Antecedents of higher-order chromatin structure:

Insights from integrative modelling

First Year Report

Benjamin L. Moore

Supervisors: Colin A. Semple and Stuart Aitken

MRC HGU, IGMM

May 28, 2015

Contents

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 compartments 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 rules and mechanisms for higher-order chromatin assembly.

Glossary

An **eigenvector**, as used in this work, is a lower-dimensional summary of Hi-C interaction matrices that captures the broad open and closed chromatin states along a chromosome. Formally, an eigenvector is any non-zero vector $\vec{x}$ that satisfies $\mathbf{A}\vec{x} = \lambda\vec{x}$ for a given square matrix $\mathbf{A}$ and corresponding eigenvalue λ. In principal components analysis, the first principal component eigenvector represents the axis upon which the original data can be projected while retaining maximal variance.

The **glasso** (or Graphical LASSO, least absolute shrinkage and selection operator) builds a conditional dependence graph from a set of related variables. By increasing the tuning parameter, a greater number of variables will be filtered as conditionally independent from the other remaining variables. The glasso is conceptually related to an earlier algorithm [?] that built a graphical model by applying lasso regression to each variable in turn, using all others as predictors.

HMMs (hidden Markov models) are models which represent non-independent observations according to underlying unobserved states. They were used in this work as each 1 Mb region often spans only part of a chromosome compartment, meaning adjacent regions are more likely to share the same compartment state.

Random Forest is a powerful machine learning method that builds a predictive model and is particularly suited to high-dimensional data. The Random Forest algorithm grows an ensemble of bifurcating decision trees from a bootstrapped sample of a training set. Further randomness is introduced by picking a subset of available variables at each vertex to test for maximal separation of child nodes. The resulting forest can then be used for classification, through vote aggregation, or regression, by averaging leaf node values across the ensemble.

Regularisation (in the context of machine learning), is the process of imposing a penalty on a model's complexity. Regularisation was employed to reduce a model with many variables to a more understandable and parsimonious model with fewer variables.

1. Introduction

The advent of chromosome conformational capture (3C) based methods has produced a wealth of chromosome topological data which offer 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.[? ? ?]

A high-profile example of computational analyses leading to new biological insight can be found in Dixon *et al.*[?] 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 lower resolution, Lieberman-Aiden *et al.*[?] identified "A" and "B" nuclear compartments, made up of regions of between 1 and 5 megabases which showed properties typical of euchromatin and heterochromatin, respectively. 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.[?]

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[?] algorithm which predicts states such as active promoters and enhancers, using a range of histone marks and other underlying features.[?] Similarly a Random Forest-based algorithm was recently developed to predict enhancers from histone modification data.[?] At the opposite end of the spectrum, theoretical mechanistic models of chromatin folding such as the "strings and binders switch" model[?] and the "fractal globule" model[? ? ?] have both produced simulated data that reflects empirical 3C observations and potentially describe the polymer dynamics of chromatin folding. 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.

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

2. Methods

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

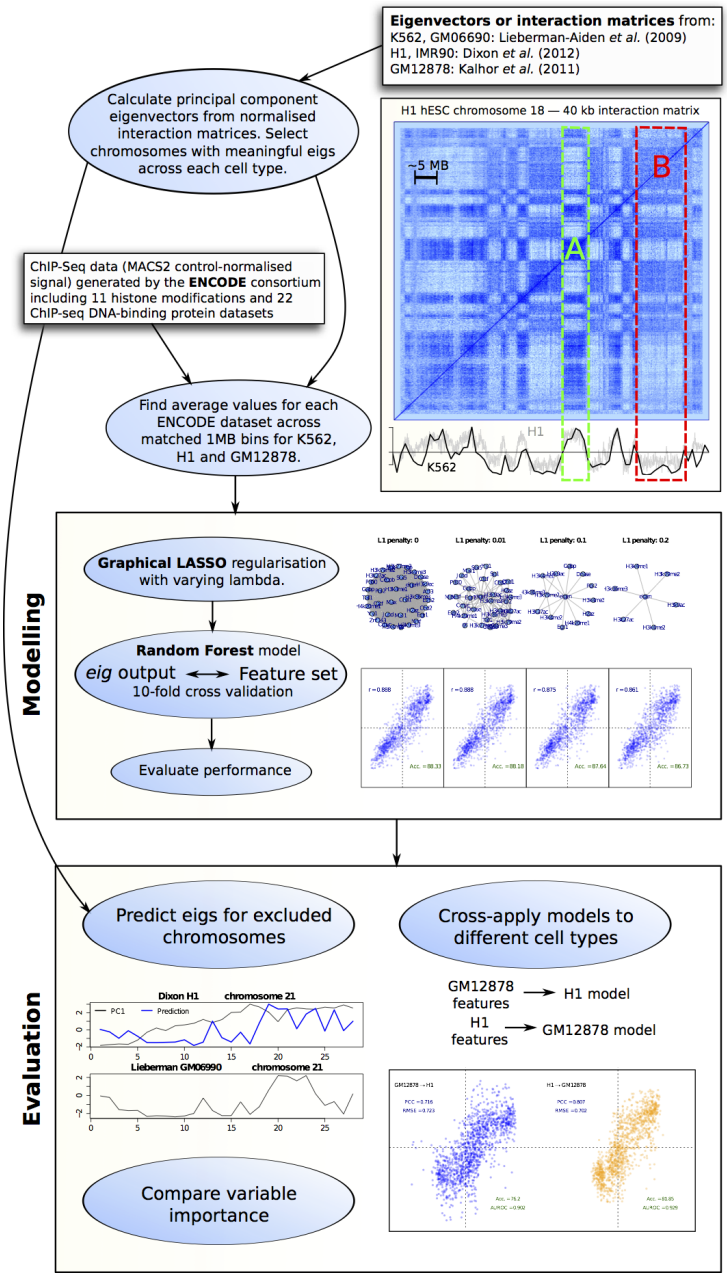


Figure 2.1: **Workflow schematic.**

2.1 Input data

2.1.1 Eigenvectors

Genome-wide intrachromosomal eigenvectors were extracted from published materials for cell types GM06690,[?] K562[?] and GM12878[?]. Eigenvectors for cell types H1 and IMR90 were calculated via principal component analysis applied to published 40 Kb resolution interaction matrices.[?] Those eigenvectors mapped to previous reference genome builds (hg18/GRCh36) were transferred onto hg19 co-ordinates (GRCh37) using the UCSC LiftOver tool.[?]

