1 INTEGRATIVE MODELLING AS A TOOL TO EXPLORE BIOLOGICAL SYSTEMS

1.1 INTRODUCTION

Large-scale chromatin data has recently been produced by multiple consortia, most notably ENCODE (Section **??**) but also the NIH Roadmap Epigenomics project.[1] The breadth and depth of this new data offers unprecedented opportunities to advance our understanding of the complex biology of the chromatin landscape. To this end, studies have already enjoyed success in integrating these data through modelling techniques, with the subsequent dissection of these models revealing novel insights into complex biological phenomena.

Recent studies have shown convincingly that local chromatin state measurements can accurately predict expression levels of genes on a genome-wide basis. Tippmann *et al.*[2] designed a linear model to predict steady-state mRNA levels in mouse embryonic stem cells based on just four predictors: 3 histone modifications (H3K36me3, H3K4me2 and H3K27me3) and Pol-II occupancy. Remarkably, the linear model was found to explain 84.6% of an estimated 91% maximal variance that could be explained (as calculated through a detailed determination of noise).[2] 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."[2] 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.[3]

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.[4] The authors here developed a two-stage model which first attempts to classify each transcription start site (TSS) into an 'on' or 'off' state using a powerful ensemble classifier technique called Random Forests (RF). The second stage of the model used the same range of histone modifications as regressors in a simple linear modelling framework to quantify predicted expression. This approach proved very successful, producing a median Pearson correlation coefficient (PCC) between predicted and empirical expression levels using 10-fold cross-validation of 0.83 across all cell lines and expression level technologies.[4] Additionally, this study highlighted

cap analysis of gene expression (CAGE) as the technology, relative to RNA-seq and RNA-PET, which produced the most predictable expression response. CAGE uses 5′ capped transcripts to generate short, specific tags which precisely identify TSS positions as well as quantifying the abundance of a given transcript.[5,6]

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

1.2 REPRODUCING DONG *et al.*

We reimplemented the published modelling framework of Dong *et al.*[4] to replicate their results and analyse the strengths and caveats of their approach. We were able to reproduce the reported results and highly accurate models of transcriptional output (Fig. XX). In doing this, we were surprised to find that the two-step classification then regression (firstly assessing a gene as 'on' or 'off' and then predicting its expression level) added little additional accuracy relative to a simple linear regression model (Fig. 1).

1.2.1 *Bestbin* method

An innovative element of the modelling approach used in Dong *et al.*[4] is the 'bestbin' method of matching chromatin measurements to the expression of a given TSS. This strategy first bins normalised signal intensities into 40×100 bp bins encompassing 4 kb around the TSS, and adds an additional bin representing the remaining gene body. Then the correlation between the signal of a given mark and the expression of a TSS across all genes is measured, then the bin producing the highest correlation is designated as the 'bestbin' and that bin's normalised ChIP-seq signal intensity is taken forward for the full model. This was shown to raise the correlation (between predicted and observed expression) by 0.1 in the simple regression model, an increase in accuracy of almost 13%, relative to simply taking the average value across all bins.[4]

1.2.2 Model adjustments

We attempted to improve the accuracy of predicted expression values produced by Dong *et al.*[4] through increasing the number of informative regressors. While Dong *et al.*[4] included broad coverage of different histone modifications, they did not investigate the impact of higher order chromatin data. For this reason, we matched

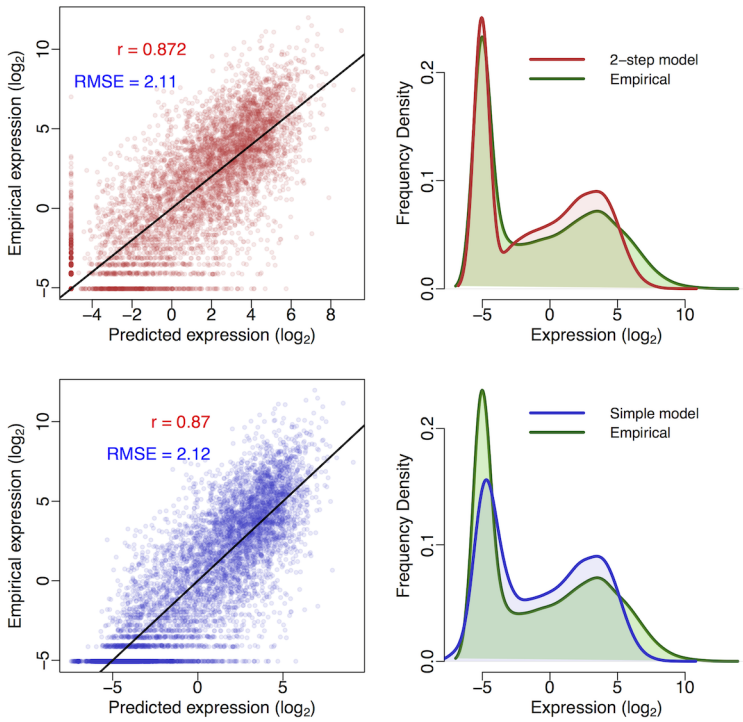

Figure 1: Comparison of a published two-step classification-regression model of transcription with a simple linear regression model. Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for the two-step model of Dong *et al.*[4] and simple multiple linear regression are shown (*left*) along with frequency distributions of predicted and observed expression levels (*right*). Scatterplots are annotated with Pearson's correlation coefficient (r) and the root mean squared error (RMSE); the black trendlines describe $y = x$. Following 10-fold cross validation, overall correlation coefficients were: linear model $0.87 \pm 1.77 \times 10^{-5}$; Two-step model $0.872 \pm 9.89 \times 10^{-5}$.

