Unravelling higher order genome organisation [working title]

Benjamin L. Moore

July 6, 2015

INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

CONTENTS

1 INTRODUCTION 3
- 1.1 Genome organisation 3
  - 1.1.1 C-methods and Hi-C 3
  - 1.1.2 Hi-C variants 4
  - 1.1.3 Chromosome compartments 5
  - 1.1.4 Topological domains 5
  - 1.1.5 Other proposed structures 7
- 1.2 Models of chromatin folding 8
  - 1.2.1 Fractal globule 8
  - 1.2.2 Strings and binders switch 8
  - 1.2.3 Looping 8
  - 1.2.4 Cell cycle changes 8
- 1.3 Criticisms of C-methods 9
- 1.4 Machine learning in genomics 9
  - 1.4.1 ENCODE 9
- 1.5 Aims 10

2 METHODS 11
- 2.1 Hi-C data 11
  - 2.1.1 Mapping 11
  - 2.1.2 Filtering 11
  - 2.1.3 Correction 11
  - 2.1.4 Eigenvector calculation 12
- 2.2 ENCODE features 12
  - 2.2.1 Clustering input features 12
- 2.3 Modelling 13
  - 2.3.1 Random Forest 13
  - 2.3.2 Model performance 13
  - 2.3.3 Other modelling approaches 14
  - 2.3.4 Graphical lasso 14
- 2.4 Variable regions 15
  - 2.4.1 Stratification by variability 15
  - 2.4.2 Enhancer enrichment 15
- 2.5 Boundaries 15
  - 2.5.1 TADs 15
  - 2.5.2 Compartments 16
  - 2.5.3 MetaTADs 16
- 2.6 Giemsa band comparison 16
- 2.7 Nuclear positioning 17
- 2.8 Gene ontology analysis 17

3 REANALYSIS OF HI-C DATASETS 18
- 3.1 Introduction 18
- 3.2 Hi-C reprocessing 18
- 3.3 Compartment profiles 18
- 3.4 Domain calls 18
- 3.5 Variable regions 21
- 3.6 Nuclear positioning 21

4 INTEGRATIVE MODELLING AS A TOOL TO EXPLORE BIOLOGICAL SYSTEMS 22
- 4.1 Introduction 22
- 4.2 Reproducing Dong *et al.* 23
- 4.3 Modelling FANTOM5 CAGE timecourse data 24

4.3.1 Dissecting the *best bin* approach 27
4.4 Modelling higher order chromatin 27
4.4.1 Predictive model 27
4.4.2 Cross-application 27
4.4.3 Variable importance 29
4.4.4 Importance of resolution 29
4.4.5 Other modelling approaches 29
4.4.6 Non-independence 29
4.4.7 Correlating input features 29

5 CHROMATIN DOMAIN BOUNDARIES 32
5.1 Introduction 32
5.2 TAD and compartment boundaries 33
5.2.1 CTCF and YY1 33
5.2.2 Repeats 33
5.3 De novo boundary prediction 33
5.4 MetaTAD boundaries 33
5.5 Other boundaries 33
5.5.1 Giemsa bands 33
5.5.2 Superboundaries 33

6 4C AND 5C ANALYSIS 34
6.1 Introduction 34
6.2 4C of the ZRS enhancer 34
6.2.1 3D modelling 34
6.3 5C in the HoxD region 34

7 DISCUSSION 35
7.1 Conclusion 36

8 APPENDICES 38

1 INTRODUCTION

1.1 GENOME ORGANISATION

It's oft-stated that the DNA within each human cell would extend for two metres fully extended. Instead that same length of DNA packs into a cell nucleus with a diameter in the order of micrometers (μm). This is achieved through a complex organisation hierarchy, ranging from how chromosomes are arranged in the nucleus, down to how DNA is wrapped around nucleosomes.[1]

Briefly, DNA exists mostly as a left-handed double-helix of hydrogen bonded purines and pyrimidines. These in turn are wrapped around histone octamers, proteins with tuneable DNA packing properties. These wrapped histones can be visualised as "beads on a string" in transcriptionally active regions, and possibly as a more-compact 30 nanometre fibre, though this is disputed.[2]

1.1.1 C-methods and Hi-C

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

With the advent DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[5] was the first sequencing-based method of 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[6,7] (both named 4C), whereby interactions were measured for a specific "viewpoint" fragment against all other restriction fragments genome-wide. The same year a many-to-many assay (5C) allowed measurements for all restriction fragments within a specified region.[8]

The final step was an all-versus-all assay, capable of assaying pairwise interaction frequencies between all restriction fragments of a genome. This assay was published by Lieberman Aiden *et al.*[9] and named Hi-C (Fig. 1). The Hi-C method added biotin tagging to pull-down only ligated fragments

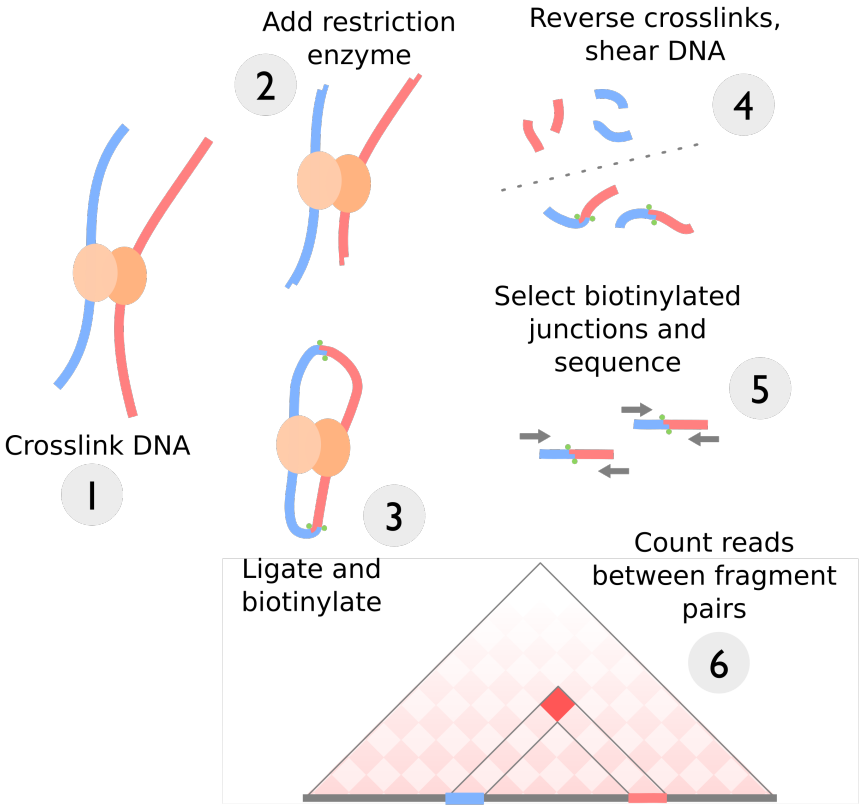

Figure 1: Steps in the Hi-C assay. Schematic of the Hi-C experimental procedure as described in Lieberman Aiden *et al.*[9]

for sequencing. At the time of publication, resolution of Hi-C data for analysis was limited by sequencing depth, given the huge number of restriction fragments produced by a 6-cutter enzyme (HindIII and NcoI were used in[9]) but the falling costs of sequencing and proven utility of the assay meant subsequent Hi-C papers incrementally increased their sequencing depth, to a point where analysis could be performed at the level of individual restriction fragments, genome-wide.[10–13]

1.1.2 Hi-C variants

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

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

Hi-C is a population-level assay, as the retrieved interaction counts are from a huge number of different cells. As well as building population-averaged models of genome structure, it is also of interest to probe cell-to-cell variability through single-cell approaches. For instance, it's been estimated

that long-range contacts identified with C-methods may occur in as few as 10% of cells at any one time.[4]

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

Capture-C is 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.[22] 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.[9]

Use of a cell population also averages away cell-cycle effects, with the vast majority of results coming from cells during interphase (around 97%[2]). Naumova *et al.*[2] 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.[13] The principle difference is that fixation and ligation now happen in place, within intact cell nuclei.

