Antecedents of higher-order chromatin structure:

Insights from integrative modelling

**Benjamin L. Moore**

**Supervisors: Colin A. Semple and Stuart Aitken**

**MRC HGU, IGMM**

**30/08/2013**

**Abstract**

**Recent advances in chromosome conformational capture technology have permitted genome-wide assessment of higher-order chromatin structure in a variety of cell types. This structural information in conjunction with data produced by the ENCODE consortium offers an unprecedented opportunity to quantitatively investigate the relationship between locus level chromatin features (such as DNA methylation, histone modification and transcription factor binding) and higher-order chromatin organisation.**

**Hi-C genome-wide pairwise interactions can be reduced to an eigenvector summary metric that captures the arrangement of the genome into nuclear â€˜compartmentsâ€™ that have been shown to represent two distinct fractions of chromatin: gene dense, transcriptionally active regions and relatively gene poor, inactive regions. However the relationships between such higher-order phenomena and locus level features remain controversial and have not been quantitatively studied. Similarly, the extent to which such datasets intersect, and how they relate to one another across cell types, is poorly understood.**

**We have built genome-wide, quantitative models describing higher-order chromatin structure based on the underlying constellations of locus level features, such as the levels of histone modifications and DNA-binding proteins. In three very different cell types, Random Forest based regression models achieved high predictive accuracy even when regularised to as few as 6 predictive features (e.g. r = 0.86). Two histone marks, H3K79me2 and H3K4me2, were consistently identified as important predictors of compartment identity across all 3 cell-types, suggesting a heightened significance for these specific modifications with regard to higher-order chromatin structure. However the models otherwise proved to be surprisingly cell type specific, with largely inconsistent influential variables, and notably reduced predictive power when a model for a particular cell type was applied to other cell types.**

**This statistically rigorous modelling approach offers new insights into the contribution of locus level features to nuclear organisation in diverse cell types, and produces testable hypotheses that may enable a greater understanding of higher-order chromatin structure. In addition, the overall modelling accuracy on regions totalling more than 1.3 GB of the human genome implies the presence of general mechanisms of higher-order chromatin assembly.**

# 1. Introduction

The advent of chromosome conformational capture (3C) based methods has produced a wealth of chromosome topological data which offers insights into the causal factors and biological outcomes related to three-dimensional genome structure. Interpretation of these contact maps, however, remains challenging and requires the development of innovative statistical and computational analysis methods.1,2,3

A high-profile example of computational analyses leading to new biological insight can be found in Dixon *et al.*4 wherein the authors characterised “topological domains” (also known as topological associating domains or TADs), a megabase-scale feature of genome organisation conserved between human and mice. At even lower resolution, Lieberman-Aiden *et al.*5 identified “A” and “B” compartments, regions of between 1 and 5 megabases which showed properties typical of euchromatin and heterochromatin, respectively. Thus the combination of these two insights has lead to a model of higher order chromatin structure whereby groups of TADs assemble into alternating A and B compartments, reflecting broadly active and inactive chromosomal regions.1

The link between epigenetic features and local chromatin state has been analysed computationally in a number of publications, notably the Hidden Markov Model-based ChromHMM6 which predicts states such as active promoters and enhancers, using a range of histone marks and other features.7 Similarly a Random Forest-based algorithm was recently developed to predict enhancers from histone modifications.8 At the opposite end of the spectrum, theoretical mechanistic models of chromatin folding such as the “strings and binders switch” model9 and the “fractal globule” model5,10,11 have both produced simulated data that reflects empirical 3C observations and potentially describe the polymer dynamics of chromatin folding. However as an intermediary between these two approaches, it is not yet known how local chromatin features may be related to higher order genome organisation.

With this aim in mind, the recent comprehensive ChIP-seq datasets produced by the ENCODE consortium12 combined with Hi-C genome-wide contact maps in a number of human cell types4,5,13 present a remarkable opportunity to investigate this potential relationship between local chromatin features and higher order structure. In this work, a machine-learning approach was employed to model the compartmental characteristics of large genomic regions based on their aggregate levels of various histone marks and DNA binding proteins. Dissection of the resulting models were then be used as a means of gleaning biological insight regarding the importance of contributory factors and in highlighting differences between cell types.

# 2. Methods

An overview of the analysis pipeline implemented in this work is shown below (Fig. ).

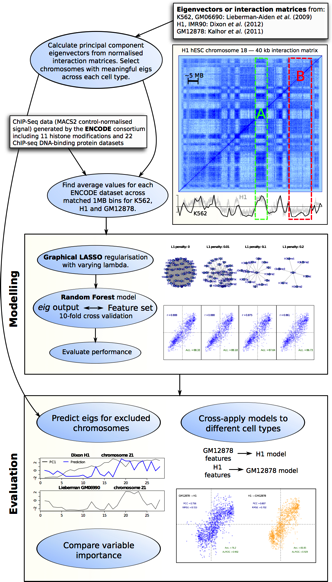


Figure 2.1: **Workflow schematic.**

## 2.1 Input data

### 2.1.1 Eigenvectors

Genome-wide intrachromosomal eigenvectors were extracted from published materials for cell types GM06690,5 K5625 and GM1287813. Eigenvectors for cell types H1 and IMR90 were calculated via principal component analysis applied to published 40 kb resolution interaction matrices.4 Those eigenvectors mapped to previous reference genome builds (hg18/GRCh36) were transferred onto hg19 co-ordinates (GRCh37) using the UCSC LiftOver tool.14

The eigenvectors were then averaged into 1 MB bins, matching the same co-ordinates across cell types. Megabase bins with less than an average of 80% eigenvector coverage were excluded. Eigenvectors were then standardised on a per cell type basis to leave comparable values.

