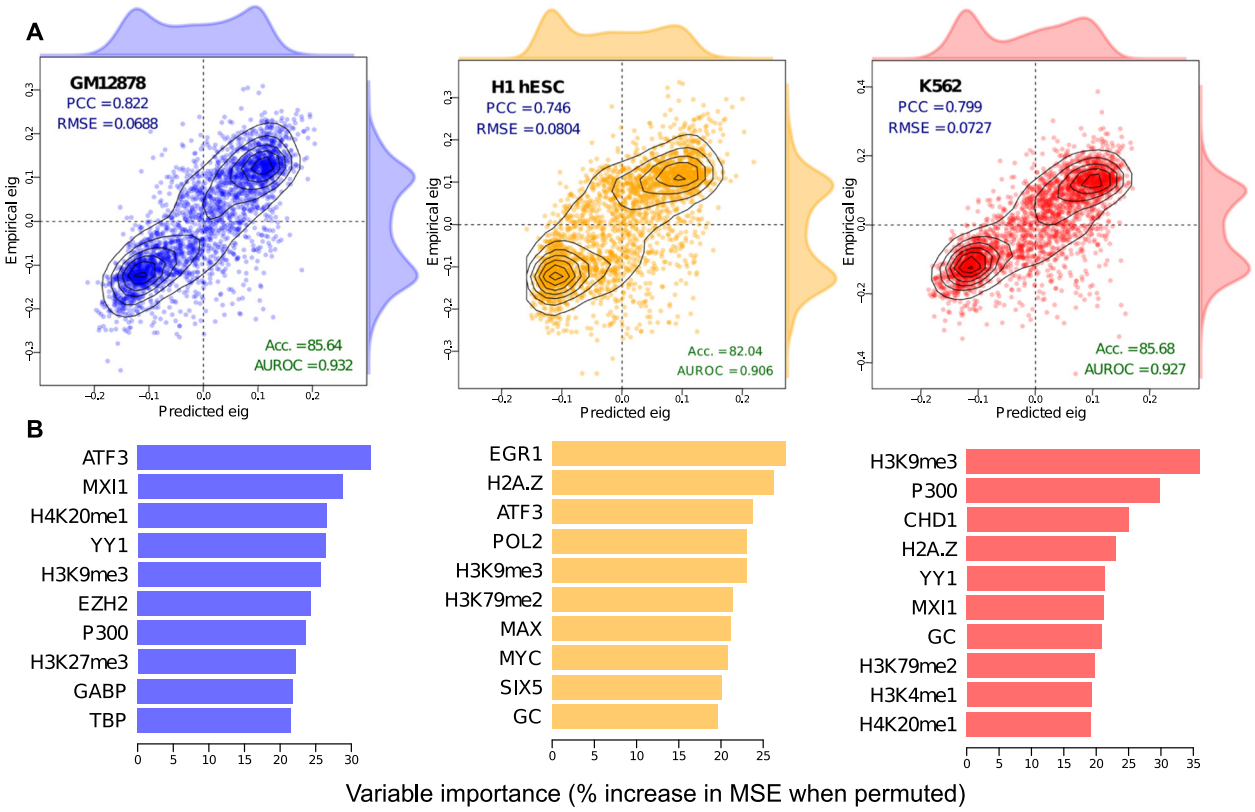

Figure 2 Accurate models of higher-order chromatin state built from locus-level features. **(A)** Model predictions (Predicted eig) are compared to observed values (Empirical eig). Various metrics are used to measure the accuracy of regression modeling – the Pearson correlation coefficient (PCC) and 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.