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 excluded. Pearson correlation coefficients were 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.[?] After this filtering, 11 chromosomes remained (1-3, 6, 11-16 and 18), a total of 1311 $\times$ 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.[? ?] However, these concerns are lessened for the purposes of characterising large 1 Mb blocks in the most general terms (i.e. eigenvectors reflecting compartmentalisation), 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,[?] 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 MACS2[?] 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 (RF) regression,[?] an established machine learning approach, was used as implemented in the R package `randomForest`.[?] The RF algorithm 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,[? ?] typically providing low bias and low variance predictions without the need for variable selection.[?] Additionally the RF method represents an example of "algorithmic modelling" [?] in that it makes no assumptions about the underlying data model. Parameters of $mtry = \frac{n}{3} \approx 11$ and $ntrees = 200$ were assumed as they are known to be largely insensitive;[? ?] 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.[? ?]

2.2.2 Graphical lasso

Regularised models made use of the Graphical LASSO[?] (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.[?]

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.[?] The algorithm minimises a negative log-likelihood (Eqn. **??**[?]) 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.

$$\underset{\mathbf{\Theta} \prec \mathbf{0}}{\text{minimise}}\ f(\mathbf{\Theta}) := -\log\det(\mathbf{\Theta}) + \text{tr}(\mathbf{S\Theta}) + \lambda\|\mathbf{\Theta}\|_1 \tag{2.1}$$

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 $A > 0; B \leq 0$, 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 produce simulated datasets to calculate the significance of the 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".[?] 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.*[?], 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 ($p \geq 0.1$).[?] 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[? ?] and even higher primates.[?]

3. Results

3.1 Concordant compartmentalisation across cell types

Genome compartments proved well-conserved across human cell types (Fig. **??**), with Pearson correlation coefficients between eigenvectors from all five 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. **??**).

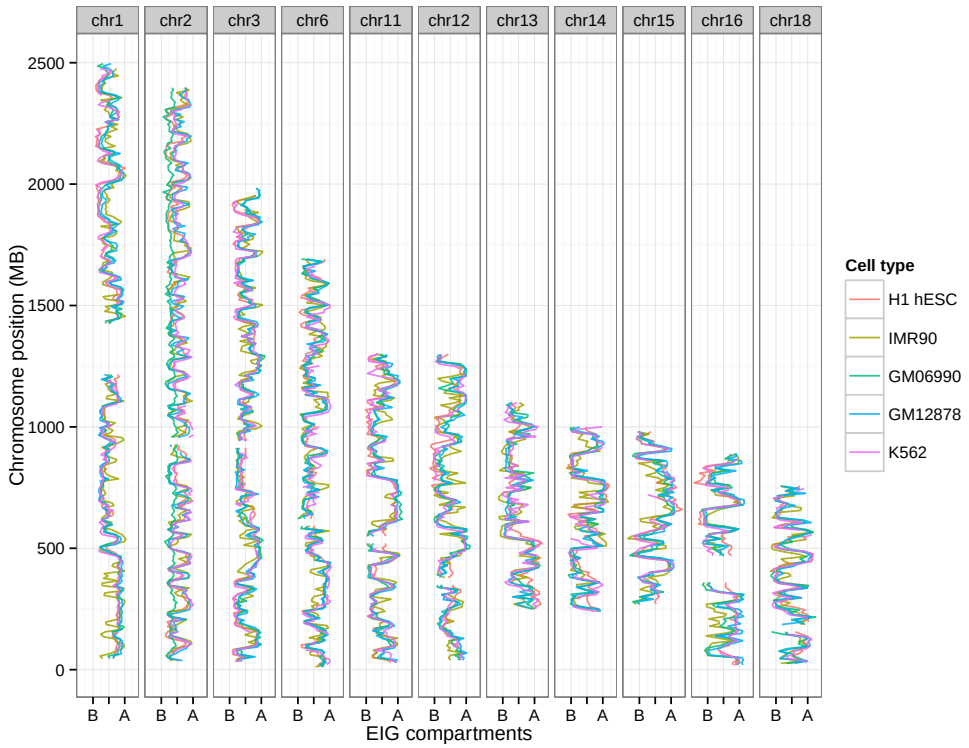


Figure 3.1: **Highly concordant A/B compartmentalisation over selected chromosomes from 5 cell types.** Principle component eigenvectors plotted along their respective chromosomes for five different cell types. "A" and "B" labels reflect compartments with positive and negative eigenvectors respectively,[?]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

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 ($p \approx 0$) between predicted and empirical eigenvector values (Fig. **??**). These correlations suggest accuracies approaching those of successful quantitative models of transcriptional output constructed using locus level chromatin features.[?]

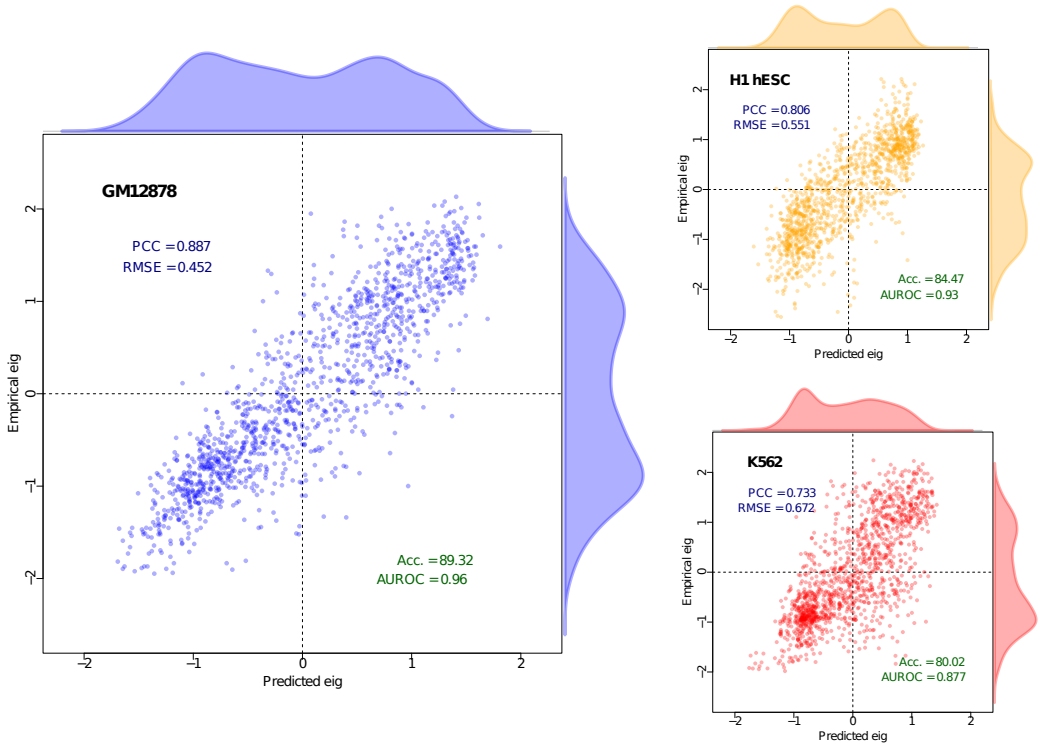


Figure 3.2: **Accurate models of eigenvector values across three cell types.** Predicted eigenvector values are plotted against their actual recorded values for each cell type. Evaluation metrics shown are Pearson's correlation coefficient (PCC), root mean-squared error (RMSE), and classification evaluators (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 variables, it was then of interest to identify which of these were most influential in the Random Forest (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 variables, 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 LASSO[?] (least absolute shrinkage and selection operator; hereafter glasso) calculates an estimate analogous to a measure of pairwise conditional independence between nodes,[?] and was selected over competing methods for several resons: (a) due to the geometry of L_1-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;* (c) fast speed of execution.[? ?]