Chromosomes in which the calculated first principal component eigenvector did not reflect A/B compartmentalisation were filtered. Pearson correlation was calculated between cell types per chromosome, and those with an average coefficient greater than one standard error above the population mean were selected (Fig. ). A minority of chromosomes meeting this criteria were excluded based on visual inspection, where they showed insufficient agreement with an observable plaid pattern exhibited by the Hi-C interaction matrix, as described previously.5 After this filtering, 11 chromosomes remained (1-3, 6, 11-16 and 18), a total of 1311×1 MB bins.

It is worth briefly noting the caveats associated with the Hi-C datasets used in this work. Firstly, each Hi-C interaction matrix represents data from a population of cells, hence cell-to-cell variability is masked; also a number of biases inherent to the procedure have been identified.15,16 However, the significance of these concerns is lessened for the purposes of characterising large 1 MB blocks, particularly given that enriched interactions are harshly normalised via correlation (and only a principal component is then taken forward).

### 2.1.2 Locus-level features

Genome-wide ChIP-seq datasets for: 22 DNA binding proteins and 10 histone marks were made available by the ENCODE consortium,12 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 processed using MACS217 to produce fold-change relative to input DNA. GC content was also calculated and used in the featureset.

## 2.2 Modelling

### 2.2.1 Random Forest

In full models, Random Forest regression18 was used as implemented in the R package randomForest.19 Parameters of and *ntrees*=200 were assumed as they are known to be largely insensitive;20 this was verified with the dataset used in this work (Fig. ).

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.21

### 2.2.2 Graphical lasso

Regularised models made use of the Graphical LASSO22 (least absolute shrinkage and selection operator) as a method of -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.22

More specifically, the graphical lasso regulates the number of 0s in the inverse covariance matrix, , also known as the precision matrix. Then if element , the variables and can be said to be conditionally independent, given the remaining variables.23 The algorithm minimises a negative log-likelihood (Eqn. 23) 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.

(2.1)

## 2.3 Analysis

### 2.3.1 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 predicted and observed eigenvectors (determined by 10-fold cross-validation), and the root mean-squared error (RMSE)of the same data. Classification error, when predictions where thresholded into *A*>0;*B*≤0, was also calculated using accuracy (% correct classifications or True Positives) and area under the receiver operating characteristic (AUROC) curve (Fig. ). Together these give a comprehensive overview of the model performance, both in terms of regression accuracy of the continuous eigenvector, and in how that same model could be used to label discrete chromatin compartments.

### 2.3.2 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 (Section ).

Hidden Markov Models (HMMs) were fit using the Baum-Welch algorithm to eigenvectors of each cell type. These HMMs were then used to simulate dummy data to calculate the significance of observed variability (Fig. ). The resulting distribution of MAD values was fit by a Weibull extreme value distribution, of which two-tailed quantiles were then used to determine significance cutoffs (Fig. ).

### 2.3.3 Nuclear positioning of chromatin compartments

Previously published data on chromosome positioning preference within the nucleus was used to label each chromosome as “inner”, “middle” or “outer”.24 Chromosomes whose DAPI hybridisation signals were significantly enriched () in the inner nuclear shell, as defined by Boyle *et al.*24, made up the “inner” group and included chromosomes 1 and 16. Similarly the “outer” group had enriched signals () 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*≥0.1).24 The significance of the difference in distribution of eigenvectors in the inner vs. 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 is greater-than or equal-to .

This chromosomal positioning data was measured in lymphoblastoid cells though nuclear architecture is though to be largely conserved between cell types25,26 and even higher primates.27

# 3.Results

## 3.1 Concordant compartmentalisation across cell types

Genome compartments proved well-conserved across human cell types (Fig. ), with Pearson correlation coefficients between cell types ranging from 0.57–0.85 (Fig. ). When A/B compartmentalisation was called using an HMM, 72.6% of 1 MB blocks were estimated as being in the same underlying state in H1, K562 and GM12878 cell types (Fig. ).

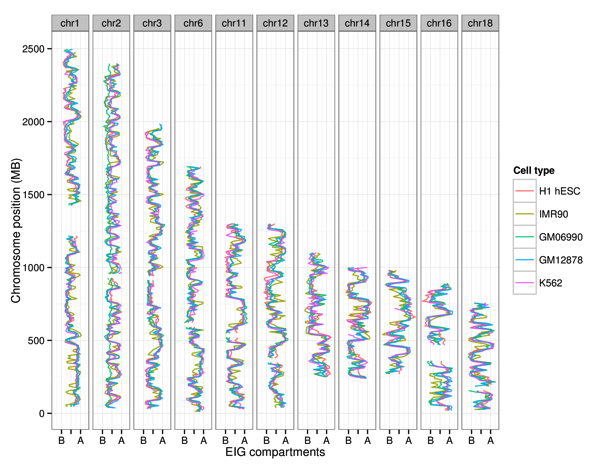
 .9

Figure 3.1: **Highly concordant A/B compartmentalisation over selected chromosomes from 5 cell types.** Principle component eigenvectors plotted along their respective chromosomes for five different cell types. “A” and “B” labels reflect compartments bins with positive and negative eigenvectors respectively,5 after being orientated to positively correlate with Pol2 binding data.13

## 3.2 Accurate models of higher-order structure

Cell-type specific models of higher order structure proved highly accurate in predicting the compartment identity of individual 1 MB blocks, producing Pearson’s correlation coefficients (PCC) of 0.73–0.89 (*p*≈0) between predicted and empirical eigenvector values (Fig. ).