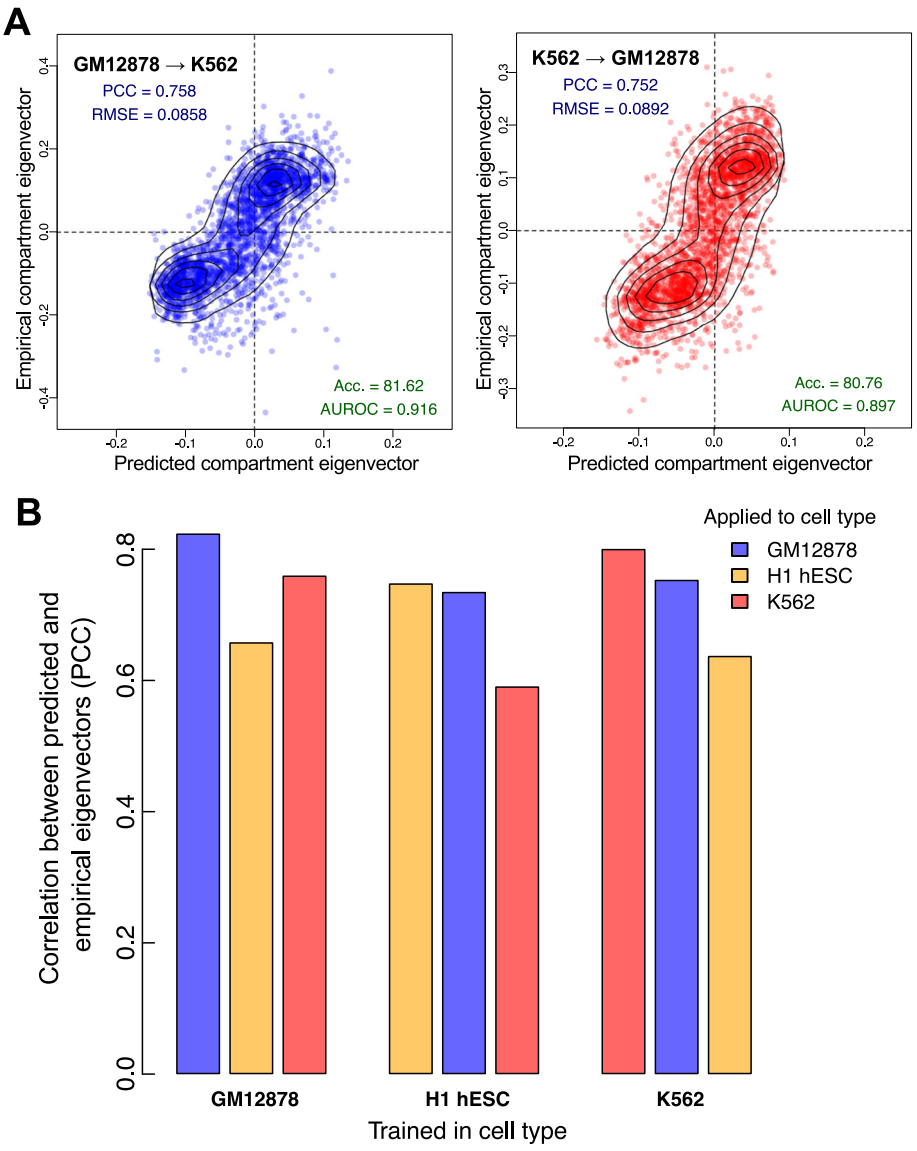

Figure 3 Models trained in one cell type can generalize to others. Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. **(A)** The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross (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