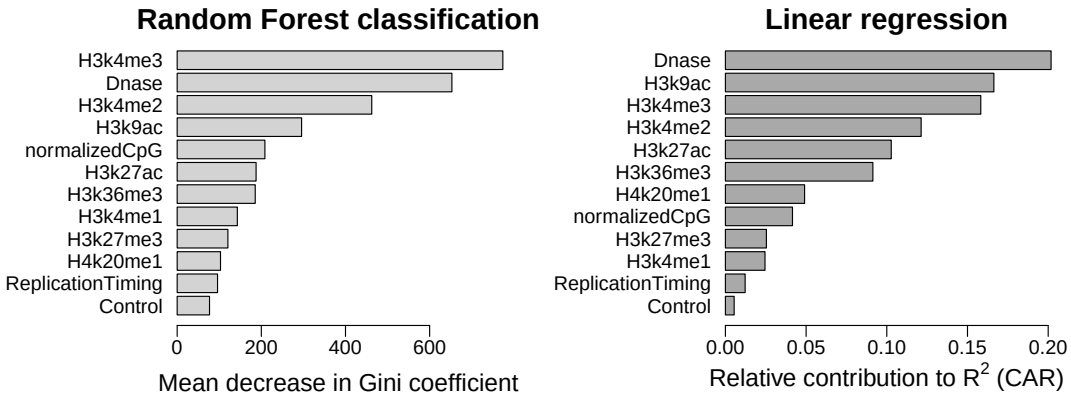

Figure 2: Relative importance metrics for variables in both stages of a reimplementation of a published model for predicting transcriptional output. Variable importance is measure by decrease in Gini coefficient for the RF classification step, and by CAR R^2 decomposition[7] for the linear regression step.

the TSS positions used in Dong *et al.*[4] with previously-published genome-wide replication timing ratios measured in BG02 ESCs.[8] This data is of a different origin to the transcriptional data in this case (which was record in H1 hESC) but replication timing is thought to be largely conserved between cell types.[9]

We then used these values as an additional regressor in both the two-step classification regression model and the simple linear model but saw no significant improvement in either model's accuracy (*data not shown*). The reasons for this are likely that the data were relatively low-resolution (1 Mb), from a imperfectly matched cell line and also that the existing model is already achieving such accurate results that they must already be accounting for most of the maximal explainable variance in gene expression given experimental and biological noise. With this in mind, additional regressors would be expected to yield diminishing returns. Even so, on closer examination, the replication timing data appeared only slightly more informative than the control ChIP-seq input measurements when evaluated with relative importance metrics (Fig. 2), implying that large-scale chromatin domains and do not have significant influence on the expression of the genes resident within them.

1.3 MODELLING FANTOM5 EXPRESSION DATA

Using unpublished FANTOM5 CAGE data and the approach established above, I next attempted to model gene expression at timepoint zero (t_0) of a differentiation timecourse of Human H1 embryonic stem cells (H1 hESC) to CD34+ hematopoietic stem cells.

The first stage of the analysis was to map each CAGE cluster to a representative TSS. FANTOM5 robust gene mapping[10] provided corresponding Entrez Gene IDs for gene-associated CAGE clusters, and we selected the most expressed cluster to represent the expression level of its mapped gene. We then compared these to Ensembl TSS annotations (v69) and discarded those tag clusters centered on a point > 50 bp from an annotated TSS associated with the mapped Entrez Gene ID, thereby removing enhancers and other non-genic transcribed regions.

Next we retrieved a number of genome-wide histone modification datasets from the ENCODE and NIH Roadmap consortia which were measured in H1 hESC cells, taking these to be reflections of the chromatin state t_0. I implemented the previously-described 'bestbin' strategy[4] (Section 1.2) to objectively select the most-correlated binned signal for each chromatin H1 hESC mark. To explore this approach, we analysed the stability of chosen bestbins by calculating them on 200 sets of 1000 randomly selected TSS samples, with each sample representing approximately 8% of the dataset (Fig. 3).

Figure 3 shows that bestbin selections are often consistent, indicating there are predictably informative regions relative to a TSS for each chromatin factor. Further-

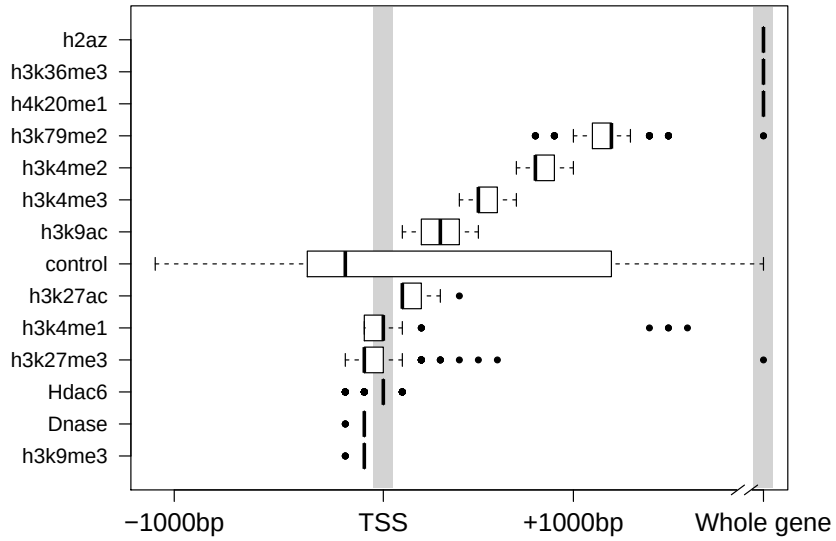

Figure 3: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, DNase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 kb flanking 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.

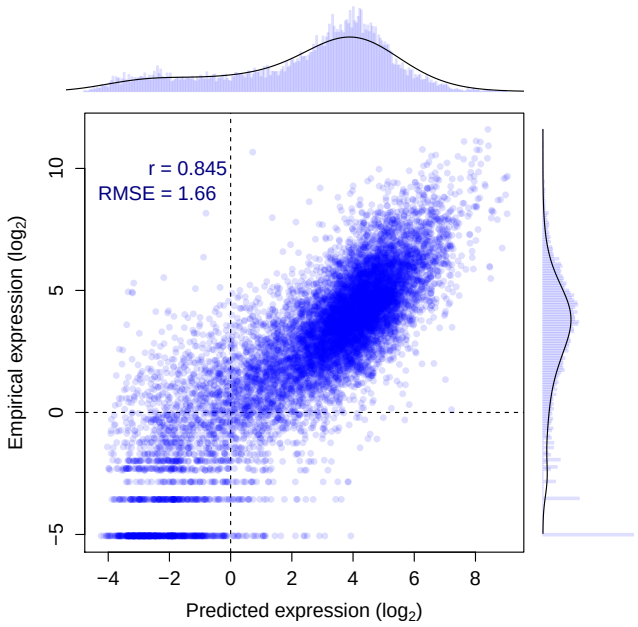

Figure 4: Random Forest predictions of FANTOM5 expression data. RF model predictions are plotted against their empirical values. The marginal distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson's correlation coefficient (r) and the root mean-squared error (RMSE) are shown (*inset*).

more, 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.[2,11,12] The negative control variable (ChIP-seq input) shows no strong location bias, as expected (Fig. 3).