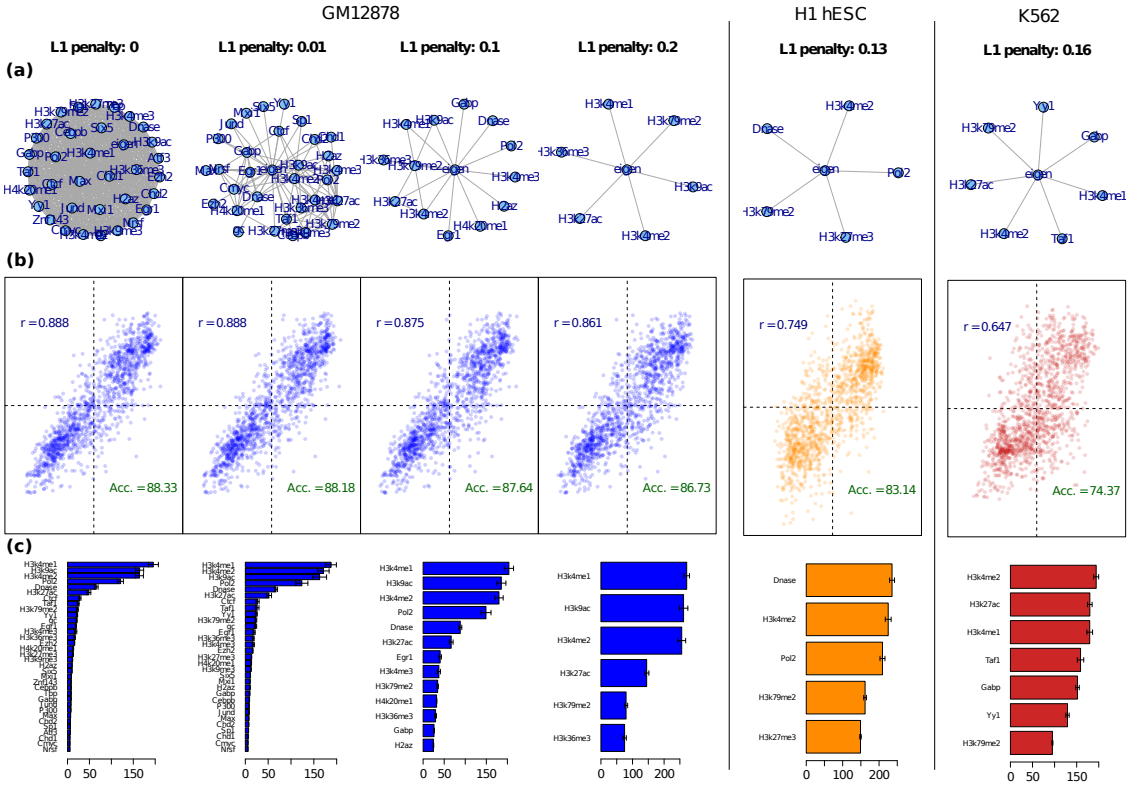


Figure 3.3: **Parsimonious models reveal influential features in predicting nuclear architecture** (a) Graphical models produced by the glasso algorithm[?] for varying regularisation parameter λ. (b) Corresponding RF results using these reduced feature sets as in Fig. **??** along with (c) variable importance estimates in terms of mean decrease in accuracy (see Methods **??**). Full λ sequences for K562 and H1 are given in the supplementary materials (Figs. **??**, **??**).

*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.

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.

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 selecting equally-sized random subets of variables, it can be shown that the size of the intersection between all three sets is significantly larger than would be expected by chance ($p = 9.6 \times 10^{-3}$), yet overall there remains a surprising disparity between cell types given the observed correlations of the response variable (Results **??**).

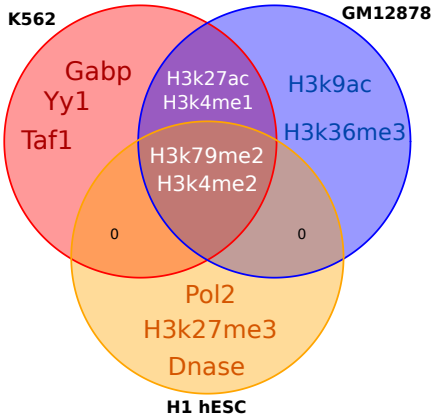


Figure 3.4: **Influential features vary among cell types.** Venn diagram showing the intersections of features remaining after regularisation in parsimonious models of nuclear architecture (Fig. **??**).

3.4 Invariant regions of higher order structure 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 structural variability, or relatively cell type invariant blocks, could be significantly more accurately predicted relative to high structural variability regions (GM12878: $t_{16} = 2.1, p = 0.051$, H1: $t_{17} = 4.4, p = 3.7 \times 10^{-4}$; K562: $t_{12} = 15, p = 3.8 \times 10^{-9}$; 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 ($t_{17} = 7.1, p = 2 \times 10^{-6}$).