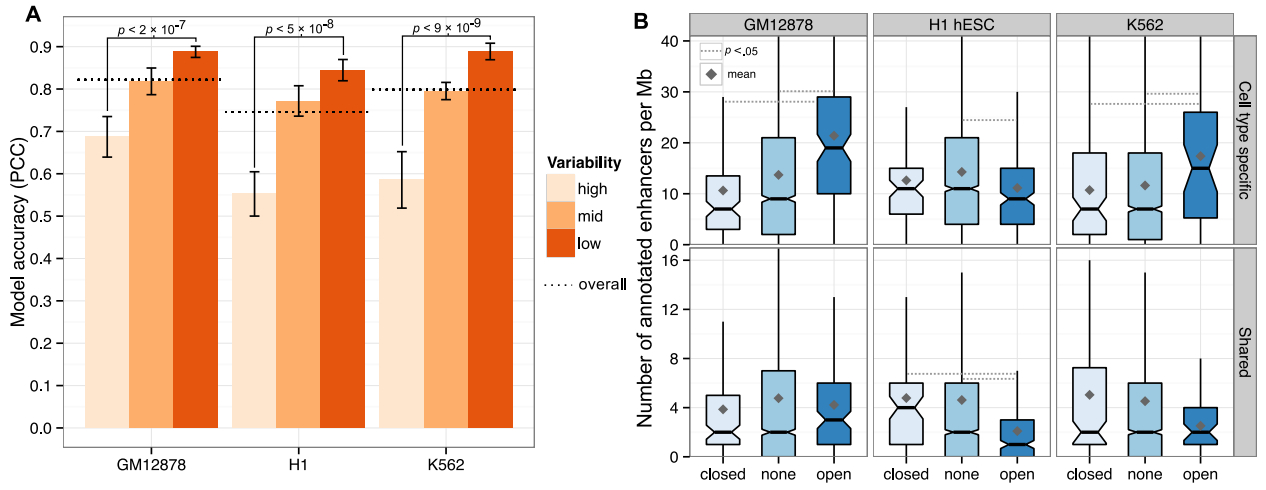

Figure 4 Regions with variable structure are less successfully modeled and are associated with cell-type-specific enhancer activity. **(A)** Model accuracy is significantly different between low- and high-variability regions, defined as, respectively, the lowest and highest thirds of the distribution 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

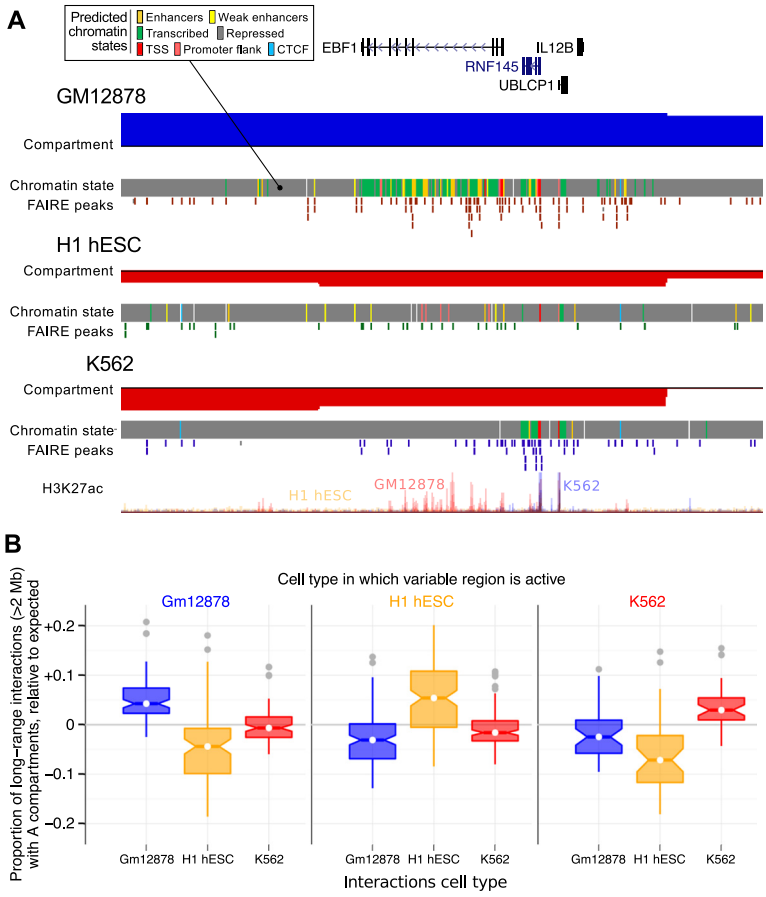

Figure 5 Structurally variable regions indicate cell-type-specific biology. Regions occupying the active A nuclear compartment in one cell type, but the repressed B compartment in the other two, were selected and ranked by the number of predicted active enhancers (Figure 4). **(A)** The region chr5:158–159 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