Having matched a variety of genome-wide H1 hESC chromatin datasets to the FANTOM5 timecourse expression data, we then built a regression model using a Random Forest (RF) approach.[13] This method outperforms a simple linear model in initial comparisons and is able to capture non-linear relationships as well as interactions without them being explicitly specified.[14]

Figure 4 shows the resulting predictions of a preliminary RF model against the actual recorded expression over a test set of approximately 11,000 TSS. This model was built with 15 predictors including control ChIP-seq input, though some of these could be removed without loss of accuracy. The model predictions evaluated with 10-fold cross validation show a significant correlation with measured CAGE levels ($PCC = 0.845 \pm 1 \times 10^{-4}$, $p < 2 \times 10^{-15}$), and the model is able to explain around 71% of the variance in the expression response (for comparison a linear model resulted in $PCC = 0.825 \pm 3.2 \times 10^{-5}$, $p < 2 \times 10^{-15}$). This result is worse than that of Dong *et al.*[4] 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.[4] The RMSEs, when

normalised by the range of observed values, compare more favourably (0.11, compared with Dong *et al.*'s: 0.14).

A possible explanation for our lower modelling accuracy relative to that of Dong *et al.*[4] is that in our case, while both chromatin data and expression timecourse were measured in H1 hESC cells, the experiments took place at different institutes and using unstandardised protocols and cell cultures. For comparison, a previous study using chromatin measurements from a number of different sources to predict expression in a matched cell-type reported a predictive correlation of 0.77.[15] The ENCODE consortium, on the other hand, went to some lengths to standardise protocols and minimise batch effects between samples.[16] Additionally, Dong *et al.*[4] implemented a pseudocount optimisation step whereby an additional count added to each binned signal intensity prior to log transformation to maximise expression correlation. In the model presented above, a fixed psuedocount of 1 was used to avoid introducing positive bias towards higher correlation. Another difference between the two approaches is our use of a single-step model; Dong *et al.* found a small increase in correlation using their classification-regression approach but with the model implemented herein (Fig. 4) 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. 4).

1.4 MODELLING HIGHER ORDER CHROMATIN

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

Previous publications have identified several correlates which track compartment eigenvector profiles to varying degrees,[17,18] 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.

1.4.1 Predictive model

We built Random Forest regression models (Methods **??**) to predict compartment eigenvector profiles genome-wide in three human cell types. Models were found to have high predictive accuracy, with Pearson correlation between predicted and observed compartment eigenvectors in the range of 0.82–0.75 (Fig. 5), comparable to that achieved by Dong *et al.*[4] in the prediction of transcription.

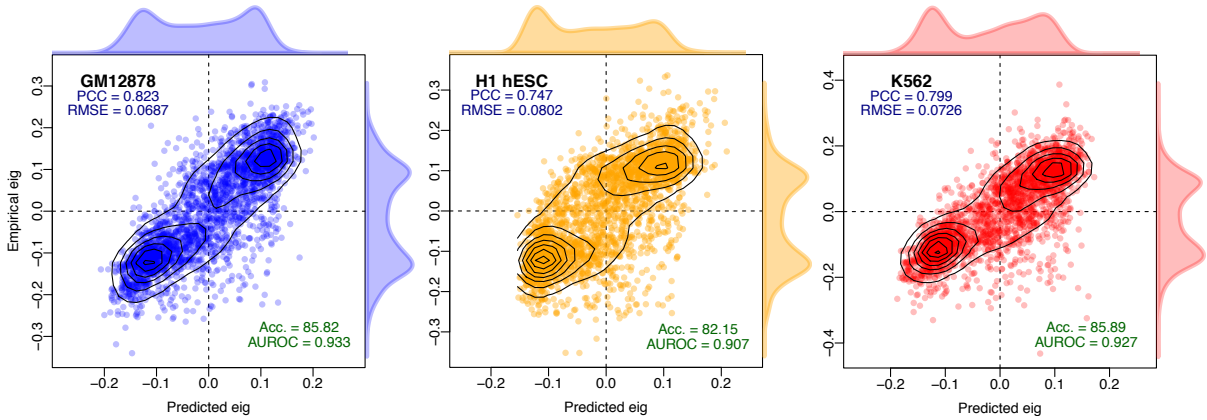

Figure 5: Compartment eigenvector model predictions are highly correlated with observed values. Pearson correlation coefficient (PCC) and root mean-squared error (RMSE) report the degree of success of the regression model, whereas accuracy (Acc.) and area under the receiver operating characteristic (AUROC) give the classification accuracy of binarized outcomes.

Our predictive models were also assessed in terms of classification performance, i.e. did the model correctly assign each block to an A or B compartment. Instead of retraining a classifier and building parallel models, instead for an estimate of classification accuracy we threshold our regression predictions (Methods **??**). We found our Random Forest models achieved high classification accuracy with upwards of 82% of the all genomic bins correctly assigned in each cell type (Fig. 5).

This predictive performance underlines the strong connection between locus-level features and higher order chromatin structure previously noted by Lieberman-Aiden *et al.*[17] Given such highly-predictive models can be generated, it is then of interest to dissect said models in an attempt to understand the nature of this captured relationship.

1.4.2 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.