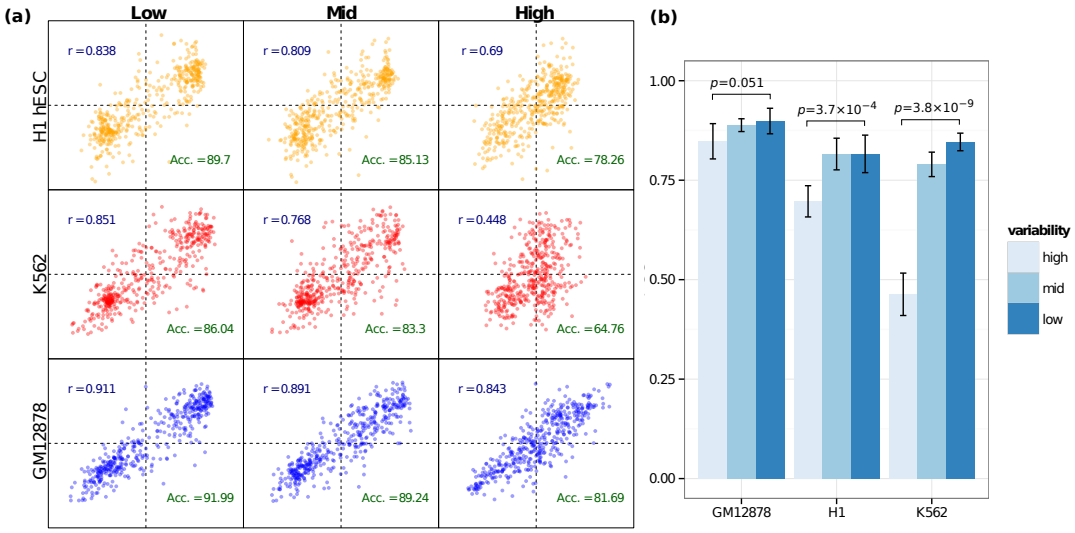


Figure 3.5: **Regions of variable structure are more difficult to model than those that are stable across cell types.** (a) Scatterplots comparing predicted and empirical eigenvector values for subsets split by variability. (b) Bar chart showing the average Pearson's correlation coefficient (PCC) over all 10 folds, with 95% confidence intervals indicated. "Low" variability regions, the third 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 that 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, or those regions most affected by artefacts of the Hi-C data processing.

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

3.5 Models differ between cell types

The cell type specific models of higher order structure were cross-applied, such that the locus-level chromatin features of one cell type were used as a feature set in a RF regression model trained in a different cell type. In each instance of cross-application, the models' performance in predicting the chromatin state in foreign cell types decreased (Fig. **??**).

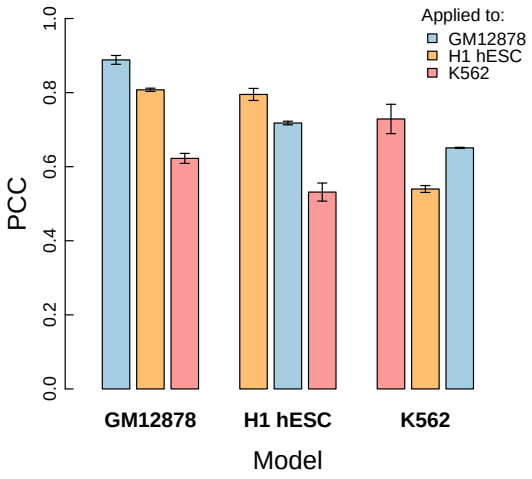


Figure 3.6: **Cross-application of models results in decreased prediction accuracy** The PCC between predicted and empirical eigenvector values is shown (with 95% confidence intervals) for models comparing their performance using feature sets from the same cell type against those from the other two.

In each case of cross-application, the results reflect the degree of cell type specificity of the model (Fig. **??**). For example, GM12878 higher order structure is more accurately predicted by the H1 model than K562 structure ($t_{10} = 14.6,\ p = 6 \times 10^{-8}$). Overall, K562 feature sets result in the lowest prediction accuracies using GM12878 or H1 models (Fig. **??**).

3.6 Models generalise to unseen chromosomes

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

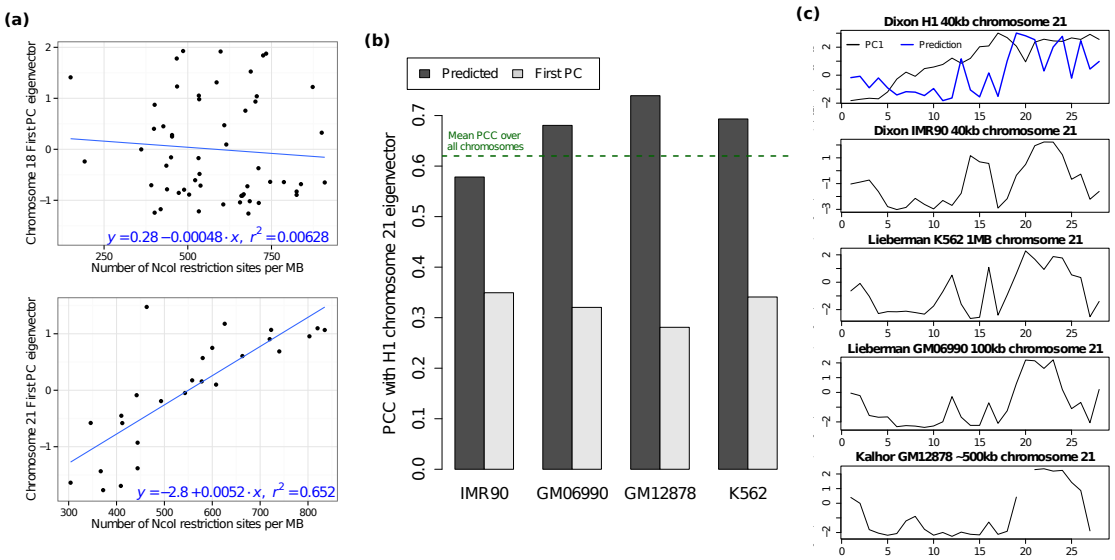


Figure 3.7: **Prediction of eigenvectors for a chromosomes whose first principal component does not describe its compartmentalisation.** (a) The first principal component of the H1 chromosome 21 interaction matrix is largely explained by the number of restriction sites per megabase ($R^2 = 0.65$; *cf.* chromosome 18: $R^2 < 0.01$). (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). This led to chromosome 21 being excluded from the main analysis (see Methods **??**). However, by applying the RF regression model for H1 developed using other chromosomes to the features on chromosome 21, new predicted eigenvector values for H1 chromosome 21 were produced (Fig. **??**c). The predicted eigenvector values for chromosome 21 in H1 proved much more similar to eigenvectors from the same chromosome in other cell types (Fig. **??**b). Hence this prediction appears to reflect the genuine compartmentalisation of chromosome 21 in H1 with reasonable accuracy.

4. Discussion

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

The relationship between locus-level chromatin features and higher order structure remains poorly understood. This work has shown that strong correlations 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.[?] A means of distinguishing between these two explanations could be a biological perturbation study, with specific factors (e.g. 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 causal role for such factors in higher order chromatin structure. Reversing the changes by the addition of these factors to deficient cells would strengthen the case further.