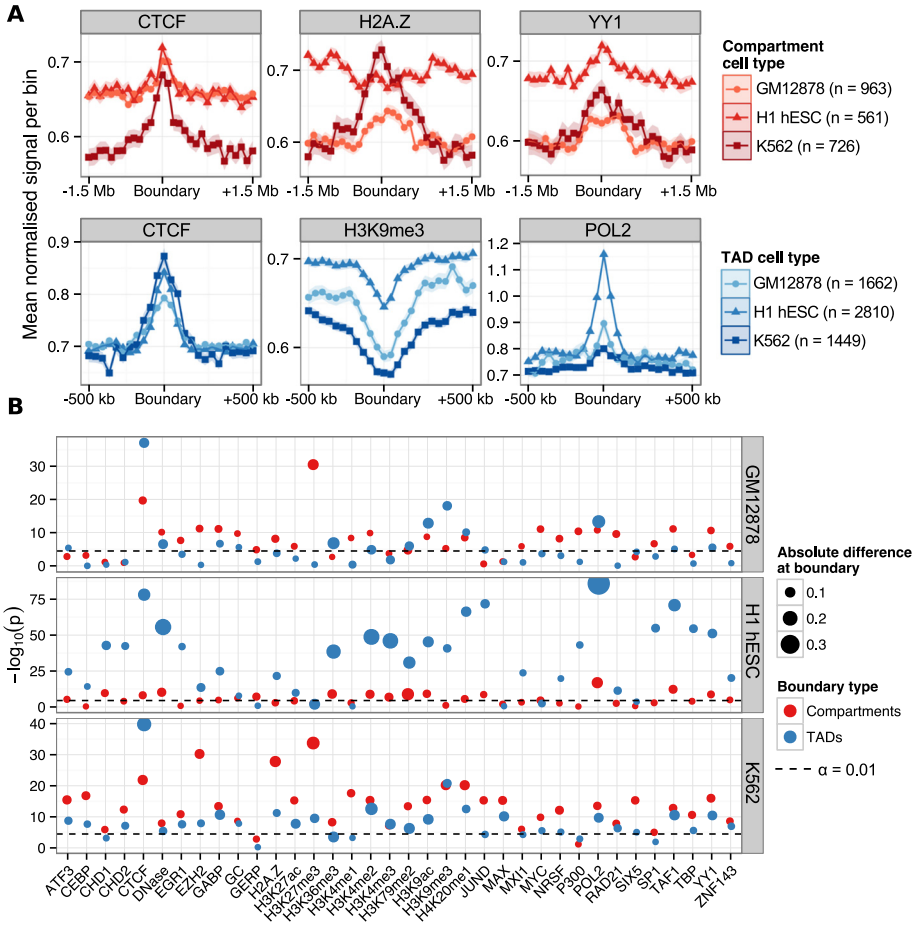

Figure 6 Chromatin features underlying TAD and compartment boundaries. **(A)** Selected profiles for locus-level features are shown for TAD boundaries (CTCF, H3K9me3 and POL2) and compartment boundaries (H2A.Z, H3K4me2 and YY1), as a mean normalized ChIP-seq signal relative to input chromatin per bin (±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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[37] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall Ka, Phillippy KH, Sherman PM, *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, **41**(Database issue): D991–5.

[38] Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. *Nucleic acids research*, **39**(Database issue): D19–21.

[39] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4): 357–9.

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

[41] Zhang Y, Liu T, Meyer Ca, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**(9): R137.

[42] Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)*, **22**(12): 1540–2.

[43] Breiman L (2001) Random Forests. *Machine learning*, **45**(1): 5–32.

[44] Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News*, **2**(December): 18–22.

[45] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, **43**(6): 1947–58.

[46] Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology*, **88**(11): 2783–92.

[47] Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, **7**: 3.

[48] Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD (2011) Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic epidemiology*, **35 Suppl 1**(Suppl 1): S5–11.

[49] Breiman L (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3): 199–231.

[50] Hastie T, Tibshirani R, Friedman J (2009) *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition. ISBN 978-0-387-84858-7.