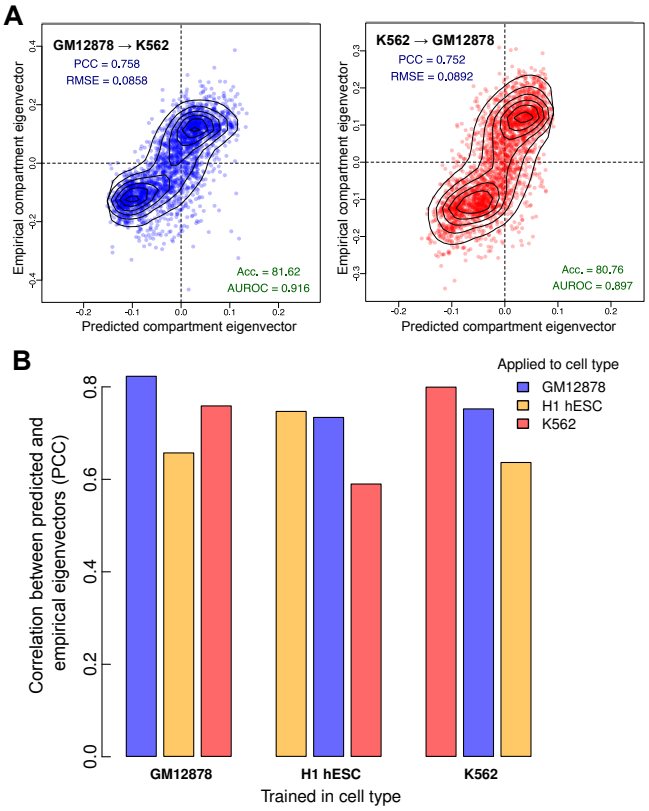

Figure 6: **Models of higher order chromatin structure learned in one cell type can be cross-applied to others.** Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. (A) The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross (PCC = 0.75). (B) In each case, predictive accuracy decreased on cross-application but there remains significant agreement between predicted and empirical values. Acc., accuracy; AUROC, area under the receiver operating characteristic curve; PCC, Pearson correlation coefficient; RMSE, root mean-squared error.

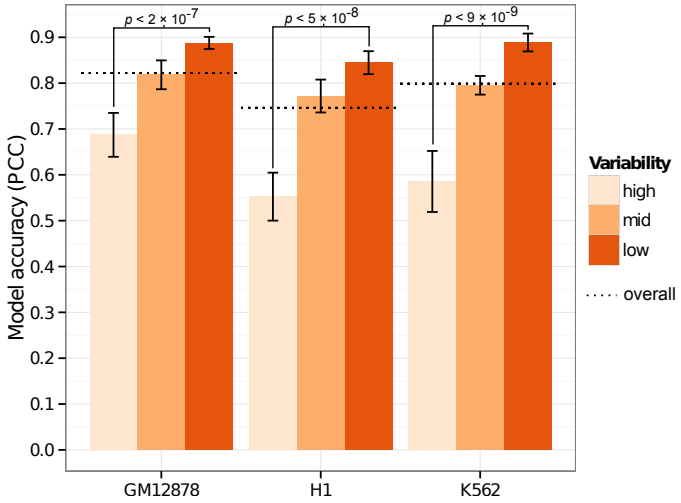

Figure 7: Genomic regions that vary across cell types are modelled less successfully than static regions. Genome-wide compartment eigenvectors were partitioned into thirds according to their median absolute deviation (MAD) across the three cell types under study. Models were fit independently to each third, and the modelling accuracy is compared.

We found cross-application between cell types was possible and with similarly-high levels of accuracy (Fig. 6). 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.

1.4.3 Between-cell variability

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

Indeed we found the most variable regions across cell types were then most difficult to predict through our Random Forest modelling framework (Fig. 7). In each cell type, the third of the genome with the most consistent compartment eigenvectors across cell types could then most accurately be modelled in that cell type, and conversely the most variable third shown significantly depleted predictability (Fig. 7). This latter result suggests these variable regions could either be those which are noisiest, where the eigenvector is least capturing compartment structure, or where cell-type specific biology is influencing compartment structure in each case, in ways not captured by out input feature set and low resolution modelling pipeline.

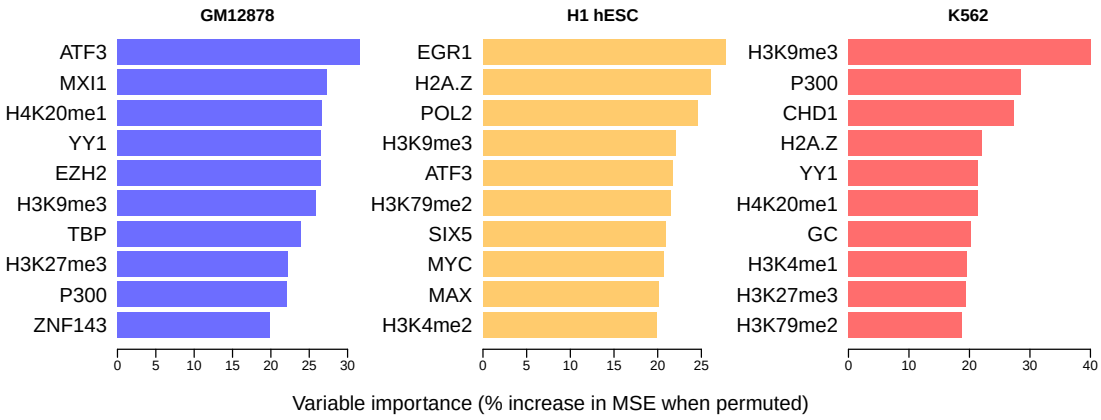

Figure 8: Variable importance per cell type specific model. Variable importance for each Random Forest model was measured in terms of percentage increase in mean squared error when permuted (Methods XX) and the top 10 ranking variables are shown for each model.

1.4.4 Variable importance

Having built accurate predictive models, we next dissect the relative variable contributions made from our range of input features and compare these across cell types. An overview on the top 10 most highly-ranked features in cell type specific models shows some agreement but also substantial differences between cell types (Fig. 8)

Only one input feature, H3k9me3, is present in the top 10 most important variables of each model (Fig. 9). H3k9me3 is one of the few features to be negatively correlated with compartment eigenvectors (e.g. Fig. 11). Of those shared between two cell type models, H3k27me3 is also a repressive mark and deposited by polycomb repressive complex 2 (PRC2)[19] while H2A.Z is a histone variant again linked to polycomb-regulated genes and essential for embryonic development.[20] Furthermore EZH2, the catalytic subunit of PRC2,[21] is also included in the feature set but only highly ranked in the GM12878 cell type model. As another example, MYC and MAX are found in the top 10 influential variables in H1 hESC, while MXI1 is found to be an informative variable in GM12878. This is in keeping with recent results suggesting MYC binds open chromatin as a transcriptional amplifier in embryonic stem cells,[22,23] with MAX and MXI1 long being known as antagonistic co-regulators of MYC.[24] These biological relationships between variables may help explain the observed differences between models: different representatives of correlated clusters of input variables may be being selected in each model (see Section 1.4.5).