1.1.3 Chromosome compartments

2.5.2

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

These compartments were identified through a continuous eigenvector profile, derived from a normalised Hi-C contact matrix.[9] Importantly, this measure holds more information than a simple two-state classification, rather the continuous values can be interpreted as relative levels of compaction or activity.[16,17]

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

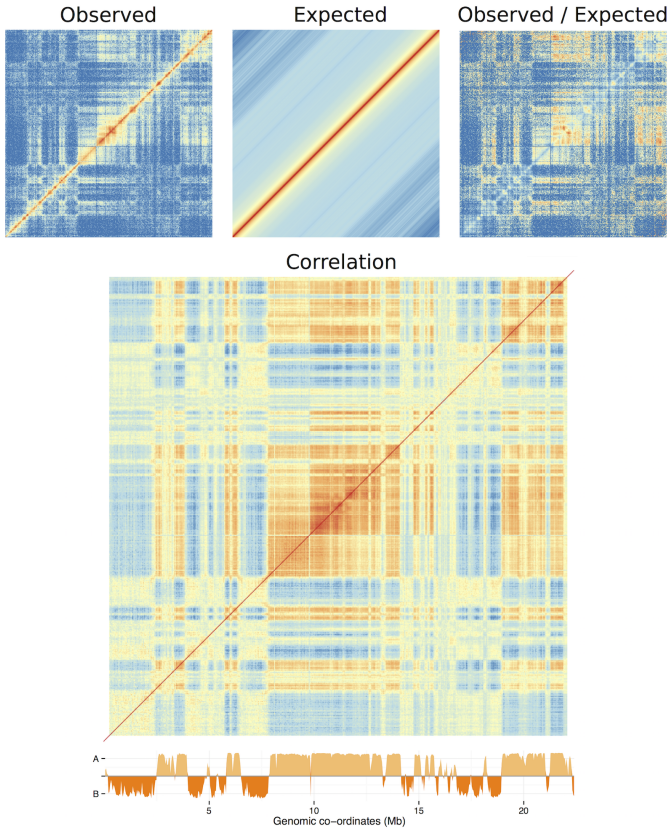

Figure 2: Derivation of A/B compartment profile from Hi-C data. Observed interaction frequencies (O) are averaged along super-diagonals to give a distance-normalised expected matrix (E). The Pearson correlation of the O/E matrix then can undergo eigenvector expansion; in most cases eigenvector $\mathbf{v}$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

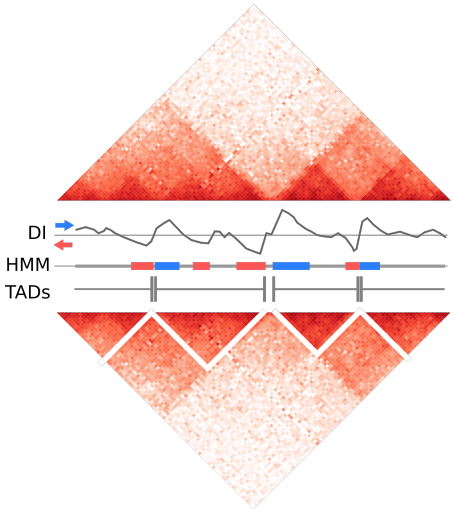

Figure 3: Dixon *et al.*[10] pipeline for calling topological associating domains (TADs). First a directionality index (DI) is calculated for each bin based on the ratio of upstream:downstream contacts. Secondly a Hidden Markov Model (HMM) is used to infer the most likely state sequence that emitted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

the level of coverage requires an exponential increase in the total amount of sequencing required.[9,23]

In experiments totalling around two billion total sequencing reads, Dixon *et al.*[10] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. 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.

Dixon *et al.*[10] 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.*[24] 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.[25]

Recent high-resolution studies have been able to resolve ever-smaller levels of sub-structure. Rao *et al.*[13] refined the concept of chromosome compartments to "sub-compartments", dividing simple A/B divisions into a total of 5 subtypes. The authors were also able to identify "contact domains" of median size 185 kb, many of which were associated with identifiable individual looping events.[13] 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[26] and the "fractal globule" model[9,27,28] have both produced simulated data that reflects empirical C-method observations and potentially describe the polymer dynamics of chromatin folding.

1.2.1 Fractal globule

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

1.2.2 Strings and binders switch

Subsequent modelling techniques integrated known biological phenomena as well as polymer models. This formed the basis of Barbieri *et al.*'s[26] "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).[29] 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[2] 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.[3] More recent studies have switched to a four-cutter restriction enzyme, for example MboI,[13] which increases this upper-bound on resolution to the order of hundreds of basepairs (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.[13]

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",[30] 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,[31] yet it remains unclear to what extent these two methods are compatible.

An additional and separate issue identified with C-methods, specifically 3C in this instance, emerges from reports that the observed ligation frequency is as low as 1% of expected values in a model system,[32] 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.[33] 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[34] algorithm which predicts states such as active promoters and enhancers, using a range of histone marks and other underlying features.[35] Similarly a Random Forest-based algorithm was developed to predict enhancers from histone modification data.[36] 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[37] combined with Hi-C genome-wide contact maps in a number of human cell types[9,10,20] 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. Dissec-

tion 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)[38] or the Short Read Archive (SRA)[39] 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[16] and `bowtie2`[40] with the `--very-sensitive` flag. An iterative mapping approach was used to maximise the number of aligning fragments.[16] 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:[16]

- 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`.[16] 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[14] 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	20
H1 hESC	331×10^6	GSE35156	10
K562	36×10^6	GSE18199	9
Cortex	373×10^6	GSE35156	10
mESC	476×10^6	GSE35156	10
IMR90	355×10^6	GSE35156	10

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).

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.*,[9] as opposed to the related eigenvectors calculated in Imakaev *et al.*[16] 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.[9,16] 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[37,41] 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[42] 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 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.[43]

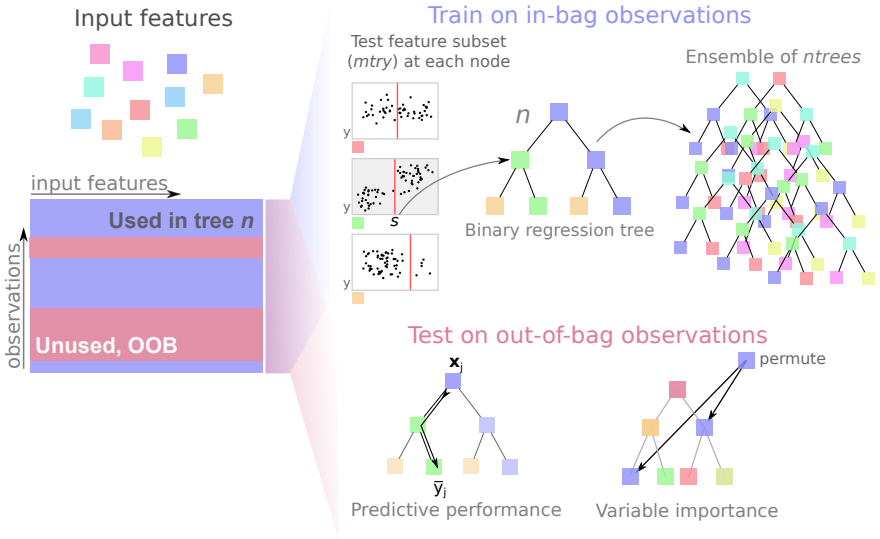

Figure 4: Random Forests overview. Random Forests are an ensemble of bagged, de-correlated classification or regression trees first described by Breiman.[44]

2.3 MODELLING

2.3.1 Random Forest

Random Forest (RF) regression,[44] was used as implemented in the R package `randomForest`.[45] 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,[46,47] typically providing low bias and low variance predictions without the need for variable selection.[48,49]