[51] Tobias RD (1995) An Introduction to Partial Least Squares Regression. *Proc. Ann. SAS Users Group Int. Conf. 20th*, pp. 1250–1257.

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

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

[54] Ashburner M, Ball Ca, Blake Ja, Botstein D, Butler H, Cherry JM, Davis aP, Dolinski K, Dwight SS, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**(1): 25–29.

[55] Huang BDW, Lempicki R (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, (301): 1–43.

[56] Furey TS (2003) Integration of the cytogenetic map with the draft human genome sequence. *Human Molecular Genetics*, **12**(9): 1037–1044.

[57] Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis Ja, Bickmore Wa (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, **10**(3): 211–9.

[58] Chambers EV, Bickmore Wa, Semple CA (2013) Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS computational biology*, **9**(4): e1003017.

[59] de Wit E, Bouwman BaM, Zhu Y, Klous P, Splinter E, Verstegen MJaM, Krijger PHL, Festuccia N, Nora EP, *et al.* (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, pp. 1–7.

[60] Tanabe H, Müller S, Neusser M, von Hase J, Calcagno E, Cremer M, Solovei I, Cremer C, Cremer T (2002) Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(7): 4424–9.

[61] Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, Palstra RJ, Wendt KS, Grosveld F, *et al.* (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols*, **8**(3): 509–24.

[62] Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B (2013) R3Cseq: An R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Research*, **41**(13): 1–12.

[63] Gentleman RC, Gentleman RC, Carey VJ, Carey VJ, Bates DM, Bates DM, Bolstad B, Bolstad B, Dettling M, *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10): R80.

[64] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Publishing Group*, **12**(2): 115–121.

[65] Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics.

[66] Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome biology*, **10**(7): R79.

[67] Ay F, Bailey TL, Noble WS (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research*.

[68] Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **66**(1): 187–205.

[69] Storey J (2015) *qvalue: Q-value estimation for false discovery rate control*. R package version 2.0.0.

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

[71] Splinter E, de Wit E, van de Werken HJG, Klous P, de Laat W (2012) Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods*, **58**(3): 221–230.

[72] Gao F, Wei Z, Lu W, Wang K (2013) Comparative analysis of 4C-Seq data generated from enzyme-based and sonication-based methods. *BMC genomics*, **14**(1): 345.

[73] Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539): 331–336.

[74] Wickham H (2009) *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6.

[75] Wickham H, Francois R (2015) *dplyr: A Grammar of Data Manipulation*. R package version 0.4.2.

[76] Van Rossum G (1995) *Python reference manual*.

[77] Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6): 841–842.

[78] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16): 2078–2079.

[79] Kent WJ, Zweig aS, Barber G, Hinrichs aS, Karolchik D (2010) BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*, **26**(17): 2204–2207.

[80] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The Human Genome Browser at UCSC The Human Genome Browser at UCSC. *Genome Research*, pp. 996–1006.

[81] Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita Pa, Wang T, Nguyen N, Paten B, Zweig AS, *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**(7): 1003–1005.

[82] Kuhn RM, Haussler D, James Kent W (2013) The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, **14**(2): 144–161.

[83] Hu G, Cui K, Northrup D, Liu C, Wang C, Tang Q, Ge K, Levens D, Crane-Robinson C, Zhao K (2013) H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, **12**(2): 180–192.

[84] Li Z, Gadue P, Chen K, Jiao Y, Tuteja G, Schug J, Li W, Kaestner KH (2012) Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell*, **151**(7): 1608–1616.

[85] Huang DW, Sherman BT, Lempicki Ra (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**(1): 1–13.

[86] Lozzio BB, Lozzio CB, Bamberger EG, Feliu AS (1981) A multipotential leukemia cell line (k-562) of human origin. *Experimental Biology and Medicine*, **166**(4): 546–550.

[87] Bickmore Wa (2013) The spatial organization of the human genome. *Annual review of genomics and human genetics*, **14**: 67–84.

[88] Bernstein BE, Stamatoyannopoulos Ja, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra Ma, Beaudet AL, *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**(10): 1045–8.