To assess the significance of observed intersections (Fig. 9), the variable selection process could be modelled with, for example, a multivariate hypergeometric distribution or via simulation. Simulation was used here for simplicity: each intersection was calculated under 10,000 variables draws with uniform distribution and empirical p-values were then calculated accordingly. Under the assumption that variables are

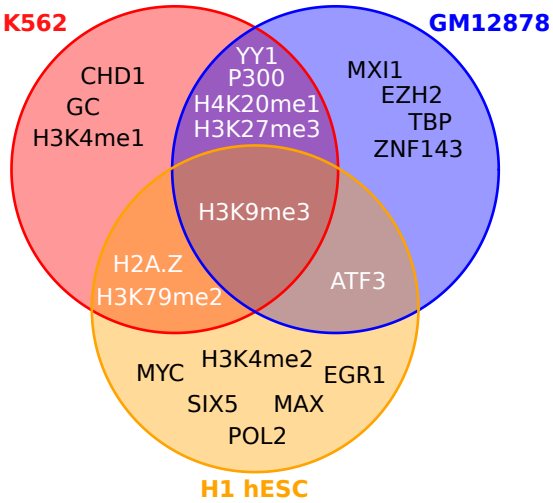

Figure 9: Intersections of the top 10 ranked variables in the cell type specific models. Venn diagram illustrating intersections between sets of ten most influential variables per cell type specific Random Forest regression model of compartment eigenvector (Fig. 8).

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

In addition to rankings, raw variable importance metrics can be compared between cell-type specific models (Fig. 10). This shows that variables such as CTCF have a relatively small but highly consistent variable importance across the three cell type specific models, whereas other features like ATF3 are highly influential in one cell type but not the other two. Absolute differences in these figures should not be over interpreted and will be affected to some degree by data quality, eigenvector calculation and other sources of noise. Nevertheless there are observations which may reflect biological phenomena, such as the higher relative importance of P300 in both hematopoietic cell line models, potentially reflecting its activity as a histone acetyl transferase that regulates hematopoiesis[25] and a noted involvement with CTCF in chromatin looping.[26]

1.4.5 Correlating input features

We have an *a priori* expectation of multicollinearity in our feature set, for example between those that each broadly correlate with transcriptional activity (including POL2,

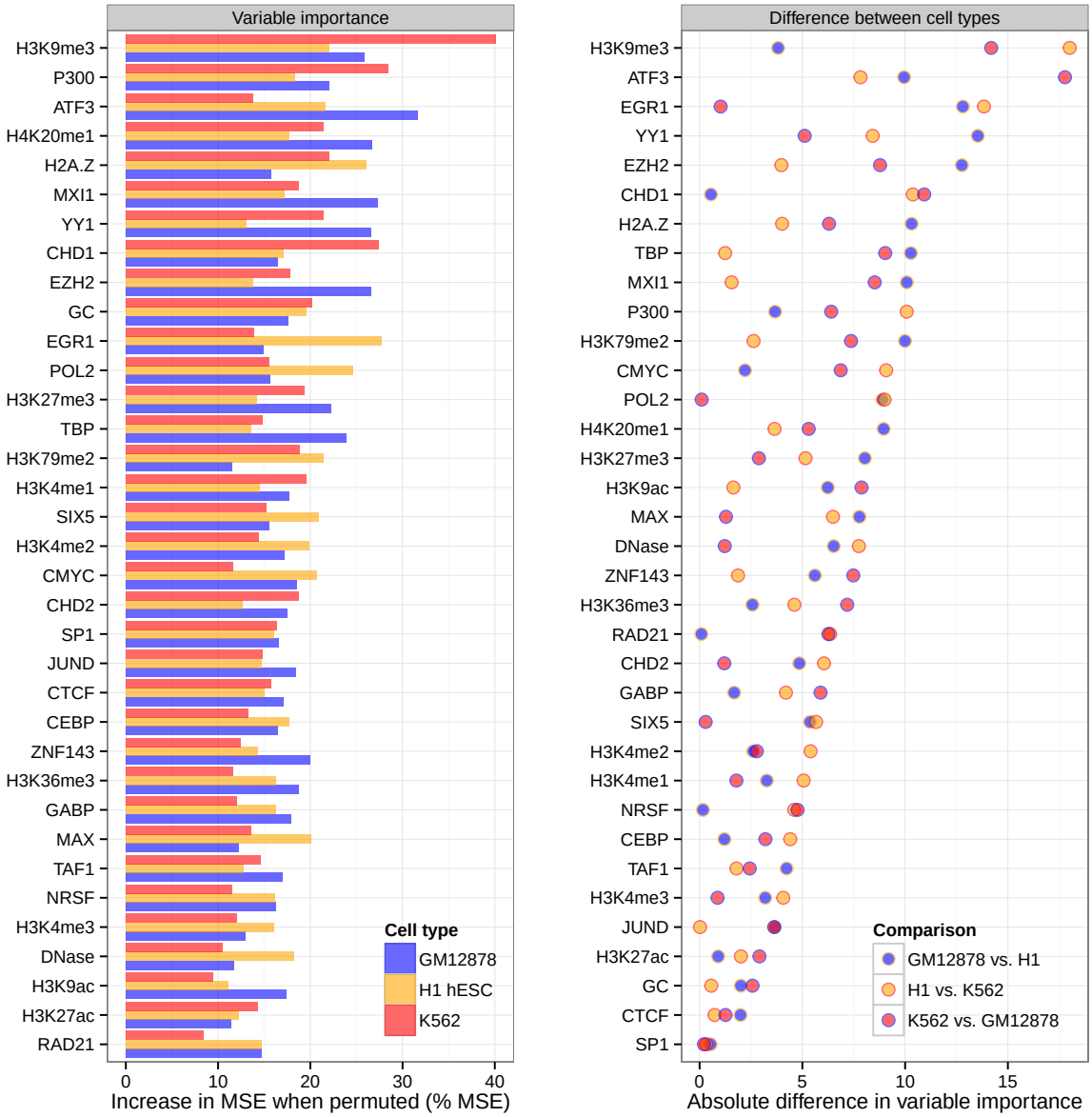

Figure 10: Comparison of variable importance between three cell type specific Random Forest models. Variable importance for each Random Forest model was measured in terms of percentage increase in mean squared error when permuted (Methods XX). Results are shown sorted by mean variable importance (*left*) and by largest absolute difference in pairwise comparisons (*right*).

