Unravelling higher order genome organisation [working title]

Results 2 Predictive modelling

Benjamin L. Moore

July 7, 2015

igmm
INSTITUTE OF GENETICS & MOLECULAR MEDICINE

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

Recent studies have shown convincingly that local chromatin state measurements can accurately predict expression levels of genes on a genome-wide basis. Tippmann *et al.*,[?] designed a linear model to predict steady-state mRNA levels in mouse (*Mus musculus*) embryonic stem cells based on just four predictors: 3 histone modifications (H3K36me3, H3K4me2 and H3K27me3) and Pol-II occupancy. Remarkably, the linear model was found to explain 84.6% of an estimated 91% maximal variance that could be explained (as calculated through a detailed determination of noise). An additional finding of this study was that mRNA half-life and microRNA mediated transcript degradation both had relatively minor influence on steady-state mRNA levels, with the authors concluding that"the lion's share of regulatory contribution is at the level of mRNA synthesis and predictable from chromatin alone."[?] 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.[?]

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.[?] The authors here developed a two-stage model which first attempts to classify each

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

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

1.2 REPRODUCING DONG *et al.*

Following on from Dong *et al.*,[?] I first reimplemented the published ENCODE modelling framework to ensure I could replicate their results. In doing so I was also able to analyse the strengths and caveats of their approach; surprisingly the two-step classification then regression (firstly assessing a gene as 'on' or 'off' and then predicting its expression level) added little additional accuracy relative to a simple linear regression model (Fig. 1).

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

I attempted to improve the accuracy of predicted expression values produced by Dong *et al.* through two methods: increasing the number of informative regressors and increasing the complexity of the model by adding interaction terms and/or non-linear components. While Dong *et al.* included broad coverage of different histone modifications, they did not investigate the impact of higher-order chromatin data.

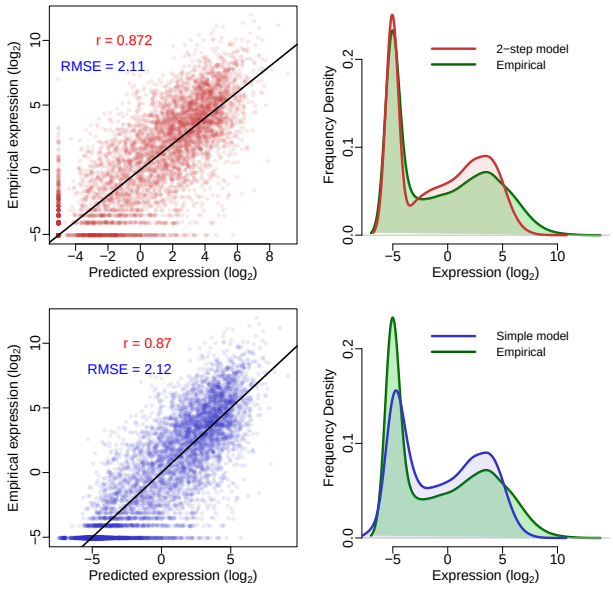

Figure 1: Comparison of classification-regression model (*upper*) with simple linear regression model (*lower*) recalculated following Dong *et al.*[?] Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for both methods are shown (*left*) along with frequency distributions of predicted and observed expression levels (*right*). Scatterplots are annotated with Pearson's correlation coefficient (r) and the root mean squared error (RMSE); the black trendlines describe $y = x$. Following 10-fold cross validation, overall correlation coefficients were: linear model $0.87 \pm 1.77 \times 10^{-5}$; Two-step model $0.872 \pm 9.89 \times 10^{-5}$. All correlations were statistically significant with $p < 1 \times 10^{-15}$ under the assumption of a t-distributed r with $d.f. = 7998$.