Interestingly, the DOT1 histone methyltransferase which methylates H3K79 has previously been linked with DNA stability in yeast.[?] A study of the mammalian orthologue DOT1L reported that despite correlating with transcription, downregulation of this enzyme left most genes transcribed at their normal rate.[?] The same study linked the "parallel nature of H3K4 methylation and H3K79 methylation",[?] implicating co-operation with the other histone

mark (H3K4me2) found in all parsimonious models (Fig. **??**). Similarly, factors disporportionately important in the nuclear architecture of a single cell type might be manipulated to study the effects on cell type specific structures.

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 position along a chromosome arm.[?] This type of explanation ties in with the "dog-on-a-lead" model of chromosome topology,[?] 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.[? ?] It goes on to postulate that some regions, such as centromeres, are dominant over their chromosomal constraints and hence have disproportionate influence on local genomic interactions.

This model of chromosome topology is consistent with the observed high correlation of compartments across cell types (Fig. **??**), with constrained chromosomal territories dominant in organising the invariant regions (encompassing perhaps three quarters of the genome) while cell type specific chromatin states are responsible for the observed variable regions. Indeed, some support for this hypothesis can be found by contrasting relative variable importance metrics between high and low variability models (Figs. **??**, **??**), where the predictive power of GC content is decreased in all cell types when comparing low variation regions with those that are highly variable. This could also 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.

5. Conclusion

We have shown that higher order chromatin structure can be accurately predicted using aggregate locus-level chromatin features. Of these, H3K4me2 and H3K79me2 appear to be of particular importance across all cell types. We also note that despite a general concordance of compartment states across cell types, there exists a surprising degree of divergence between cell-type specific models, both in terms of relative variable importance and according to 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 activity.

6. Additional figures

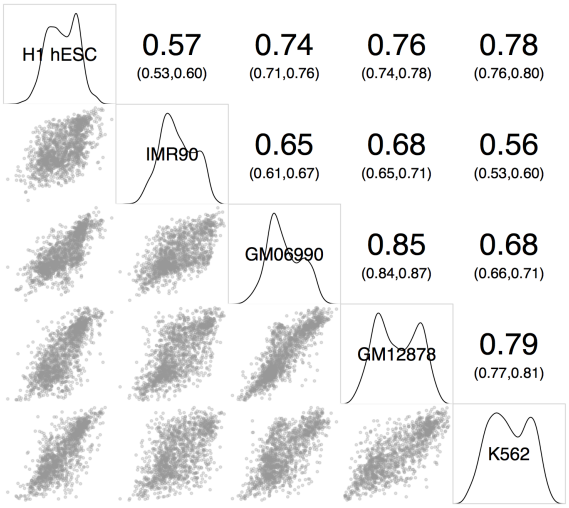


Figure 6.1: **General concordance of eigenvectors across cell types.** Correlogram showing the mean correlation of eigenvectors across selected chromosomes of five human cell types. Pearson's correlation coefficient is shown (*upper*) along with kernel density estimates of each cell type's eigenvector distribution (*diagonal*) and scatterplots comparing megabase blocks from each cell type (*lower*).

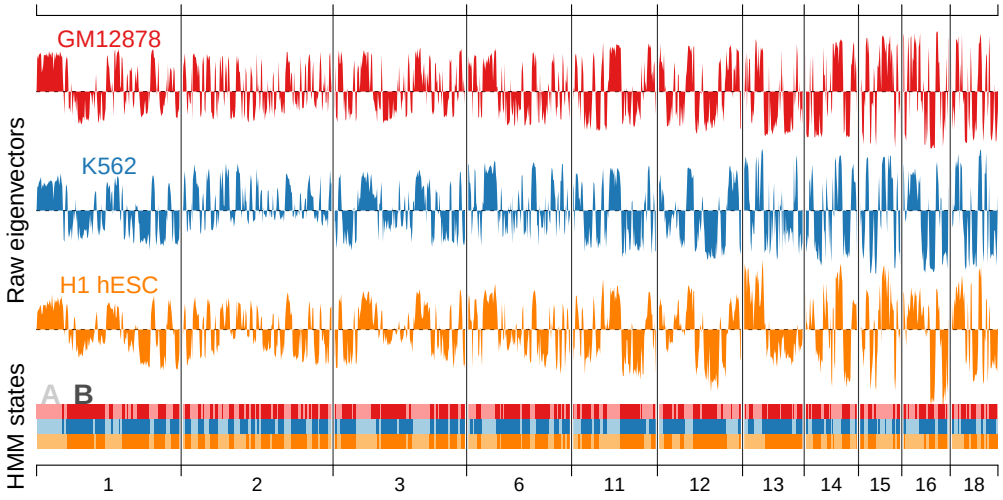


Figure 6.2: **Eigenvectors and HMM state calls for selected chromosomes across three cell types.** 72.6% of HMM state calls (*lower*) are in agreement across 1 Mb blocks in three human cell types: GM12878, K562 and H1 hESC.

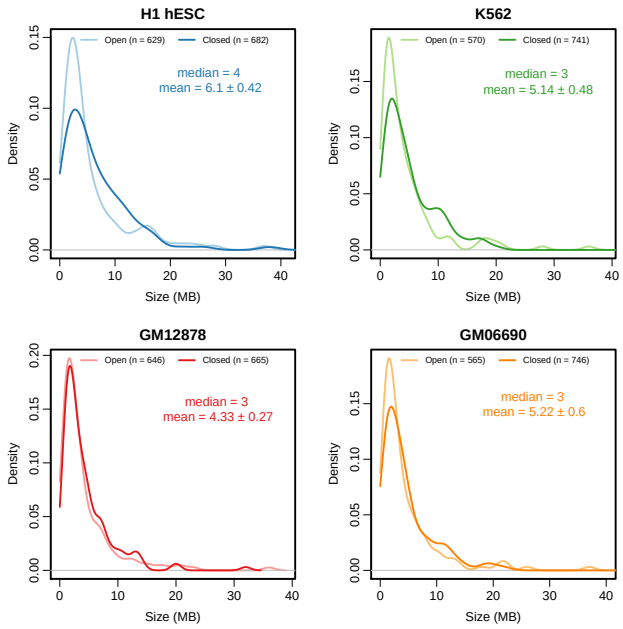


Figure 6.3: **Characterisation of open and closed compartment sizes in various cell types.** Density plots showing the size distributions of open and closed chromosome compartments. Means are shown with 95% confidence intervals. n refers to the number of 1 Mb blocks classified as either open or closed.