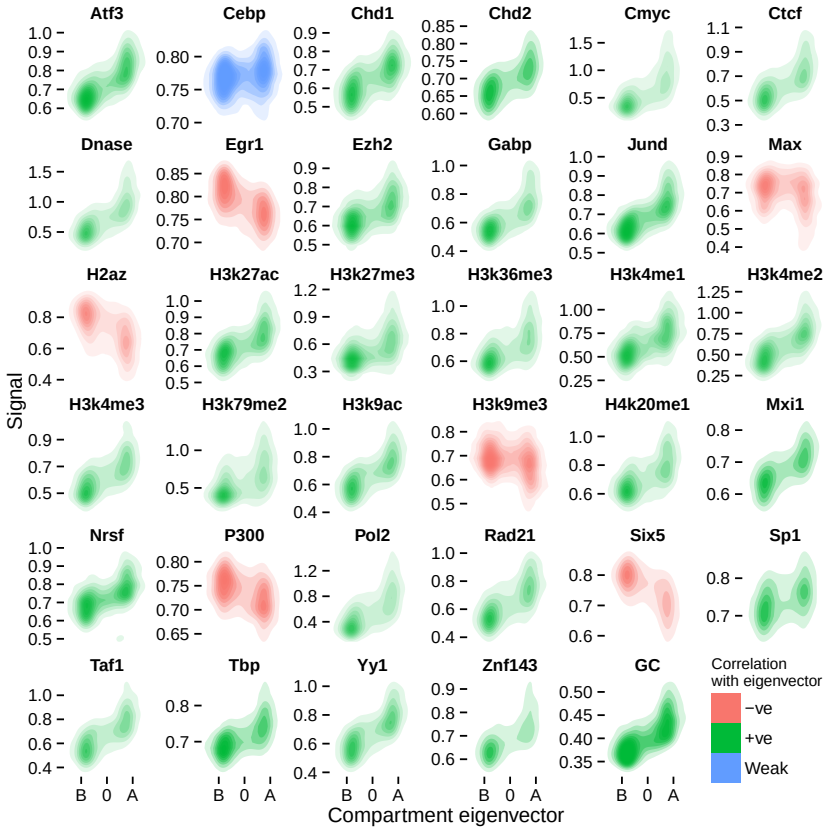

Figure 11: Correlations of individual features with compartment eigenvector in the H1 hESC cell type. Two-dimensional kernel density estimates show the density of points in a scatterplot of compartment eigenvector (x-axis) against each input feature individually (y-axes). Features with a PCC against eigenvector of above or below 0.1 are coloured as positive or negative, respectively.

H3K36me3, GC content). To explore these relationships, we performed unsupervised clustering of our feature sets in each cell type (Fig. 12).

We found as expected pervasive multicollinearity across our feature sets, with the majority of input variables in each model falling into a persistent "active" cluster containing regions with high DNase hypersensitivity, POL2 binding and histone modifications H3K36me3 as well as GC content (Fig. 12).

Outliers are also present. H3K9me3, noted for high variable importance in each model (Fig. 8) and the only feature ranked within the top 10 in each model (Fig. 9) is a clear outgroup in the H1 hESC and GM12878 correlation heatmaps, and in K562 forms a stable cluster only with the P300 transcription factor (Fig. 12). This suggests H3K9me3 is providing orthogonal information to many of the other input variables, and likely explains its high variable importance.

1.5 TECHNICAL CONSIDERATIONS

1.5.1 Resolution

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

Model accuracy when applied to higher resolution input features proved to be similarly high, with empirical PCC being 88 to 95% as high as that at 1 Mb native resolution (Fig. 13).

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.

1.5.2 Other modelling approaches

Random Forest (RF) was *a priori* chosen as an appropriate and powerful modelling tool for this work. Other methods could have been used and should be compared. Here we compare our RF approach with two other options: multiple linear regression and partial least squares regression.

Our results confirm RF as a suitable and powerful approach for modelling our relationships of interest in this work (Fig. 14), with both the highest cell-type specific performance (PCC between predicted and observed $= 0.790$) and on cross-applications (mean PCC $= 0.689$).