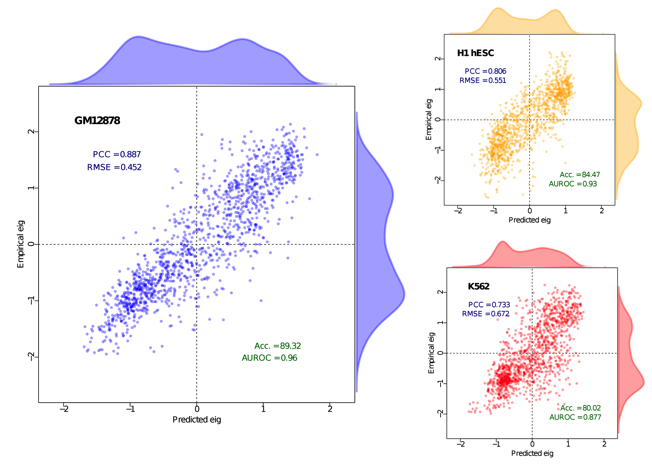


Figure 3.2: **Accurate models of eigenvector values across three cell types.** Predicted eigenvector values are plotted against their actual recorded values for each cell type. Evaluation metrics shown are Pearson’s correlation coefficient (PCC), root mean-squared error (RMSE), and classification evaluators (with correct classification defined as either >0 in both test and training set or both <0 — the top-right and bottom-left quadrants of the above plots) accuracy (% true positives) and area under the receiver operating characteristic curve (AUROC). Kernel density estimates describe the distribution of their opposite axes.

## 3.3 Parsimonious models highlight common features

Having established that the compartment property of higher order chromatin structure can be accurately predicted using a feature set of 34 predictors, it was then of interest to identify which of these were most influential in the RF models.

To this end, standard variable importance metrics produced by the RF models, such as mean decrease in accuracy, can be calculated and compared between models. However, in this instance there exists strong multicollinearity between predictors, as well as several individual high correlations between input feature and output eigenvector. For this reason a form of tuneable regularisation was desirable, allowing the dense models to be restricted to a small number of influential features which composed an interpretable model. The graphical LASSO22 (least absolute shrinkage and selection operator; hereafter glasso) calculates an estimate analogous to a measure of pairwise conditional independence between nodes,28 and was selected over competing methods for several resons: (a) due to the geometry of -regularisation, the resulting precision matrix is sparse, hence the glasso can be used to removes conditionally independent variables with respect to a regularisation parameter; (b) under Gaussian Markov Random Field (GRMF) theory, the precision matrix estimate relates to an interpretable graphical output;[[1]](#footnote-1) (c) fast speed of execution.20,22

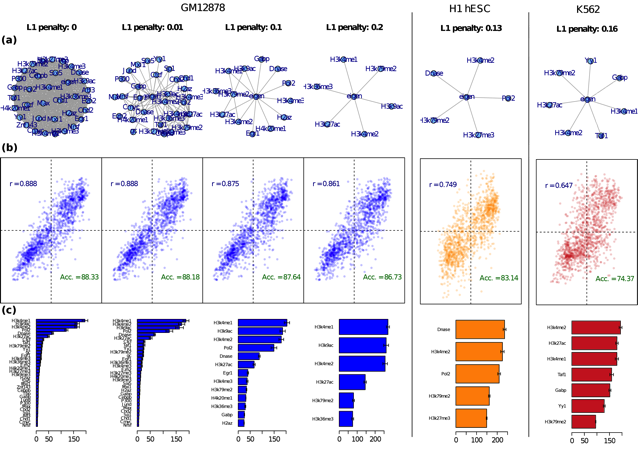


Figure 3.3: **Graphical Lasso -based regularisation evolves parsimonious graphical models for feature selection.** (a) Graphical models produced by the glasso algorithm22 for varying regularisation parameter λ. (b) Corresponding RF results using these reduced feature sets as in Fig. along with (c) variable importance estimates in terms of mean decrease in accuracy (see Methods ). Full λ sequences for K562 and H1 are given in the supplementary materials (Figs. , ).

Glasso regularisation was used to produce models with ≈5 features and these were then used to retrain Random Forest regression models (Fig. ). While in each case the model performance slightly deteriorates with increased regularisation, the remaining variables offer insight into the primary antecedents of chromosome compartmentalisation in each cell type.

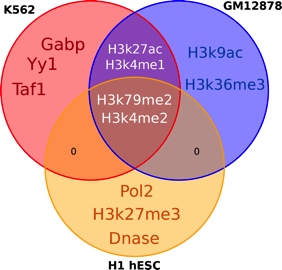


Figure : **Overlap of remaining variables in regularised models** Venn diagram showing the intersections of features kept in parsimonious models (Fig. ).

Surprisingly, the remaining features in the regularised models are largely inconsistent between cell types (Fig. ). Two histone marks, H3K4me2 and H3K79me2, are present in each of the regularised models and another two, H3k27ac and H3k4me1, remain in both K562 and GM12878 cell type models (Fig ). The remaining variables were specific to individual cell type models. By randomly sampling variables it can be shown that the size of the intersection between all three sets is significantly larger than would be expected by chance (), yet overall there remains a surprising disparity between cell types given the observed correlations of the response variable (Results ).