Additionally the RF method represents an example of "algorithmic modelling"[50] 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;[49,51] 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).[47,49]

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, 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 regres-

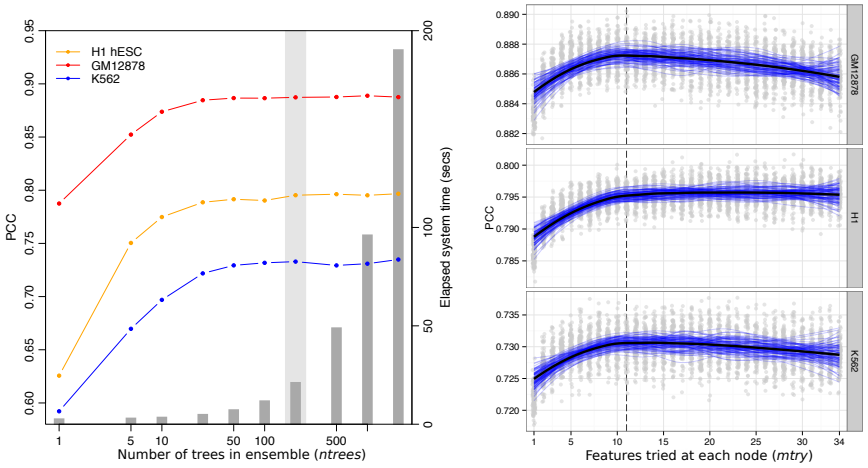

Figure 5: Confirmatory Random Forest parameter optimisation. Two user-facing Random Forest paramters are known to be insensitive over a broad range.[51] Optimisations for *ntrees* (the number of trees in the forests) and *mtry* (the number of features tested at each node) are shown for three different models, with typical values of 200 trees and $\frac{1}{3}$ of input variables highlighted.

sion 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.[52] The `plsdepot` R implementation of PLS regression was used in this work.

2.3.4 Graphical lasso

Regularised models made use of the Graphical LASSO[53] (least absolute shrinkage and selection operator) as a method of L_1-norm based regularisation, implemented via the `glasso` R package. The graphical lasso provides tuneable regularisation which is capable of feature selection via minimising regression parameters to 0. It was chosen in this case due to the multicollinearity of the featureset, the algorithm's fast speed of execution and the intuitiveness a graphical model presents.[53]

More specifically, the graphical lasso regulates the number of 0s in the inverse covariance matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, also known as the precision matrix. Then if element $\theta_{ij} = 0$, the variables X_i and X_j can be said to be conditionally

independent, given the remaining variables.[54] The algorithm minimises a negative log-likelihood (Eqn. 1 [54]) given the tuning parameter λ, which was tuned in this case to leave a small number of variables (< 10) directly dependent on the eigenvector data.

$$\underset{\boldsymbol{\Theta} \succ \mathbf{0}}{\text{minimise}}\, f(\boldsymbol{\Theta}) := -\log\det(\boldsymbol{\Theta}) + \text{tr}(\mathbf{S}\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_1 \tag{1}$$

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[35] and SegWay[55] combined annotations.[56] 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.[56]

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.

2.5 BOUNDARIES

2.5.1 TADs

TAD boundaries were called using the software provided in Dixon *et al.*[10] using their recommended parameters. For the generation of boundary

profiles, input features were averaged into 40 kb bins spanning ±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.[9]

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.[57]

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".[58] 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.*[58], 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 ($p \geq 0.1$).[58] 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[59,60] and even higher primates.[61]

2.8 GENE ONTOLOGY ANALYSIS

Variable regions (section 2.4.1) were tested for functional enrichments using Gene Ontology (GO) annotations.[62] The DAVID tool[63] was used to compare GO terms for genes located in variable compartments with a background set of genes within all annotated compartments.

3 REANALYSIS OF HI-C DATASETS

3.1 INTRODUCTION

Since the initial publication of the Hi-C technique in 2009,[9] 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 killable (kb) resolution,[9] subsequent experiments achieved first 40 kb[10], then 10 kb[12] and most recently 1 kb[13], enabling bona fide 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[9] or IMR90 against GM12878.[10] 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,[37] hence have huge amounts of matched ChIP-seq and histone modification data available.

By combinatorial reanalysis of these cell-matched datasets, I can investigate t

3.2 HI-C REPROCESSING

Each Hi-C dataset used in this work was reprocessed using the same pipeline from raw sequencing reads. In each case, experiments used the same HindIII restriction enzyme.

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. 6), 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 precise correlations of these independent measures are in the interval $R = [.75, .8]$ (Fig. **??**; Pearson correlation coefficients, PCC).

3.4 DOMAIN CALLS

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.[20]

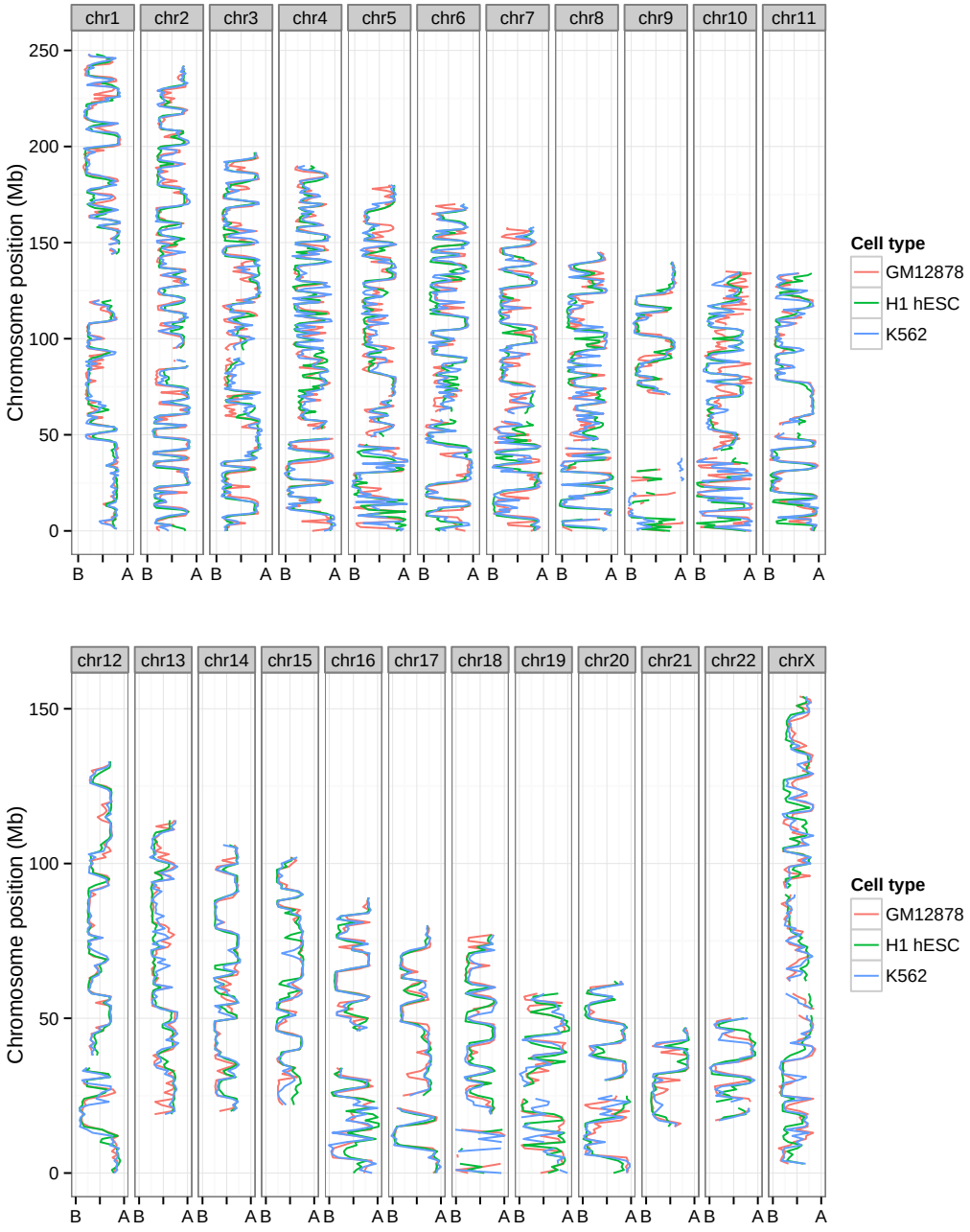

Figure 6: Compartment profiles are observably well-correlated between human cell types and across all chromosomes. Caption