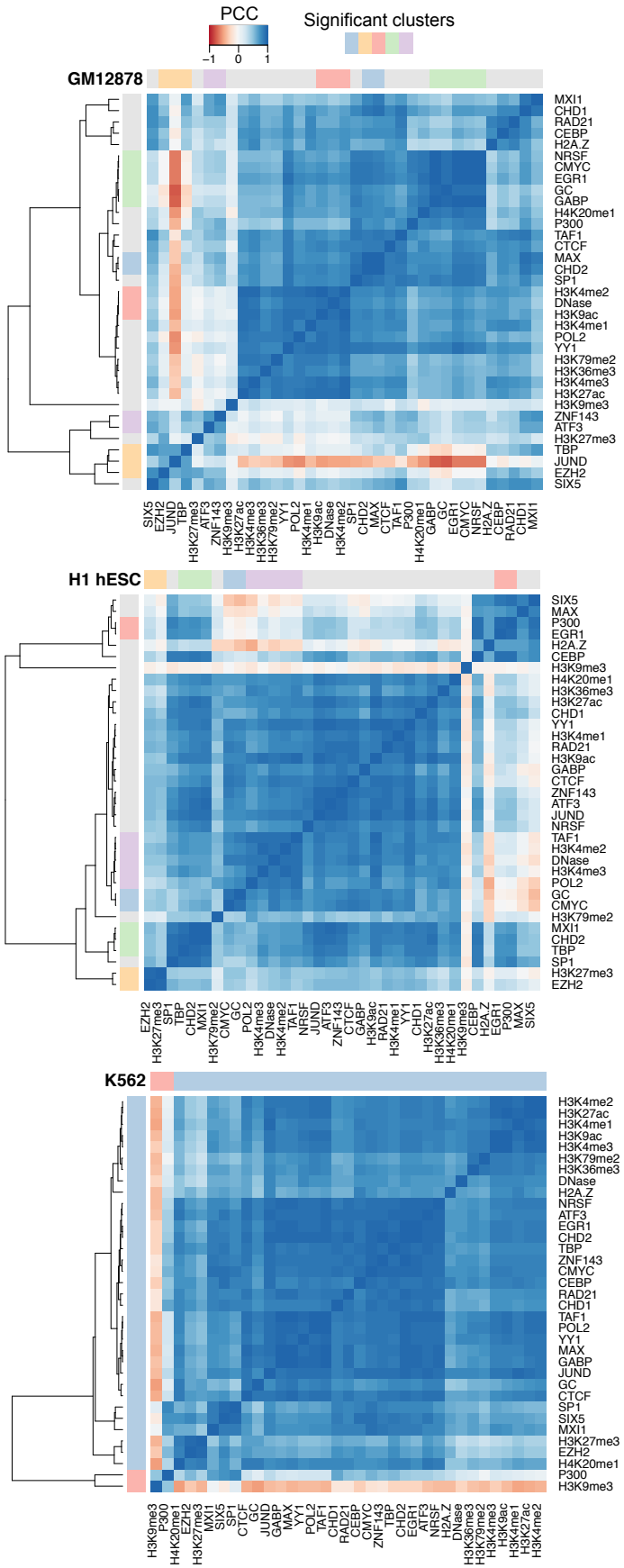

Figure 12: Correlation heatmaps of the 35 features used to model compartment eigenvectors. The Pearson correlation coefficient (PCC) of genome-wide 1 Mb bins of each feature were pairwise correlated with each other. The features were also clustered using hierarchal clustering. The significance of these clusters was determined through multi-scale bootstrap resampling, with those clusters that were stable across different sizes of resampling deemed significant, as implemented in the `pvclust` R package.[27]

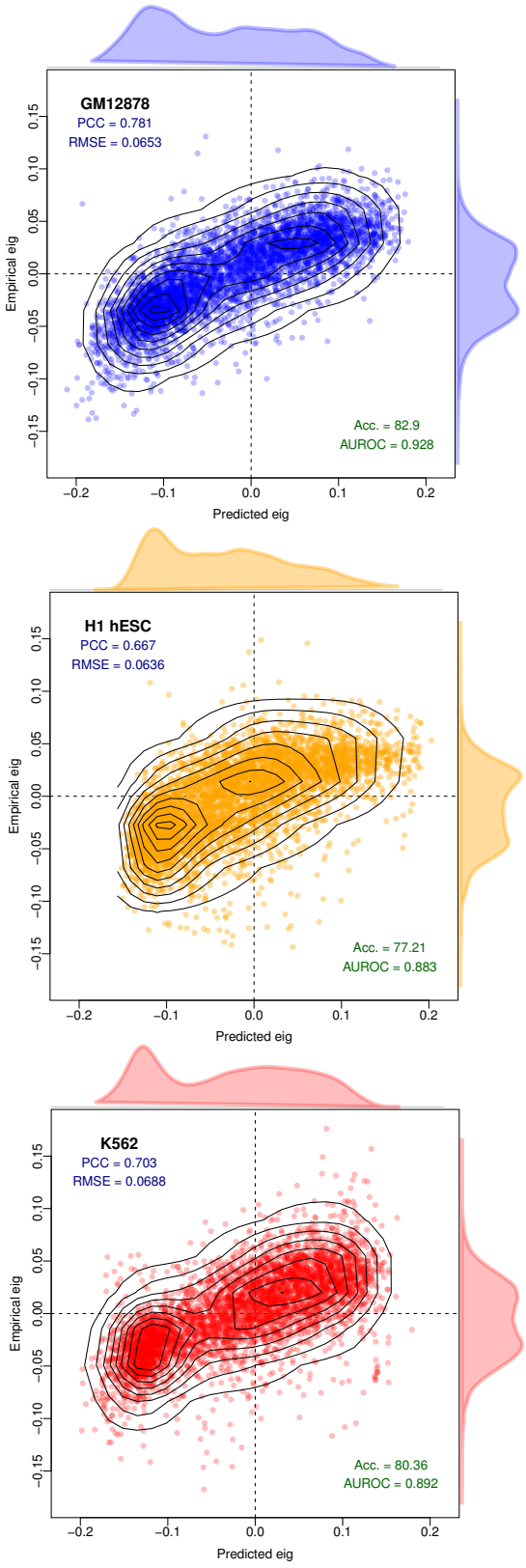

Figure 13: Models learned at 1 Mb resolution can be applied to higher resolution datasets. Despite having been trained on low resolution training sets, the Random Forest models generated can successfully predict compartment eigenvectors at higher resolution (100 kb, a 10× zoom). Eigenvectors at a higher resolution than this do not necessarily reflect A/B compartmentalisation.

Figure 14: Comparison of Random Forest performance with other modelling approaches. Heatmaps show the Pearson correlation coefficient between predicted and observed compartment eigenvectors genome-wide for three regression techniques: multiple linear regression (LM), Random Forest (RF) and partial least squares (PLS). Results are summarised in Table 1.

Table 1: Performance comparison of different modelling techniques. Comparison of mean Pearson correlation coefficient between predicted and observed compartment eigenvectors for three different modelling approaches: LM: linear regression; RF: Random Forest regression; PLS: partial least squares regression. Correlations were averaged per cell type over three cell types (cell type specific) and in the six possible crosses (cross-application) shown in Fig. 14.

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

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

Partial least squares regression is another technique which used dimensionality reduction to engineer a lower-dimension orthogonal feature set. Hence this method is well-suited to multi collinear inputs, such as our feature set. As expected, PLS regression provides highly accurate cell type specific predictions (mean PCC = 0.750) and during cross-application (mean PCC = 0.641), though in both cases produces slightly inferior results to RF models (Fig. 14).