## 3.4 Invariant regions are better described by locus-level features

The set of matched 1 MB blocks was then stratified into regions of low, mid and high variability based on the mean absolute deviation (MAD) of compartment eigenvectors (see Methods ). Modelling these regions independently revealed that low variability, or relatively cell type invariant blocks, could be significantly more accurately predicted relative to high variability regions (GM12878: , H1: ; K562: ; Fig. ). Additionally, in the K562 cell type, the subset of regions conserved with GM12878 and H1 could be significantly better predicted than all 1 MB blocks ().

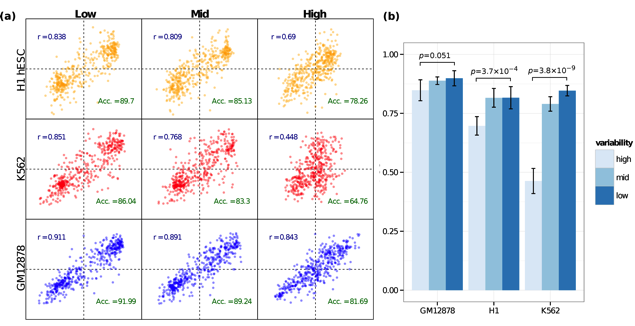


Figure 3.5: **Variable chromosomal regions are more difficult to model than those that are stable across cell types.** (a) Scatterplots comparing predicted and empirical eigenvector values for subsets split by variability. (b) Bar chart showing the average Pearson’s correlation coefficient (PCC) over all 10 folds, with 95% confidence intervals indicated. “Low” variability regions, the of the 1 MB bins with the lowest median absolute deviation across cell types, proved more amenable to predictive modelling in each cell type.

This result could indicate there exists a number of genomic sites with a fixed higher order chromatin state that is well-defined by histone marks, transcription factors and related components. Conversely, the hard-to-predict variable regions could be those under the influence of cell type specific factors which are not present in the set of predictors, or through localised chromatin events. An alternative interpretation is that the high variability regions are those in which the principal component is least accurately reflecting columns of the Hi-C interaction matrix, else regions most affected by artefacts of the Hi-C data processing.

Significantly invariant blocks (see Methods ) were tested for a range of potential genomic feature enrichments using the GREAT tool,29 but no significant results were observed (*data not shown*).

## 3.5 Models differ between cell types

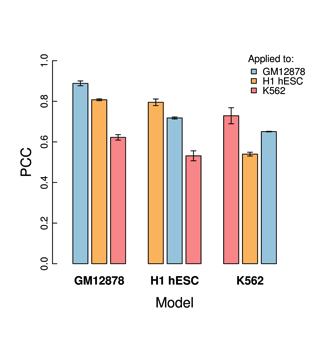
 .5

Figure 3.6: **Cross-application of models results in decreased prediction accuracy** The PCC between predicted and empirical eigenvector values is shown (with 95% confidence intervals) for models comparing their performance using feature sets from the same cell type against those from the other two.

The cell type specific models of higher order structure were then cross-applied, whereby the locus-level chromatin features of one cell type were used as a featureset in a RF regression model trained in a different cell type. In each instance of cross-application, the models performed worse at predicting the chromatin state in foreign cell types (Fig. ).

In each case of cross-application, the results reflect the relative accuracy of cell type specific models (Fig. ). For example, GM12878 blocks are more accurately predicted by the H1 model compared to K562 blocks () and are also relative to H1 blocks with the K562 model (). Similarly K562 feature sets result in the lowest prediction accuracy using GM12878 or H1 models (Fig. ).

## 3.6 Models generalise to unseen chromosomes

Given that do not appear to generalise well across cell types (Results ), it was of interest to confirm that the models were not overfit to the training data. As an example, the previously-excluded chromosome 21 was used as an external validation set (Fig. ).

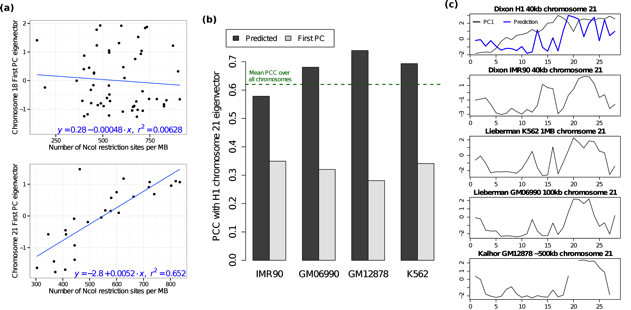


Figure 3.7: **Model can predict eigenvectors for those chromosomes whose first principal component does not describe its compartmentalisation.** (a) The first principal component of the H1 chromosome 21 interaction matrix is largely explained by the number of restriction sites per megabase (; cf. chromosome 18: ). (b) The model prediction of the H1 compartmentalisation correlates well with other cell types, in most cases above the average PCC between all autosomal chromosomes. (c) Visual comparison of predicted eigenvector with the same chromosome in other cell types.

The first principal component eigenvector of chromosome 21 in the H1 cell type can be largely explained by the number of restriction enzyme sites per MB, whereas this is not the case for other chromosomes (Fig. a). By applying the RF regression model to the features on chromosome 21, predicted eigenvectors proved more agreeable with eigenvectors from the same chromosome in other cell types (Fig. b). Hence this prediction appears to be reflect the genuine compartmentalisation of chromosome 21 in H1 with reasonable accuracy.