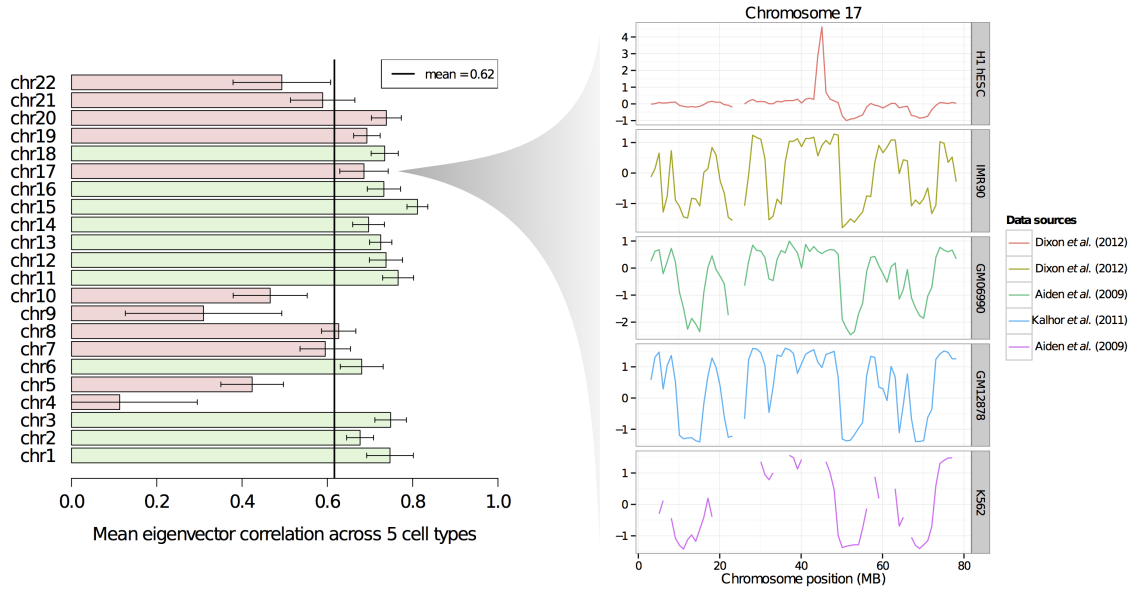


Figure 6.4: **Selection of correlated chromosomes for modelling.** Generally those chromosomes whose mean correlation across all 5 cell types was 1 standard error above the mean were taken forward as examples of properly-formed PC eigenvectors reflecting A/B compartmentalisation. Some chromosomes meeting 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*).

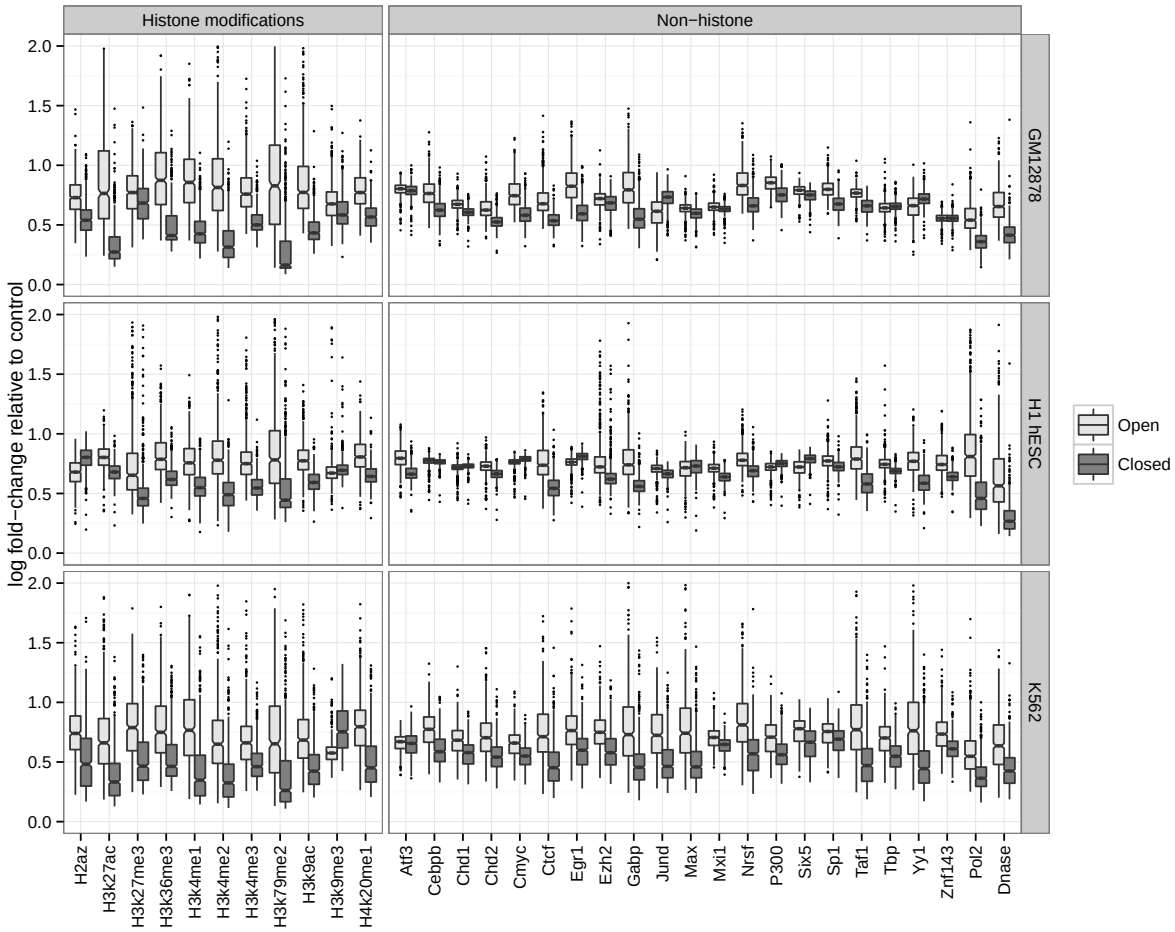


Figure 6.5: **Characterisation of open and closed compartments in terms of cell-matched locus level data.** Each variable distribution is depicted as a box-and-whisker diagram for open and closed megabase blocks. The y-axis represents the $\log_e$ fold signal change relative to ChIP-Seq input control per Mb.