PLS uses a type of dimensionality reduction, which offers another way to explore the inter-relationships between our feature set. Plotting input features against these lower-dimension components can give a revelling insight beyond simple correlations (e.g. Fig. 12). Figure 15 shows a "circle of correlations", where features are plotted

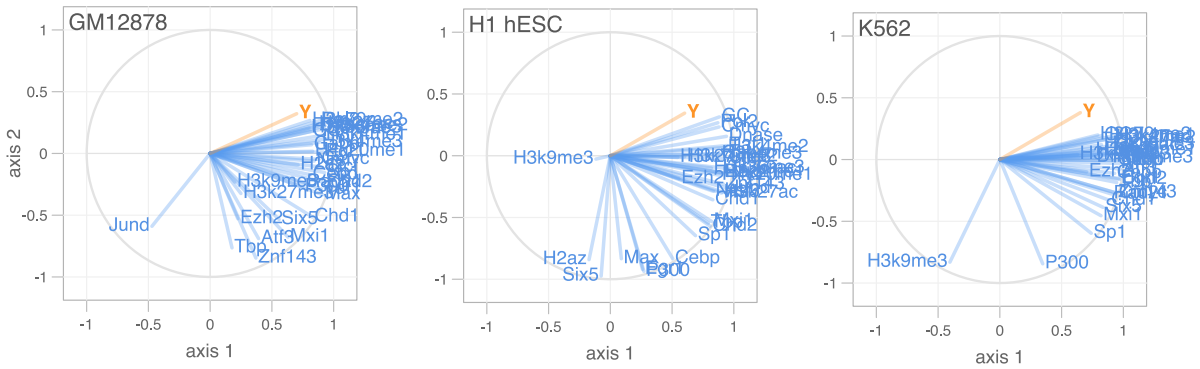

Figure 15: Circle of correlations of variables compared with PLS axes. Model variables are plotted against the first two axes used in PLS regression models per cell type. Y represents our compartment eigenvector.

onto polar co-ordinates against the first two PLS components. Interpretation of this figure is that nearby variables in the scatterplot are positively correlated, and the vector length from the circle centre is proportional to said variable's representation in the model. Negatively correlated variables point in opposite directions while uncorrelated variables are orthogonal to each other.[28] We therefore see the known multicollinearity represented as groupings of overlapping variables in each cell type, with a smaller number of orthogonal and negatively correlated variables in each cell type (Fig. 15).

1.5.3 Non-independence

As recognised through our use of Hidden Markov Models (Methods XX), consecutive bins along a chromosome are non-independent yet thus far predictive models have not considered this inter-dependence.

This is for two reasons: firstly non independence could be thought of as an artefact of bin-sizing (we have elected to use regular, fixed binning beneath the scale of compartments themselves whereas another approach could use variable bin sizes, for example per compartment, TAD or restriction fragment); secondly using information of a bin's surroundings may obscure by proxy the chromatin features which would otherwise prove predictive. As an example, knowing that bin x_{i-1} and bin x_{i+1} are in compartment state A would allow us with high confidence to predict the state of bin x_i, but without learning anything of any region's relationships with their histone modifications and bound factors.

1.6 PARSIMONIOUS MODELS FROM EXPANDED FEATURE SETS

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

Conversely, the 35 variables used thus far as model inputs are not the complete set available in each cell type, but only the subset of those assayed in all three cell types under study. The ENCODE consortium has produced a significantly greater number of datasets[16,29] in each cell type which have thus far gone unused. Here we'll explore models of higher order chromatin structure, in some cases built from over 100 variables, and then generate parsimonious models using optimal subsets guided by statistical techniques that penalise model complexity.

1.6.1 Stepwise regression

Multiple linear regression is a simple and analytically well-described modelling framework which is amenable to regularisation through a variety of methods. A simple approach is to start with a complete model and serially remove and/or add variables, then calculate a metric (here we use the Bayesian information criterion, BIC) which weighs the the model likelihood against model complexity. This process is iterated until the metric reaches a (local) minimum, thus creating a more parsimonious model which retains predictive accuracy and should be less prone to overfitting. Stepwise regression also aids interpretation by selecting representative features from collinear clusters.[30] A detailed explanation of this feature selection procedure can be found in Methods XX. It should be noted that despite its continued widespread usage, several statistical issues have been identified with the stepwise procedure for model selection.[31]

In terms of model performance alone, stepwise regression gives the highest predictive accuracy on a held-out validation set in each cell type specific model of compartment eigenvector (Table 2), however it must be said that differences in model performance across all comparisons are modest. These results do show that even expanded feature sets of up to 187 input features add little explanatory power beyond that of much less complex models with 20 or fewer input variables (Table 2).

Table 2: Performance comparison of full and optimised RF and ML models. PCC between predicted and empirical compartment eigenvectors is shown for a range of modelling scenarios, including multiple linear regression (LM) and Random Forest (RF) approaches. For model selection, two methods are used: stepwise BIC-regularised linear models and LASSO regression; in each case those same features were then also used in building a separate RF for comparison.

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

1.6.2 LASSO (ℓ_1) regression

A more modern technique for regularisation of linear models is the least absolute shrinkage and selection operator (LASSO). In brief, the LASSO is a form of ℓ_1 regularisation that penalises the sum of absolute values of standardised regression coefficients. By penalising absolute values and sums, rather than squared values as in ℓ_2 regularisation, coefficients can be shrunk to 0 thereby removing terms from the model. Thus LASSO combines coefficient shrinkage of techniques like Ridge regression with a type of feature selection as seen in stepwise regression.[32,33] A detailed explanation of this method can be found in Methods XX.

Again we can perform a simplistic comparison of model performance using LASSO regression and other techniques (Table 2). LASSO retrieves comparable numbers of informative variables to the stepwise regression technique in each cell type, and again removes the majority of input features from expanded sets as redundant or relatively uninformative.

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

1.6.3 Regularised Random Forest

Random Forest (RF) comparisons are included for comparison in Table 2 where RF models were built using model-selection procedures based on linear regression. The result of this is the linear regression-based feature selection acts as a "filter" method for feature selection, fully independent of the RF learning algorithm. A more coherent approach might be an "embedded" method, where a regularisation procedure is integrated with the learning algorithm.[34,35]

While RF is a much younger technique than linear models, a framework for Regularised Random Forests has recently been described[36] and implemented in the R package `RRF`.[?] The algorithm uses the idea that at each node in a tree, unused variable should only be included if they offer a significant information gain over those

available variables which have already been used in the tree. This differs from the standard RF algorithm where splitting decisions at each node are entirely independent of each other (Methods XX).

We found that this algorithm was unable to perform feature selection on our highly collinear feature set, instead leaving full or almost full feature sets in each case (*data not shown*) and so providing equal results to a standard RF model using expanded feature sets (Table 2).

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Consortium TF, Pmi R, Dgt C (2014) A promoter-level mammalian expression atlas. *Nature*, **507**(7493): 462–70.

[11] 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.

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

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

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

[15] 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.

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

[17] 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.

[18] 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.

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

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

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

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

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

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

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

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

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

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

[29] 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.

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

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

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

[33] 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.

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

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

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