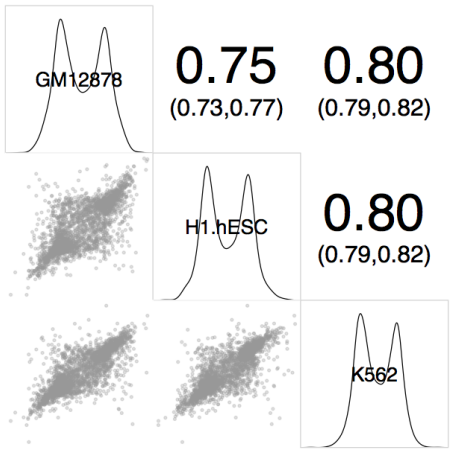

Figure 7: Compartment eigenvectors are well-correlated between human cell types Megabase resolution compartment eigenvector values are shown in a plot matrix. *Upper triangle*: Pearson correlation coefficients between pairs, with 95% confidence intervals (??); *diagonal* Kernel density estimates of eigenvector values per cell type; *lower triangle*: *x*-*y* scatterplot of values.

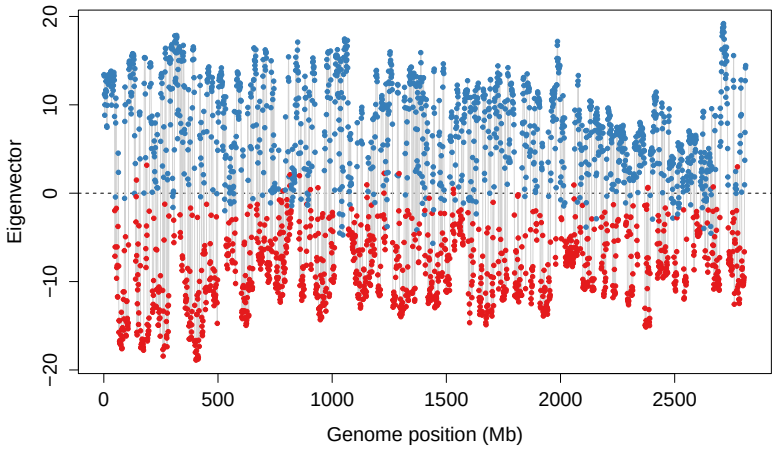

Figure 8: ?? placeholder

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 state. An examplar region is showing in Fig. 8.

3.5 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.

3.6 NUCLEAR POSITIONING

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[64] and NIH Roadmap Epigenomics[65] 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,[35,66,67] 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?"[68] 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.[68] Similarly, the interplay between locus-level factors and higher-order organisation of chromatin, while known the be an important factor in transcription, remains poorly understood mechanisatically.[69] 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.*,[70] designed a linear model to predict steady-state mRNA levels in mouse (*Mus musculus*) 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."[70] 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.[71]

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.[72] 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 (r) between predicted and empirical expression levels using 10-fold cross-validation of 0.83 across all cell lines and expression level technologies.[72] Additionally, this study highlighted cap analysis of gene expression (CAGE) as the technology, relative to RNA-Seq and RNA-PET, which produced the most

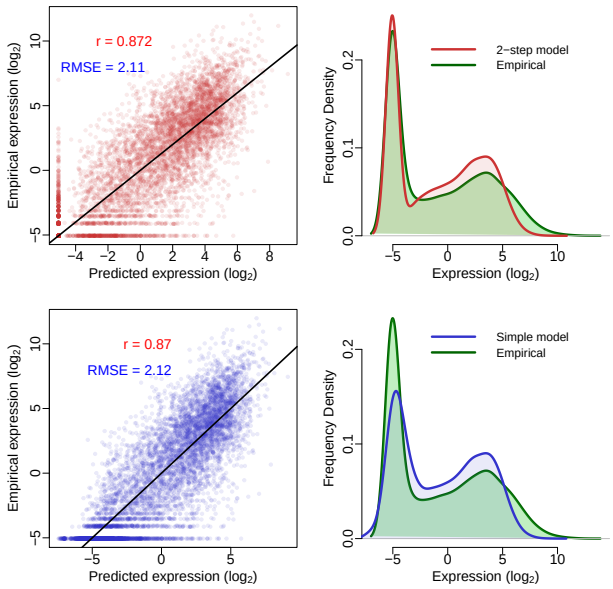

Figure 9: Comparison of classification-regression model (*upper*) with simple linear regression model (*lower*) recalculated following Dong *et al.*[72] Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for both methods are shown (*left*) along with frequency 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 $0.872 \pm 9.89 \times 10^{-5}$. All correlations were statistically significant with $p < 1 \times 10^{-15}$ under the assumption of a t-distributed r with $d.f. = 7998$.

predictable expression response. CAGE uses 5′ capped transcripts to generate short, specific tags which precisely identify TSS positions as well as quantifying the abundance of a given transcript.[73,74]

These recent publications highlight the importance and relevance of advancing our understanding of chromatin biology through a model-based approach. Each of these existing models however, treats expression levels as stationary outcome in each cell type and ignores any temporal dynamics. The huge amount of novel timecourse CAGE data produced by the FANTOM5 consortium[75] puts us in an ideal position to investigate how chromatin influences transcription beyond a simple single-point response and move towards a more complete understanding of the drivers of transcriptional flux.

4.2 REPRODUCING DONG *et al.*

Following on from Dong *et al.*,[72] I first reimplemented the published ENCODE modelling framework to ensure I could replicate their results. In doing so I was also able to analyse the strengths and caveats of their approach; 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. 9).

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 kbp around the TSS, and adds an additional bin representing the remaining gene body. Then the correlation between the signal of a given

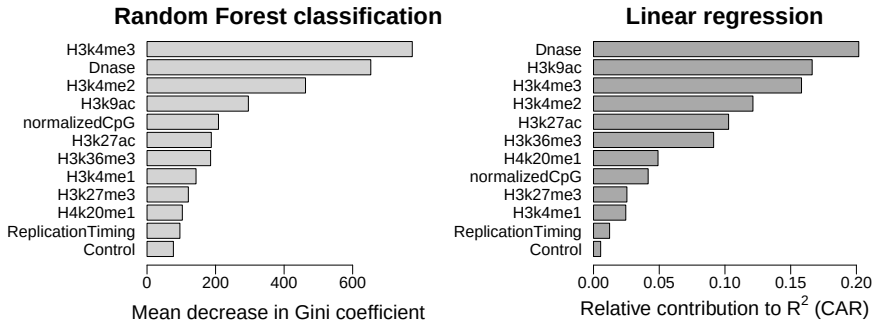

Figure 10: Relative importance metrics for variables in both the classification (*left*) and regression (*right*) stages of my reimplementation of Dong *et al.*'s two-step model.[72] The additional variable 'ReplicationTiming' shows the influence of $\log_2(early/late)$ replication timing ratio measured in the BG02 ESC cell type;[76] H1 hESC data was not available but these higher-order measurements appear to be largely conserved across cell-types.[?] For details of CAR R^2 decomposition, see Zuber and Strimmer (2010).[77]

mark and the expression of a TSS across all genes is measured — the bin producing the highest correlation is designated as the 'bestbin' and that bin's normalised ChIP-seq signal intensity in then 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.[72]

I attempted to improve the accuracy of predicted expression values produced by Dong *et al.* through two methods: increasing the number of informative regressors and increasing the complexity of the model by adding interaction terms and/or non-linear components. While Dong *et al.* included broad coverage of different histone modifications, they did not investigate the impact of higher-order chromatin data. For this reason, I matched the TSS positions used in Dong *et al.* with previously-published genome-wide replication timing ratios measured in BG02 ESCs.[76] I 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. The reasons for this are likely that the data were relatively low-resolution (1 megabase blocks), from a imperfectly matched cell line and also that the Dong *et al.* 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. However, 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. 10), implying that large-scale chromatin domains and long range interactions do not have significant influence on the expression of the genes resident within them. It would be of interest to investigate this further should more detailed higher order data become available. For example Hi-C interaction matrices have been calculated in the H1 cell line[10] and these could be compressed to principle component eigenvectors as has been done with other cell lines.[?]