[89] Nikolov M, Fischle W (2012) Systematic analysis of histone modification readout. *Molecular bioSystems*, **Advance Ac**.

[90] Sajan SA, Hawkins RD (2012) Methods for identifying higher-order chromatin structure. *Annual review of genomics and human genetics*, **13**: 59–82.

[91] Henikoff S, Shilatifard A (2011) Histone modification: cause or cog? *Trends in genetics : TIG*, **27**(10): 389–96.

[92] Tippmann SC, Ivanek R, Gaidatzis D, Schöler A, Hoerner L, van Nimwegen E, Stadler PF, Stadler MB, Schübeler D (2012) Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Molecular systems biology*, **8**(593): 593.

[93] McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL (2012) Genome-wide in silico prediction of gene expression. *Bioinformatics*, **28**(21): 2789–96.

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

[95] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26): 15776–81.

[96] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, *et al.* (2006) CAGE: cap analysis of gene expression. *Nature methods*, **3**(3): 211–22.

[97] Zuber V, Strimmer K (2011) High-Dimensional Regression and Variable Selection Using CAR Scores. *Statistical Applications in Genetics and Molecular Biology*, **10**(1): 1–27.

[98] Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, **20**(6): 761–70.

[99] RIKEN Omics Science Center (2012) FANTOM5. *http://fantom.gsc.riken.jp/*.

[100] Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics*, **41**(3): 376–81.

[101] Schaft D (2003) The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation. *Nucleic Acids Research*, **31**(10): 2475–2482.

[102] Breiman L (2001) Random forests. *Machine learning*, **45**: 5–32.

[103] Karlić R, Chung Hr, Lasserre J, Vlahovicek K, Vingron M (2010) Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(7): 2926–31.

[104] Beringer M, Ballar C, Croce LD, Viz P (2015) Role of PRC2-associated factors in stem cells and disease. **282**: 1723–1735.

[105] Creyghton MP, Markoulaki S, Levine SS, Hanna J, Lodato Ma, Sha K, Young Ra, Jaenisch R, Boyer La (2008) H2AZ Is Enriched at Polycomb Complex Target Genes in ES Cells and Is Necessary for Lineage Commitment. *Cell*, **135**(4): 649–661.

[106] Deb G, Singh AK, Gupta S (2014) EZH2: Not EZHY (Easy) to Deal. *Molecular cancer research : MCR*, **12**(5): 639–53.

[107] Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, *et al.* (2012) c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, **151**(1): 68–79.

[108] Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, Resch W, Baek S, Pruett N, *et al.* (2013) Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, **155**(7): 1507–20.

[109] Zervos AS, Gyuris J, Brent R (1993) Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell*, **72**(2): 223–232.

[110] Sun XJ, Man N, Tan Y, Nimer SD, Wang L (2015) The Role of Histone Acetyltransferases in Normal and Malignant Hematopoiesis. *Frontiers in Oncology*, **5**(May): 1–11.

[111] Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics*, **43**(7): 630–8.

[112] Abdi H, Williams LJ (2010) Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(4): 433–459.

[113] Mantel N (1970) Why Stepdown Procedures in Variable Selection. *Technometrics*, **12**(3): 621–625.

[114] Hurvich CM, Tsai Cl (1990) The Impact of Model Selection on Inference in Linear Regression. *The American Statistician*, **44**(3): 214.

[115] Tibshirani R (1994) Regression Selection and Shrinkage via the Lasso.

[116] Guyon I, Weston J, Barnhill S, Vapnik V (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**: 1157–1182.

[117] Kohavi R, Kohavi R (1997) Wrappers for feature subset selection. *Artificial Intelligence*, **97**(1-2): 273–324.

[118] Deng H, Runger G (2012) Feature selection via regularized trees. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, pp. 1–8.

[119] Deng H (2013) Guided Random Forest in the RRF Package. *arXiv*, pp. 1–2.

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

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

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

[123] Nora EP, Dekker J, Heard E (2013) Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? *BioEssays*, **35**(9): 818–828.