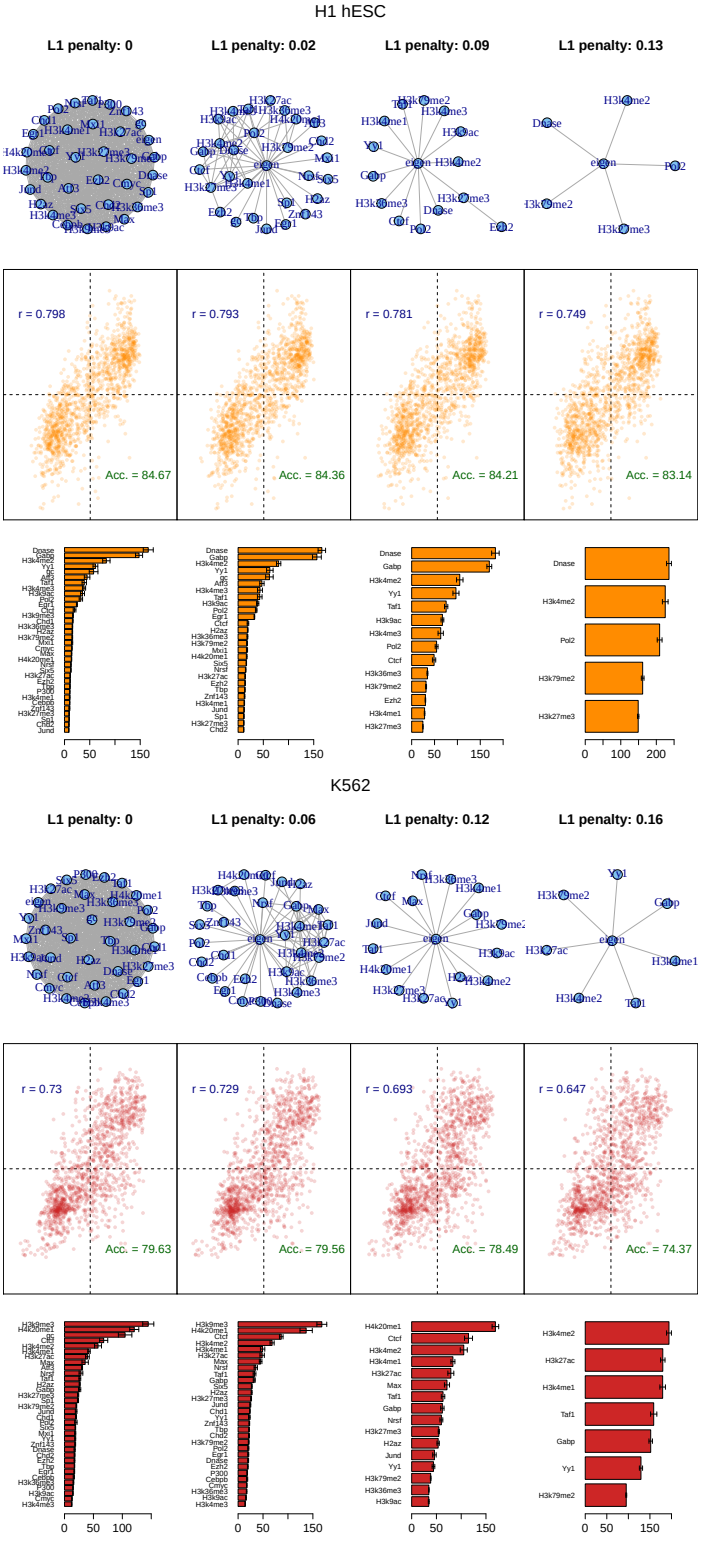
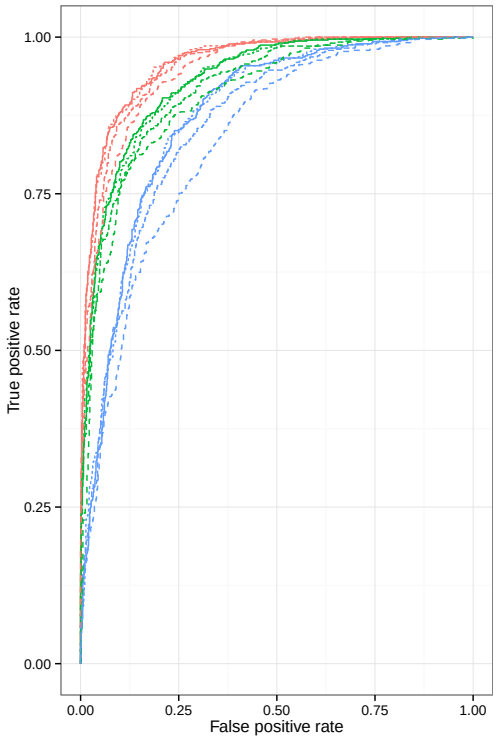


Figure 6.6: **Graphical Lasso L_1 *norm*-based regularisation evolves parsimonious graphical models for feature selection.** Graphical models produced by the glasso algorithm are shown for varying regularisation parameter λ (*upper*), as in Figure **??**. Here the results are shown for cell types H1 (*upper*) and K562 (*lower*).



cell type	model	AUC	nvars
GM12878	1	0.9583	34
GM12878	2	0.9589	31
GM12878	3	0.9520	13
GM12878	4	0.9443	6
H1 hESC	1	0.9308	34
H1 hESC	2	0.9292	28
H1 hESC	3	0.9171	14
H1 hESC	4	0.9029	5
K562	1	0.8728	34
K562	2	0.8735	30
K562	3	0.8587	16
K562	4	0.8310	7

Figure 6.7: **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.

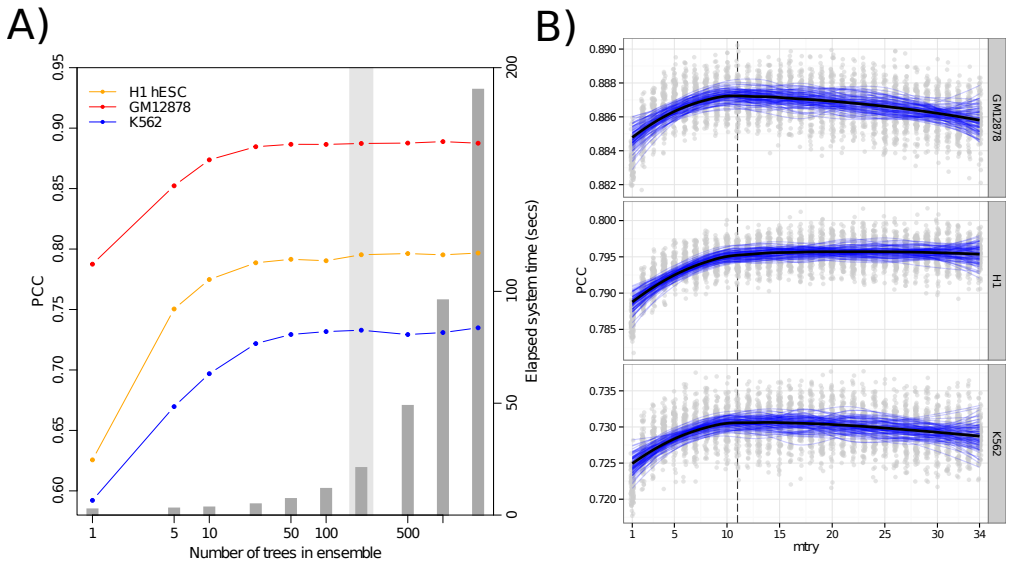


Figure 6.8: **Random Forest parameters proved largely insensitive to parameters ntrees and mtry.** The number of trees, *ntree* (A), was chosen as 200 and the number of variables tested at each node, *mtry* (B), was the default value for regression: $n/3 \approx 11$.