4.3 MODELLING FANTOM5 CAGE TIMECOURSE DATA

Using unpublished FANTOM5 data and the approach established above, I next attempted to model gene expression at timepoint zero (t_0) of a differen-

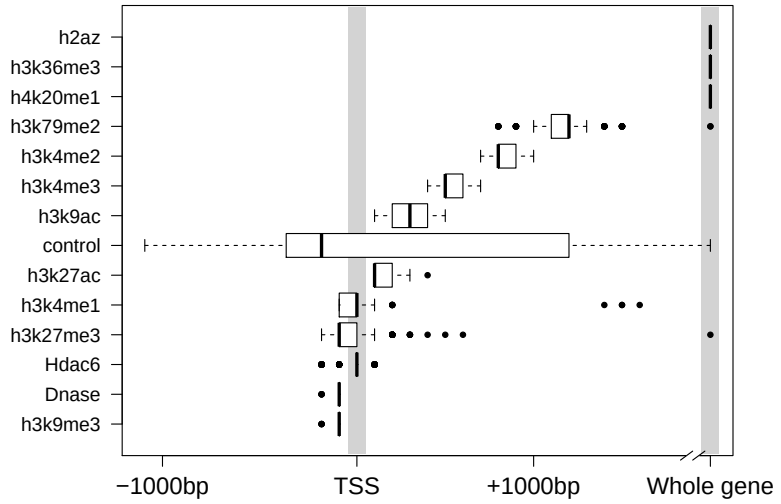

Figure 11: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, Dnase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 Kb flanking 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.

tiation 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[75] provided corresponding Entrez Gene IDs for gene-associated CAGE clusters, and I selected the most expressed cluster to represent the expression level of its mapped gene. I 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.

Next I 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[72] to objectively select the most-correlated binned signal for each chromatinH1 hESC mark. Additionally, I 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) and the result is shown in Figure 11.

This result shows that bestbin selections are often consistent, indicating there are predictably informative regions relative to a TSS for each chromatin factor (Fig. 11). 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.[70,78,79]

Having matched a variety of genome-wide H1 hESC chromatin datasets to the FANTOM5 timecourse expression data, I then built a regression model using a Random Forest (RF) approach.[80] This method outperforms a simple linear model in my initial comparisons and is able to capture non-

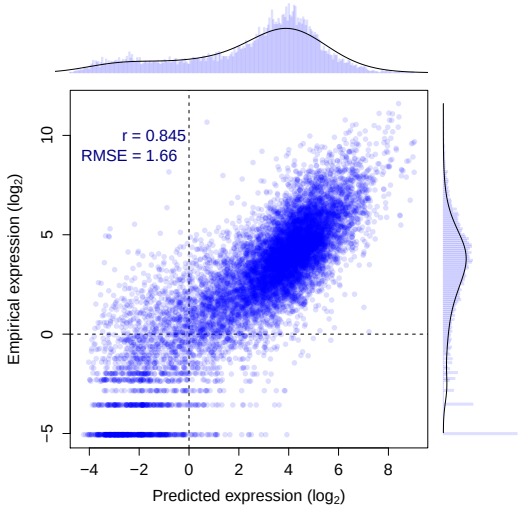

Figure 12: Evaluation of RF model predictions (x-axis) against an independent test set (y-axis). The distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson's correlation coefficient (r) and the root mean-squared error (RMSE) are also shown (*inset*).

linear relationships as well as interactions without them being explicitly specified.[48]

Figure 12 shows the resulting predictions of a preliminary RF model against the actual recorded expression over a test set of approximately 11000 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 evluated with 10-fold cross validation show a significant correlation with measured CAGE levels ($r = 0.845 \pm 1 \times 10^{-4}$; $t_{10868} = 164.4$, $p < 2 \times 10^{-15}$), and the model is able to explain around 71% of the variance in the expression response (for comparison a linear model resulted in $r = 0.825 \pm 3.2 \times 10^{-5}$; $t_{10868} = 152.2$, $p < 2 \times 10^{-15}$).

This result is worse than that of Dong *et al.* 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.[72] 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 this decrease in accuracy is that while both chromatin data and expression timecourse were measured in H1 hESC cells, the experiments took place at different institutes and likely using differing 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.[81] Additionally, Dong *et al.* implemented a pseudocount optimisation step whereby an additional count added to each binned signal intensity prior to log transformation was optimised 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. 12) 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. 12).

Having built a reasonable model of t_0 expression, the next stage of this preliminary analysis was to consider successive timepoints. In the available CD34+ differentiation dataset, this consisted of expression data recorded at three timepoints (days 0, 3 and 9—hereafter t_0, t_3 and t_9 respectively). However genome-wide expression was highly correlated between each of

these timepoints (Pearson correlation coefficients: $t_0, t_3 = 0.911$; $t_0, t_9 = 0.913$; $t_3, t_9 = 0.977$), and this high correlation meant that the genome-wide model performed essentially equally well regardless of the expression timepoint it was trained or tested on. In future analyses, higher-resolution timecourses may offer more interesting variation or alternatively genes that remain invariant throughout the timecourse could be filtered out of the dataset.

4.3.1 Dissecting the *best bin* approach

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 data assembled in this work (Chapter 1).

Previous publications have identified several correlates which track compartment eigenvector profiles to varying degrees,[9,16] 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 build Random Forest regression models (see Methods XX) to predict compartment eigenvector profiles genome-wide in three human cell types. Models were found to have high predictive accuracy, comparable to that achieved by Dong *et al.*[72] in the prediction of transcription.

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 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. 13). 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.

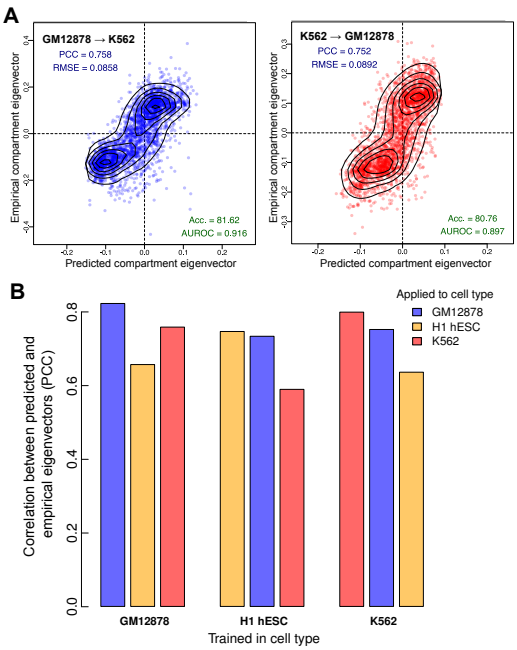

Figure 13: **Models of higher order chromatin structure learned in one cell type can be cross-applied to two others** Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. (A) The GM12878 model achieved high accuracy when applied to K562 features (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.

4.4.3 Variable importance

4.4.4 Importance of 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 resolution. 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. 14).

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.

4.4.5 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. 15), 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 imposes linear relationships between features and predictions 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

4.4.6 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.