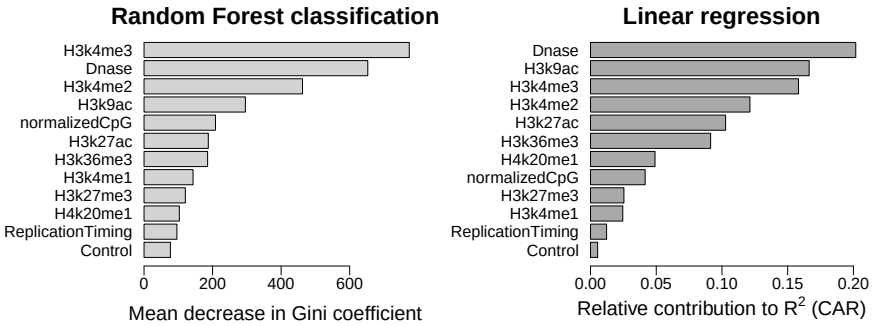

Figure 2: Relative importance metrics for variables in both the classification (*left*) and regression (*right*) stages of my reimplementation of Dong *et al.*'s two-step model.[?] The additional variable 'ReplicationTiming' shows the influence of $\log_2(early/late)$ replication timing ratio measured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

For this reason, I matched the TSS positions used in Dong *et al.* with previously-published genome-wide replication timing ratios measured in BG02 ESCs.[?] I then used these values as an additional regressor in both the two-step classification regression model and the simple linear model but saw no significant improvement in either model's accuracy. The reasons for this are likely that the data were relatively low-resolution (1 megabase blocks), from a imperfectly matched cell line and also that the Dong *et al.* model is already achieving such accurate results that they must already be accounting for most of the maximal explainable variance in gene expression given experimental and biological noise. With this in mind, additional regressors would be expected to yield diminishing returns. However, on closer examination, the replication timing data appeared only slightly more informative than the control ChIP-seq input measurements when evaluated with relative importance metrics (Fig. 2), implying that large-scale chromatin domains and long range interactions do not have significant influence on the expression of the genes resident within them. It would be of interest to investigate this further should more detailed higher order data become available. For example Hi-C interaction matrices have been calculated in the H1 cell line[?] and these could be compressed to principle component eigenvectors as has been done with other cell lines.[?]

1.3 MODELLING FANTOM5 CAGE TIMECOURSE DATA

Using unpublished FANTOM5 data and the approach established above, I next attempted to model gene expression at timepoint zero (t_0) of a 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[?] provided corresponding Entrez Gene IDs for gene-associated CAGE clusters, and I selected the most expressed cluster to represent the expression level of its mapped gene. I then compared these to Ensembl TSS annotations (v69) and discarded those tag clusters centered on a point > 50 bp from an annotated TSS associated with the mapped Entrez Gene ID, thereby removing enhancers and other non-genic transcribed regions.

Next I retrieved a number of genome-wide histone modification datasets from the ENCODE and NIH Roadmap consortia which were measured in H1 hESC cells, taking these to be reflections of the chromatin state t_0. I implemented the previously-described 'bestbin' strategy[?] to objectively select the most-correlated binned signal for each chromatinH1 hESC mark. Additionally, I analysed the stability of chosen bestbins by calculating them on 200 sets of 1000 randomly selected TSS samples (with each sample representing approximately 8% of the dataset) and the result is shown in Figure 3.

This result shows that bestbin selections are often consistent, indicating there are predictably informative regions relative to a TSS for each chromatin factor (Fig. 3). Furthermore, the selected bestbins match known biological mechanisms; for example the H3K36me3

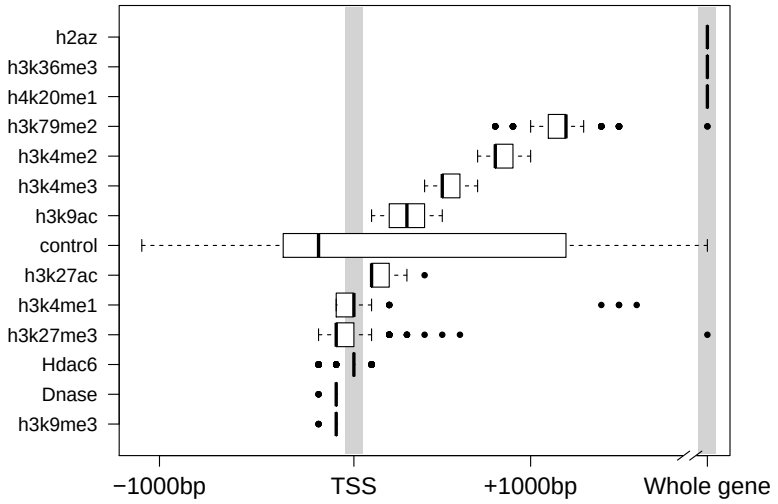

Figure 3: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, Dnase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 Kb flanking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

mark's bestbin is consistently the whole gene measurement and this mark is known to be enriched in actively transcribed exons.[? ? ?]