**Chapter 4**

# Discussion

## 4.1 Relationship between locus-level chromatin features and higher order structure

Despite the accruement of much relevant data, the relationship between locus-level chromatin features and higher order structure remains poorly understood. This work has shown that a strong correlative relationship exists between higher order chromosome compartmentalisation and aggregate levels of several histone modifications and DNA binding proteins.

Interpretations of the observed relationship could be either (a) causative, whereby specific histone modifications and other bound factors alter nucleosome dynamics and bring about a more open and active higher order structure or (b) purely correlative, such that chromatin is organised by latent factors (such as nuclear lamina and nuclear matrix proteins), with large scale active regions then painted with active marks as a side-effect of transcriptional activity.30 A means of distinguishing between these two explanations could be a biological perturbation study, with specific factors (possibly the methyltransferases responsible for H3K4me2 or H3K79me2) being downregulated in a population of cells which could then be used for Hi-C analysis. Significant changes to compartmentalisation would then indicate a (however indirect) causal role in higher order chromatin structure.

Interestingly, the DOT1 histone methyltransferase which writes to H3k79 has previously been linked with DNA stability in yeast.31 Another study reported, regarding the mammalian orthologue DOT1L, that despite correlating with transcription, downregulation left most genes transcribed at their normal rate.32 The same study linked the “parallel nature of H3K4 methylation and H3K79 methylation” 32—implicating co-operation with the other histone mark (H3K4me2) found in all parsimonious models (Fig. ).

## 4.2 Implications for models of genome topology

A previous statistical analysis hypothesised that interphase chromatin organisation at the megabase scale is driven by some combination of sequence factors and epigenetic states, along with the region’s longitude along a chromosome arm.16 This type of explanation ties in with the “dog-on-a-lead” model of chromosome topology,33 which states the chromosome (holding the “lead”) constrains genomic regions to local areas of the nucleus—but within those contraints genes and regulatory elements have some flexibility to locate preferential binding partners.26,33 It goes on to state that some features, such as centromeres, are dominant over their chromosomes, hence have disproportionate influence on local genomic interactions.

This combined input would explain the observed high correlation of compartments across cell types (Fig. ); fixed chromosomal territories could aid in organising the invariant regions while cell-type specific chromatin states are responsible for the observed variance. Indeed, some support for this hypothesis can be found by contrasting relative variable importance metrics between high and low variability models (Figs. , ), where GC content’s predictive power is decreased in all cell types when comparing low variation blocks with those that are highly variable. Likewise this could explain the loss of accuracy during cross application (Fig. ) and the partial overlap of important features in parsimonious models (Figs. , ).

## 4.3 Caveats

In this work, megabase regions were treated as independent response variables, though the HMMs designed to call compartment states highlight that this is an oversimplification (Figs. , ). Including adjacent compartment values as predictive variables yielded significant increases in model accuracy (Fig. ), but does not aid in the understanding of the relationship between locus-level features and higher order chromatin. Another important consideration of the presented models and their underlying datasets is pervasive multicollinearity, which in particular limits the power of statistical tools to delineate individual variable contributions.

**Chapter 5**

# Conclusion

We have shown that higher order chromatin structure can be accurately predicted using aggregate locus-level features. Of these, H3K4me2 and H3K79me2 may be of particular importance. We also note that despite a general concordance of compartment states across cell types, there exists surprising disagreement between cell-type specific models, both in terms of relative variable importance and through cross-application of feature sets. These observed differences could be due to a hypothesised biology of genome topology whereby cell type invariant regions can be well-defined by locus-level signals, but variable regions may be more influenced by specific active enhancers and other facets of transcription regulation machinery.

**Chapter**

# Additional figures

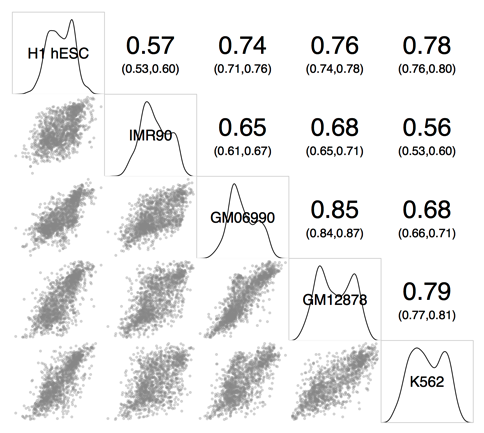


Figure 5.1: **General concordance of eigenvectors across cell types.** Correlogram showing the mean correlation of eigenvectors across selected chromosomes of five Human cell types. Pearson’s correlation coefficient is shown (*upper*) along with kernel density estimates of each cell type’s eigenvector distribution (*diagonal*) and scatterplots comparing megabase blocks from each cell type (*lower*).

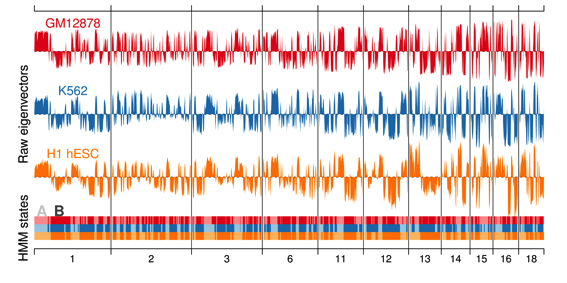


Figure 5.2: **Eigenvectors and HMM state calls for selected chromosomes across three cell types.** 72.6% of HMM state calls (*lower*) are in agreement across 1 MB blocks in three human cell types: GM12878, K562 and H1 hESC.