4.4.7 Correlating input features

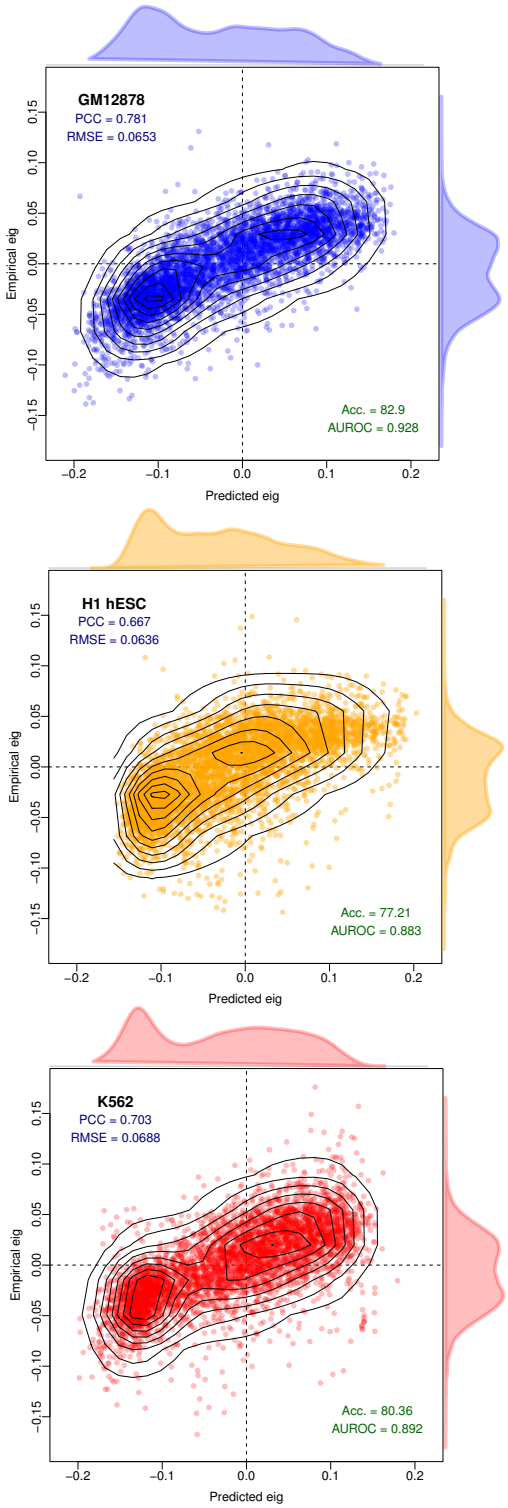

Figure 14: Models learned at 1 Mb resolution can be applied to higher resolution datasets. Despite having been trained on low resolution training sets, the Random Forest models generated can successfully predict compartment eigenvectors at higher resolution (100 kb, a 10× zoom). Eigenvectors at a higher resolution than this do not necessarily reflect A/B compartmentalisation.

Figure 15: **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). Cell type specific predictions were assessed using cross-validation (LM and PLS) and OOB estimates (RF), while cross-applied models were learnt from the full dataset of the training cell type. For PLS, the optimal number of components was also selected using cross-validation. Cell type specific models (heatmap diagonal) can be learnt with similar success via each method (mean PCC for cell type specific models: LM 0.787, RF 0.790, PLS 0.750) but on cross-application RF and PLS prove much more generalisable (mean PCC over cross-applications: LM 0.139, RF 0.689, PLS 0.641).

5 CHROMATIN DOMAIN BOUNDARIES

5.1 INTRODUCTION

Multiple studies have defined chromatin domains of different types, for example: chromosome compartments;[9] topological associating domains (TADs);[10] contact and loop domains;[13] physical domains;[25,82] and others.[24] 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 Pol2.[10] 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,[83] however an alternative theory could be that looping between enhancer elements and promoters results in an observable boundary through C-method experiments.[13] Another non-exclusive explanation is that if chromatin domains represent co-regulatory regions as is widely thought,[83–85] 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.[86] 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.[86] 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.[86]

A second experiment used CRISPR genome editing to link TAD boundary changes with limb development disorders,[87] indicating that boundary changes could provide an underlying explanation for pathogenic non-coding structural variants.[88] Similarly, domain boundaries on X-chromosomes were found to be weakened following the disruption of condensation binding sites.[89] 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.[90]

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

5.2.1 CTCF and YY1

5.2.2 Repeats

5.3 DE NOVO BOUNDARY PREDICTION

5.4 METATAD BOUNDARIES

5.5 OTHER BOUNDARIES

5.5.1 Giemsa bands

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 4C AND 5C ANALYSIS

6.1 INTRODUCTION

6.2 4C OF THE ZRS ENHANCER

6.2.1 3D modelling

6.3 5C IN THE HOXD REGION

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[72]. 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[10,59], 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[34,37,56,91] 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[59]. 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[92]. 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[93]. 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[94]. 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[10,59], 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[95] 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.[86]). 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.

8 APPENDICES

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

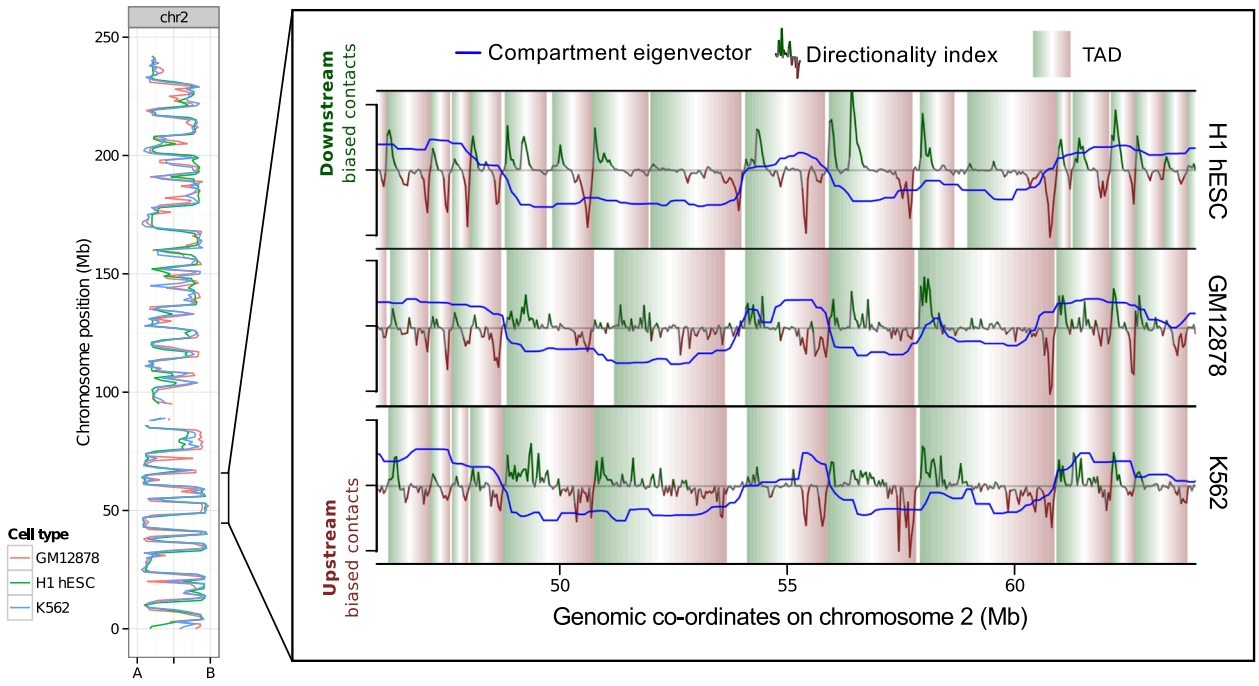

Figure 1 Concordance of chromatin structure at multiple scales over three human cell types. The eigenvector compartment profile is shown for chromosome 2 for three human cell types (left). Genome-wide Pearson correlation coefficients between eigenvectors at 1-Mb resolution are in the range 0.75 to 0.80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