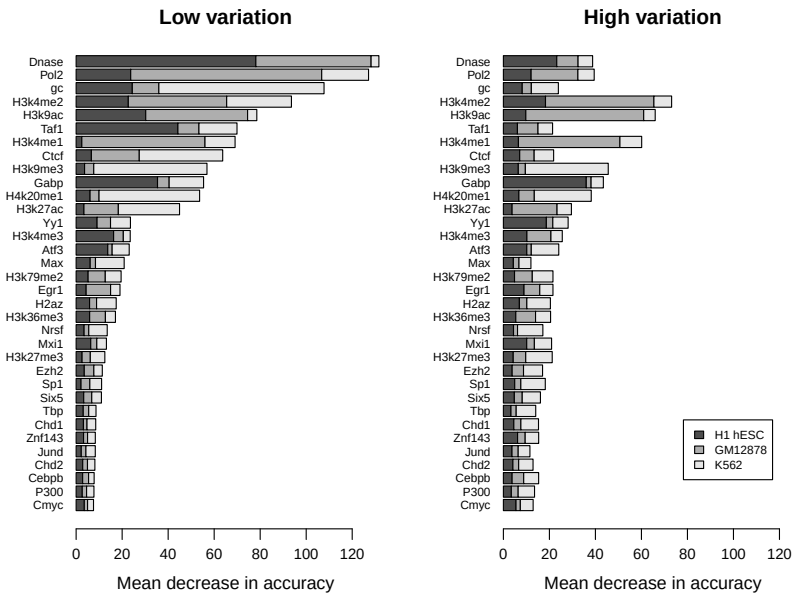


Figure 6.9: **Relative variable importance for RF features in models built with blocks that are conserved between human cell types (low variation) and those that are variable (high variation).** Mean decrease in accuracy (Methods **??**) is calculated for each variable in three human cell types and compared between the third of regions with the lowest mean absolute deviation across cell types (*left*) and the third with the highest (*right*).

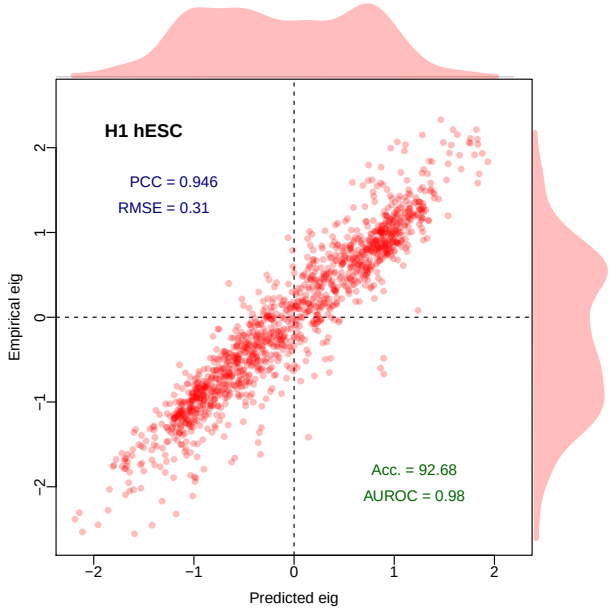


Figure 6.10: **Autoregressive terms increase model accuracy.** When adjacent eigenvector values (y_{i-1}, y_{i+1}) are used as features in predicting the state of a central Mb block (y_i) the model accuracy greatly improves. This is shown above for the H1 cell type, highlighting the known non-independence of adjacent eigenvector values.

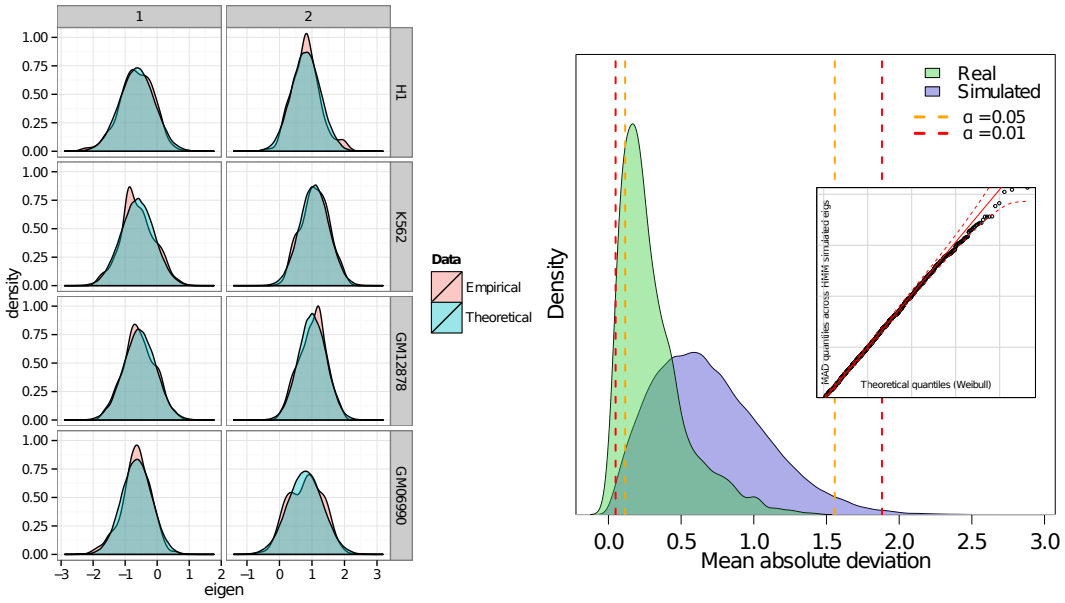


Figure 6.11: **Estimating the significance of median absolute deviations observed between eigenvectors across cell types.** Two state normal HMMs were fit to observed eigenvectors in each cell type (*left*) and the median absolute deviation was calculated at each position. The distribution of these values was approximated by a Weibull extreme-value distribution (with $k = 1.87$, $\lambda = 0.76$; QQ-plot *right, inset*) and the quantiles of this distribution were used to determine the significance of observed values (*right*).