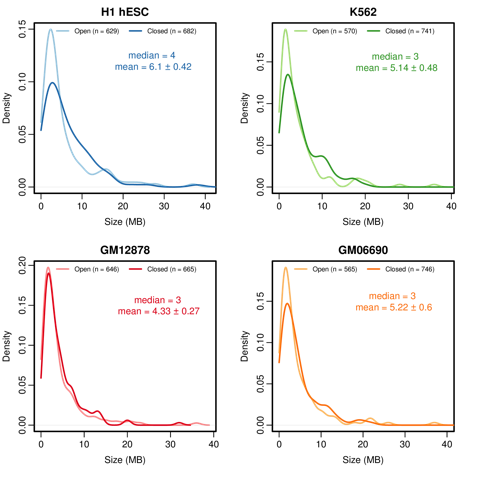


Figure 5.3: **Characterisation of open and closed compartment sizes in various cell types.** Density plots showing the size distributions of open and closed chromosome compartments. Means are shown with 95% confidence intervals. *n* refers to the number of 1 MB blocks classified as either open or closed.

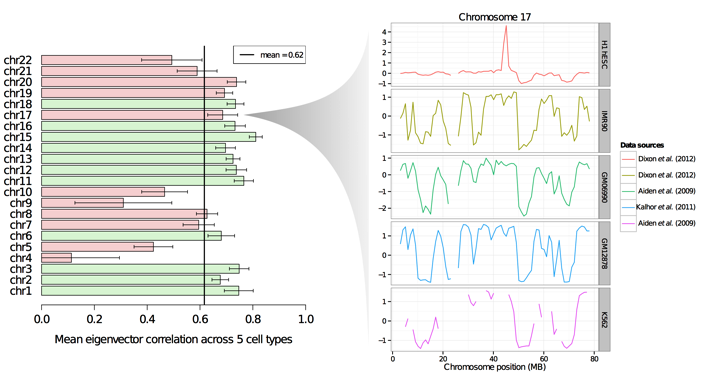


Figure 5.4: **Selecting correlated chromosomes for modelling.** Generally those chromosomes whose mean correlation across all 5 cell types was 1 standard error above the mean were taken forward as examples of properly-formed PC eigenvectors reflecting A/B compartmentalisation. Some chromosomes meeting this criterion were excluded due to obvious aberration or not reflecting the observable “plaid” pattern in the normalised interaction matrix, such as chromosome 17 (*right*).

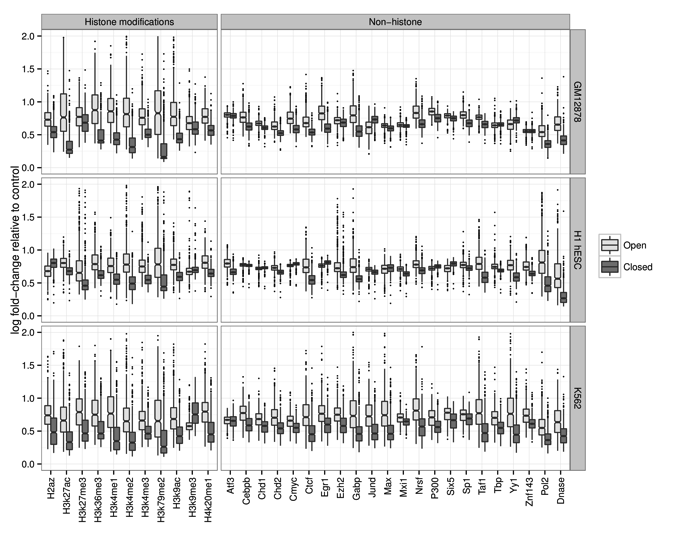


Figure 5.5: **Characterisation of open and closed compartments in terms of cell-matched locus level data.** Each variable distribution is depicted as a box-and-whisker diagram for open and closed megabase blocks. The *y*-axis represents the fold signal change relative to ChIP-Seq input control per MB.

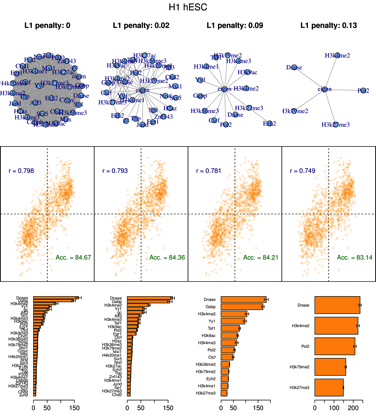
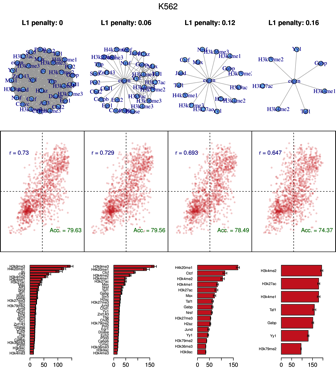
  1

Figure 5.6: **Graphical Lasso -based regularisation evolves parsimonious graphical models for feature selection.** Graphical models produced by the GLASSO [ref] algorithm are shown for varying regularisation parameter λ (*upper*), as in Figure . Here the results are shown for cell types H1 (*upper*) and K562 (*lower*).