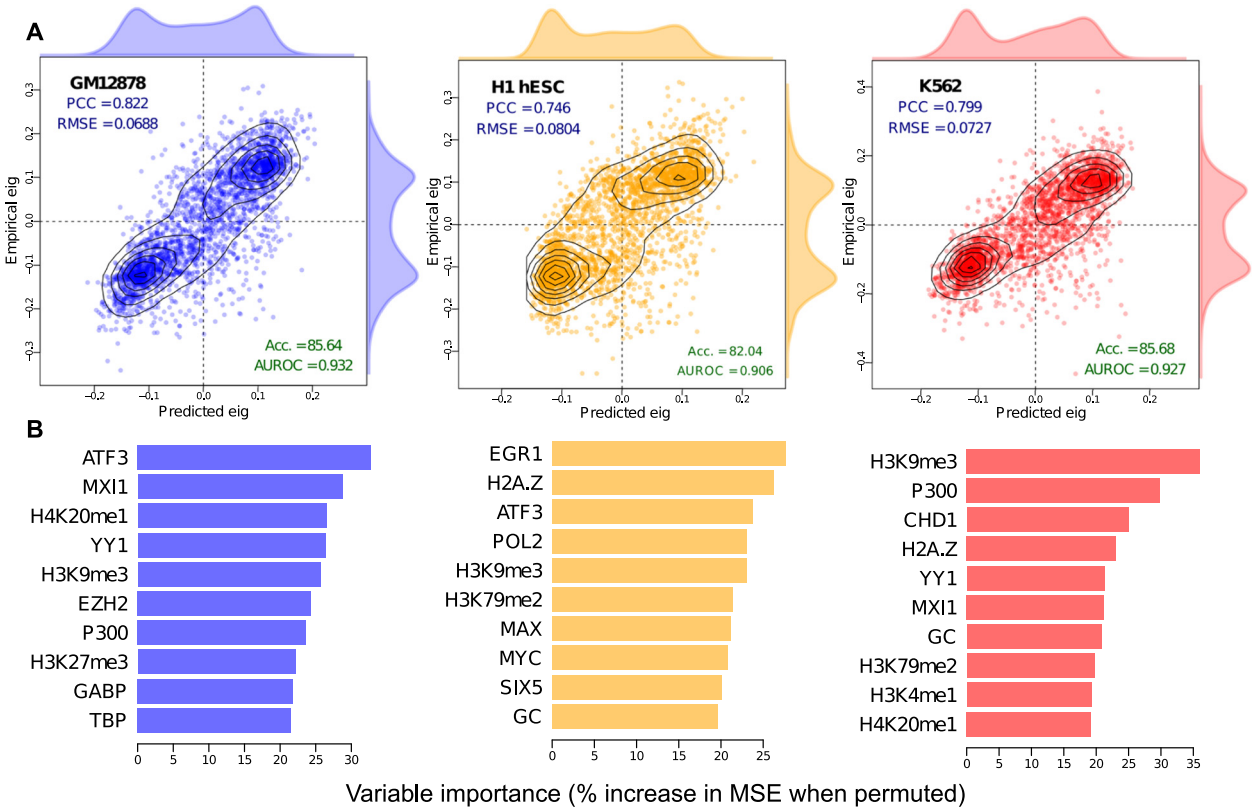

Figure 2 Accurate models of higher-order chromatin state built from locus-level features. **(A)** Model predictions (Predicted eig) are compared to observed values (Empirical eig). Various metrics are used to measure the accuracy of regression modeling – the Pearson correlation coefficient (PCC) and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

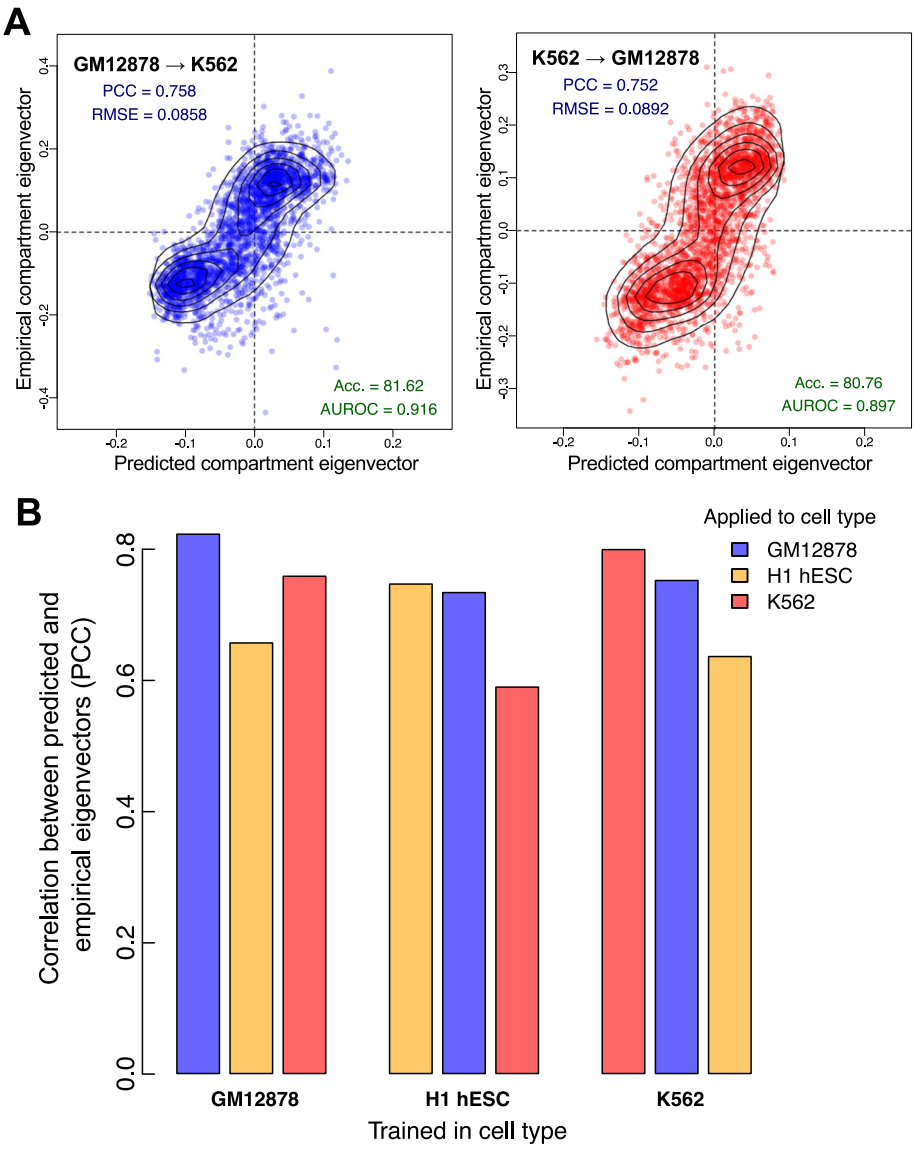

Figure 3 Models trained in one cell type can generalize to others. Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. **(A)** The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross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

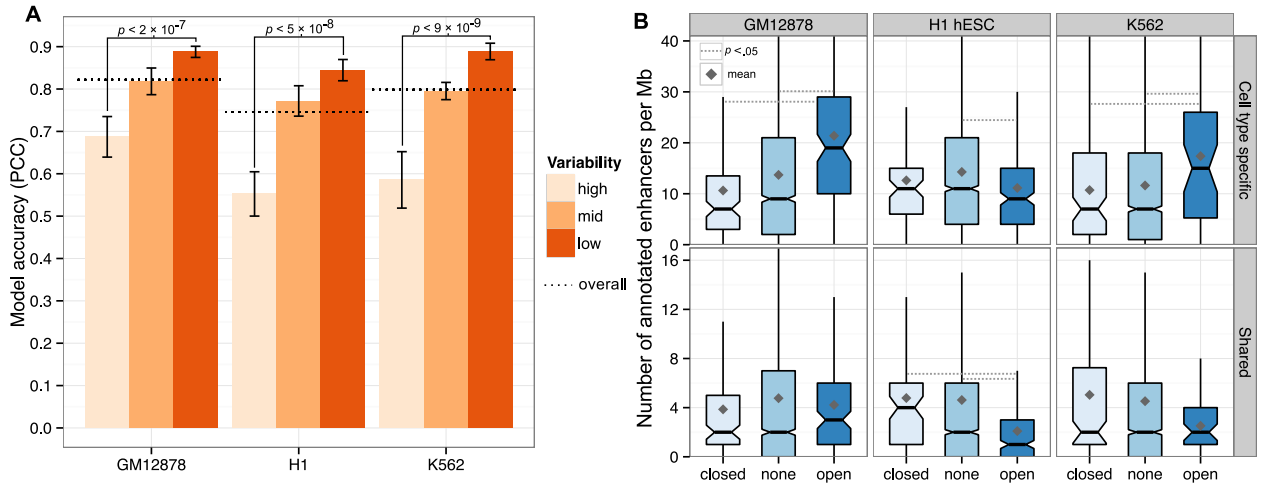

Figure 4 Regions with variable structure are less successfully modeled and are associated with cell-type-specific enhancer activity. **(A)** Model accuracy is significantly different between low- and high-variability regions, defined as, respectively, the lowest and highest thirds of the distribution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