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

[125] Lupiáñez D, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz J, *et al.* (2015) Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, pp. 1012–1025.

[126] Ren B, Dixon J (2015) A CRISPR Connection between Chromatin Topology and Genetic Disorders. *Cell*, **161**(5): 955–957.

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

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

[129] Evertts AG, Manning AL, Wang X, Dyson NJ, Garcia Ba, Coller Ha (2013) H4K20 methylation regulates quiescence and chromatin compaction. *Molecular biology of the cell*, **24**(19): 3025–37.

[130] Onder TT, Kara N, Cherry A, Sinha AU, Zhu N, Bernt KM, Cahan P, Marcarci BO, Unternaehrer J, *et al.* (2012) Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, **483**(7391): 598–602.

[131] Chai X, Nagarajan S, Kim K, Lee K, Choi JK (2013) Regulation of the boundaries of accessible chromatin. *PLoS genetics*, **9**(9): e1003778.

[132] Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H, Carr SA, Bernstein BE (2012) H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome biology*, **13**(10): R85.

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

[134] Atchison ML (2014) Function of YY1 in Long-Distance DNA Interactions. *Frontiers in immunology*, **5**(February): 45.

[135] Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, Flicek P (2013) Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome biology*, **14**(12): R148.

[136] Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, *et al.* (2013) Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome research*, **23**(12): 2066–77.

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

[138] Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, García-Díaz A, *et al.* (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science (New York, N.Y.)*, **317**(5835): 248–251.

[139] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**(1-2): 335–348.

[140] Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, (SUPPL. 25).

[141] Anderson E, Devenney PS, Hill RE, Lettice La (2014) Mapping the Shh long-range regulatory domain. *Development (Cambridge, England)*, (September): 1–10.

[142] Anderson E, Peluso S, Lettice La, Hill RE (2012) Human limb abnormalities caused by disruption of hedgehog signaling. *Trends in Genetics*, **28**(8): 364–373.

[143] Hill RE, Lettice La, B PTRS (2013) Alterations to the remote control of Shh gene expression cause congenital abnormalities Alterations to the remote control of Shh gene expression cause congenital abnormalities Author for correspondence :. (May).

[144] Laurell T, Vandermeer JE, Wenger AM, Grigelioniene G, Nordenskjöld A, Arner M, Ekblom AG, Bejerano G, Ahituv N, Nordgren A (2012) A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. *Human Mutation*, **33**(7): 1063–1066.

[145] Lettice La, Hill AE, Devenney PS, Hill RE (2008) Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Human Molecular Genetics*, **17**(7): 978–985.

[146] Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS (2013) Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, **9**(1): e1002893.

[147] Varoquaux N, Ay F, Noble WS, Vert Jp (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics (Oxford, England)*, **30**(12): i26–i33.

[148] Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J (2014) 3D genome reconstruction from chromosomal contacts. *Nature methods*, (september).

[149] Trieu T, Cheng J (2014) Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research*, **42**(7): 1–11.

[150] Peng C, Fu LY, Dong PF, Deng ZL, Li JX, Wang XT, Zhang HY (2013) The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Research*, **41**(19).

[151] Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, *et al.* (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, **147**(7): 1628–39.

[152] Zwang Y, Oren M, Yarden Y (2012) Consistency test of the cell cycle: roles for p53 and EGR1. *Cancer research*, **72**(5): 1051–4.

[153] Müller-Tidow C, Klein HU, Hascher A, Isken F, Tickenbrock L, Thoennissen N, Agrawal-Singh S, Tschanter P, Disselhoff C, *et al.* (2010) Profiling of histone H3 lysine 9 trimethylation levels predicts transcription factor activity and survival in acute myeloid leukemia. *Blood*, **116**(18): 3564–71.

[154] Hagmeyer BM, Duyndam MC, Angel P, de Groot RP, Verlaan M, Elferich P, van der Eb A, Zantema A (1996) Altered AP-1/ATF complexes in adenovirus-E1-transformed cells due to EIA-dependent induction of ATF3. *Oncogene*, **12**: 1025–1032.

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