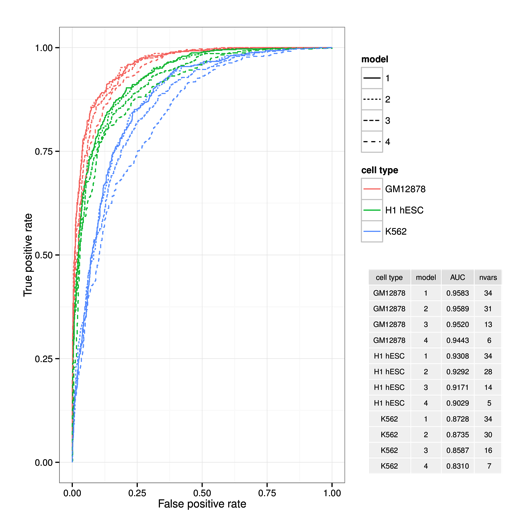


Figure 5.7: **Receiver operating curves for the three cell type models at varying levels of regularisation.** The receiver operating curves (ROC) are shown for each model regularisation (Model 1-4) and for each cell type (see Fig. ). The table (*inset*) gives the area under ROC (AUC) value as well as the number of variables (nvars) remaining in the each regularised model.

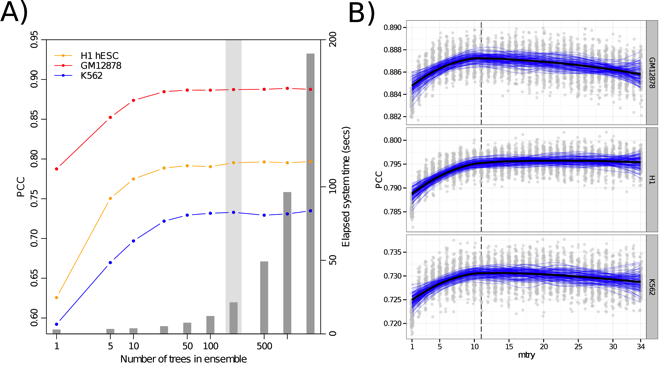
 .9

Figure 5.8: **Random Forest parameters proved largely insensitive to parameters ntrees and mtry.** The number of trees, *ntree* (A), was chosen as 200 and the number of variables tested at each node, *mtry* (B), was the default value for regression: *n*/3≈11.

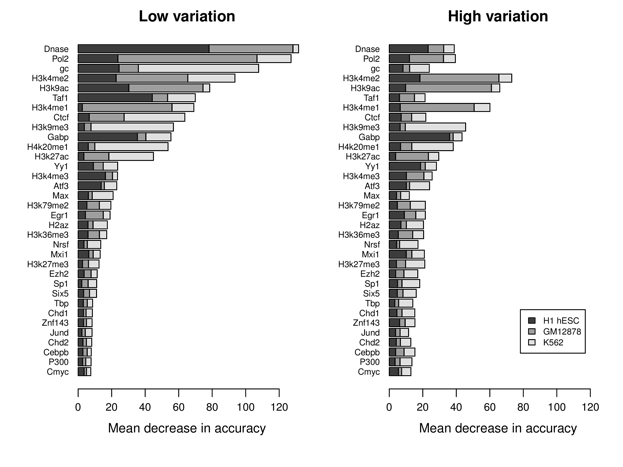


Figure 5.9: **Relative variable importance for RF features in models built with blocks that are conserved between human cell types (low variation) and those that are variable (high variation).** Mean decrease in accuracy (Methods ) is calculated for each variable in three human cell types and compared between the of regions with the lowest mean absolute deviation across cell types (*left*) and the with the highest (*right*).

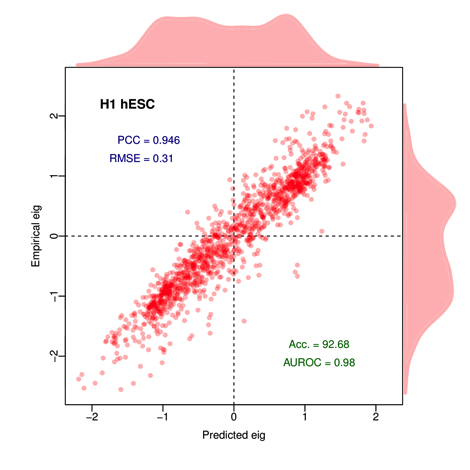


Figure 5.10: **Autoregressive terms increase model accuracy.** When adjacent eigenvector values (, ) are used as features in predicting the state of a central MB block () the model accuracy greatly improves. This is shown above for the H1 cell type, highlighting the known non-independence of adjacent eigenvector values.

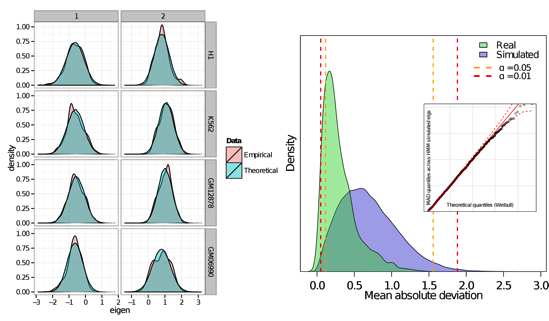


Figure 5.11: **Estimating the significance of median absolute deviations observed between eigenvectors across cell types.** Two state normal HMMs were fit to observed eigenvectors in each cell type (*left*) and the median absolute deviation was calculated at each position. The distribution of these values was approximated by a Weibull extreme-value distribution (with *k*=1.87, λ=0.76; QQ-plot *right, inset*) and the quantiles of this distribution were used to determine the significance of observed values (*right*).