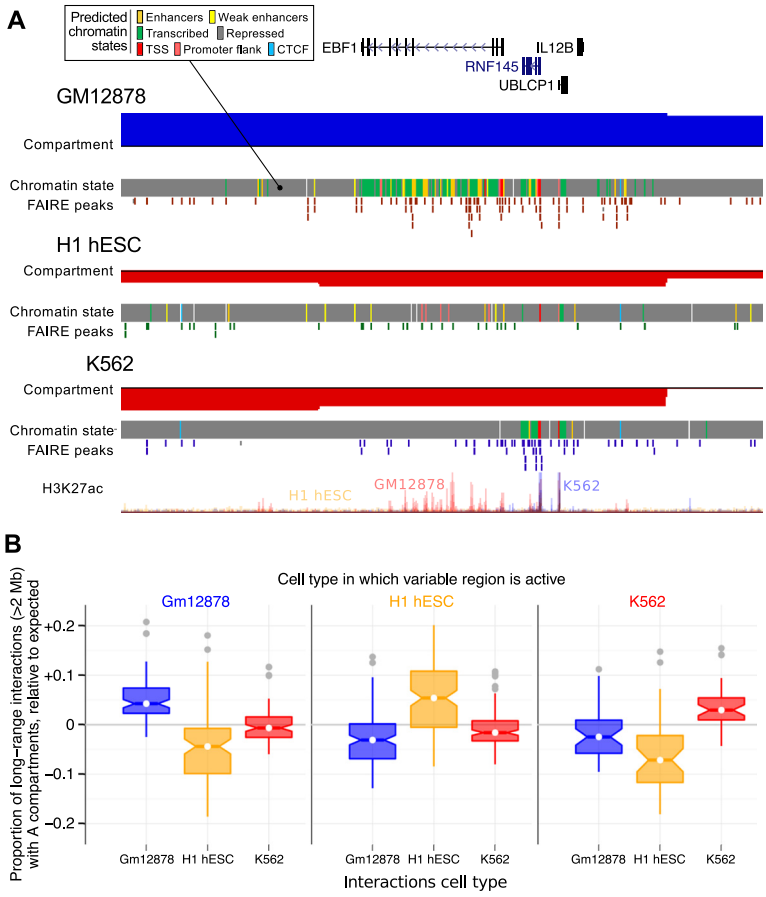

Figure 5 Structurally variable regions indicate cell-type-specific biology. Regions occupying the active A nuclear compartment in one cell type, but the repressed B compartment in the other two, were selected and ranked by the number of predicted active enhancers (Figure 4). **(A)** The region chr5:158–159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

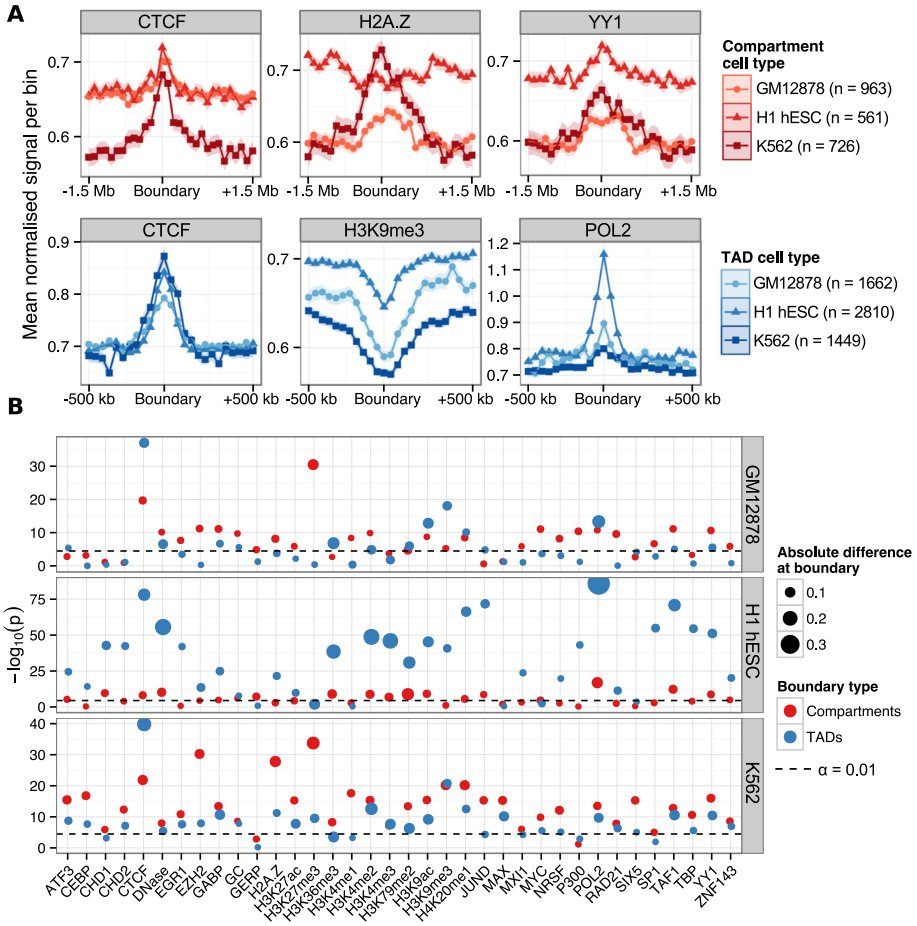

Figure 6 Chromatin features underlying TAD and compartment boundaries. **(A)** Selected profiles for locus-level features are shown for TAD boundaries (CTCF, H3K9me3 and POL2) and compartment boundaries (H2A.Z, H3K4me2 and YY1), as a mean normalized ChIP-seq signal relative to input chromatin per bin (±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] Pombo A, Dillon N (2015) Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*, **16**(4): 245–257.

[2] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] 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.

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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–.

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

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

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

[35] 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.

[36] 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.

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

[38] 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.

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

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

[41] 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.

[42] 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.

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

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

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

[46] 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.

[47] 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.

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

[49] 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.

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

[51] 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.

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

[53] Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, **9**(3): 432–41.

[54] Mazumder R, Hastie T (2012) The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, pp. 1–21.

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

[64] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan Kk, Cheng C, Mu XJ, Khurana E, Rozowsky J, *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**(7414): 91–100.

[65] 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.

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

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

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

[69] Li G, Reinberg D (2011) Chromatin higher-order structures and gene regulation. *Current opinion in genetics & development*, **21**(2): 175–86.

[70] 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.

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

[72] 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.

[73] 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.

[74] 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.

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

[76] 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.

[77] 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.

[78] 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.

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

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

[81] 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.

[82] 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.

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

[84] 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.

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

[86] 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.

[87] 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.

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

[89] 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*.

[90] 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.

[91] 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.

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

[93] 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.

[94] Hagmeyer BM, Duyndam MC, Angel P, de Groot RP, Verlaan M, Elfferich 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.

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