Bibliography

[1] Dekker, J, Marti-Renom, M. a, & Mirny, L. a. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics* **14**, 390–403.

[2] Steffen, P. a, Fonseca, J. a. P, & Ringrose, L. (2012) Epigenetics meets mathematics: towards a quantitative understanding of chromatin biology. *BioEssays : news and reviews in molecular, cellular and developmental biology* **34**, 901–13.

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

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

[5] Lieberman-Aiden, E, van Berkum, N. L, Williams, L, Imakaev, M, Ragoczy, T, Telling, A, Amit, I, Lajoie, B. R, Sabo, P. J, Dorschner, M. O, Sandstrom, R, Bernstein, B, Bender, M. A, Groudine, M, Gnirke, A, Stamatoyannopoulos, J, Mirny, L. A, Lander, E. S, & Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93.

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

[7] Ernst, J, Kheradpour, P, Mikkelsen, T. S, Shoresh, N, Ward, L. D, Epstein, C. B, Zhang, X, Wang, L, Issner, R, Coyne, M, Ku, M, Durham, T, Kellis, M, & Bernstein, B. E. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–9.

[8] 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**, e1002968.

[9] Barbieri, M, Chotalia, M, Fraser, J, Lavitas, L.-M, 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**, 16173–8.

[10] Mirny, L. a. (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**, 37–51.

[11] Grosberg, A. Y, Nechaev, S, & Shakhnovich, E. (1988) The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique* **49**, 2095–2100.

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

[13] 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**, 90–8.

[14] Hinrichs, A. S, Karolchik, D, Baertsch, R, Barber, G. P, Bejerano, G, Clawson, H, Diekhans, M, Furey, T. S, Harte, R. A, Hsu, F, Hillman-Jackson, J, Kuhn, R. M, Pedersen, J. S, Pohl, A, Raney, B. J, Rosenbloom, K. R, Siepel, A, Smith, K. E, Sugnet, C. W, Sultan-Qurraie, A, Thomas, D. J, Trumbower, H, Weber, R. J, Weirauch, M, Zweig, A. S, Haussler, D, & Kent, W. J. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**, D590–8.

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

[16] Imakaev, M, Fudenberg, G, McCord, R. P, Naumova, N, Goloborodko, A, Lajoie, B. R, Dekker, J, & Mirny, L. A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999–1003.

[17] Zhang, Y, Liu, T, Meyer, C. a, Eeckhoute, J, Johnson, D. S, Bernstein, B. E, Nusbaum, C, Myers, R. M, Brown, M, Li, W, & Liu, X. S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137.

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

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

[20] Hastie, T, Tibshirani, R, & Friedman, J. (2009) *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer), 2 edition.

[21] Cutler, D. R, Edwards, T. C, Beard, K. H, Cutler, A, Hess, K. T, Gibson, J, & Lawler, J. J. (2007) Random forests for classification in ecology. *Ecology* **88**, 2783–92.

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

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

[24] Boyle, S, Gilchrist, S, Bridger, J. M, Mahy, N. L, Ellis, J. a, & Bickmore, W. a. (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics* **10**, 211–9.

[25] Chambers, E. V, Bickmore, W. A, & Semple, C. A. (2013) Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS computational biology* **9**, e1003017.

[26] de Wit, E, Bouwman, B. a. M, Zhu, Y, Klous, P, Splinter, E, Verstegen, M. J. a. M, Krijger, P. H. L, Festuccia, N, Nora, E. P, Welling, M, Heard, E, Geijsen, N, Poot, R. a, Chambers, I, & de Laat, W. (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* pp. 1–7.

[27] 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**, 4424–9.

[28] Menéndez, P, Kourmpetis, Y. a. I, ter Braak, C. J. F, & van Eeuwijk, F. a. (2010) Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge. *PloS one* **5**, e14147.

[29] McLean, C. Y, Bristor, D, Hiller, M, Clarke, S. L, Schaar, B. T, Lowe, C. B, Wenger, A. M, & Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495–501.

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

[31] Wysocki, R, Javaheri, A, Allard, S, Sha, F, Côté, J, & Kron, S. J. (2005) Role of Dot1-dependent histone H3 methylation in G1 and S phase DNA damage checkpoint functions of Rad9. *Molecular and cellular biology* **25**, 8430–43.

[32] Steger, D. J, Lefterova, M. I, Ying, L, Stonestrom, A. J, Schupp, M, Zhuo, D, Vakoc, A. L, Kim, J.-E, Chen, J, Lazar, M. A, Blobel, G. A, & Vakoc, C. R. (2008) DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Molecular and cellular biology* **28**, 2825–39.

[33] Krijger, P. H. L & de Laat, W. (2013) Identical cells with different 3D genomes; cause and consequences? *Current opinion in genetics & development* **23**, 191–6.

1. It should be noted that in this work, rather than using the resulting sparse inverse covariance matrix to parameterise a Gaussian graphical model, instead the glasso is used as a means of feature selection to generate a non-independent subset of influential variables as input to the RF model. [↑](#footnote-ref-1)