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

Figure 4 shows the resulting predictions of a preliminary RF model against the actual recorded expression over a test set of approximately 11000 TSS. This model was built with 15 predictors including control ChIP-seq input, though some of these could be removed without loss of accuracy. The model predictions evluated with 10-fold cross validation show a significant correlation with measured CAGE levels ($r = 0.845 \pm 1 \times 10^{-4}$; $t_{10868} = 164.4$, $p < 2 \times 10^{-15}$), and the model is able to explain around 71% of the variance in the expression response (for comparison a linear model resulted in $r = 0.825 \pm 3.2 \times 10^{-5}$; $t_{10868} = 152.2$, $p < 2 \times 10^{-15}$).

This result is worse than that of Dong *et al.* who achieved cross-validated correlation coefficients of up to 0.9, but it is roughly equal to their median test set correlation of 0.83.[?] The RMSEs, when normalised by the range of observed values, compare more favourably (0.11, compared with Dong *et al.*'s: 0.14). A possible explanation for this decrease in accuracy is that while both chromatin data and expression timecourse were measured in H1 hESC cells, the experiments

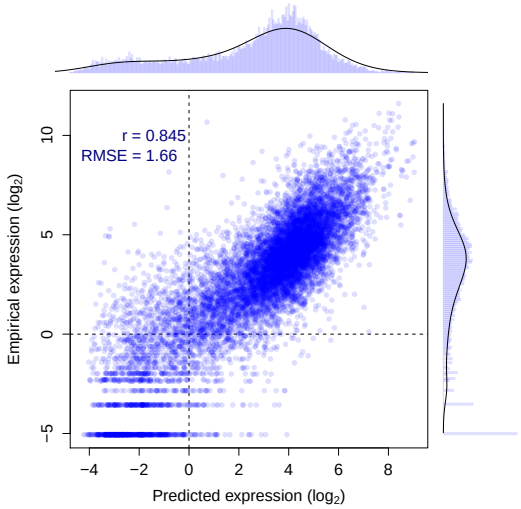

Figure 4: Evaluation of RF model predictions (x-axis) against an independent test set (y-axis). The distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson's correlation coefficient (r) and the root mean-squared error (RMSE) are also shown (*inset*).

took place at different institutes and likely using differing protocols and cell cultures. For comparison, a previous study using chromatin measurements from a number of different sources to predict expression in a matched cell-type reported a predictive correlation of 0.77.[?] Additionally, Dong *et al.* implemented a pseudocount optimisation step whereby an additional count added to each binned signal intensity prior to log transformation was optimised to maximise expression correlation. In the model presented above, a fixed psuedocount of 1 was used to avoid introducing positive bias towards higher correlation. Another difference between the two approaches is our use of a single-step model; Dong *et al.* found a small increase in correlation using their classification-regression approach but with the model implemented herein (Fig. 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).

Having built a reasonable model of t_0 expression, the next stage of this preliminary analysis was to consider successive timepoints. In the available CD34+ differentiation dataset, this consisted of expression data recorded at three timepoints (days 0, 3 and 9—hereafter t_0, t_3 and t_9 respectively). However genome-wide expression was highly correlated between each of these timepoints (Pearson correlation coefficients: $t_0, t_3 = 0.911$; $t_0, t_9 = 0.913$; $t_3, t_9 = 0.977$), and this high correlation meant that the genome-wide model performed essentially equally well regardless of the expression timepoint it was trained or tested on. In future analyses, higher-resolution timecourses may offer more interesting variation or alternatively genes that remain invariant throughout the timecourse could be filtered out of the dataset.

1.3.1 Dissecting the *best bin* approach

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 data assembled in this work (Chapter 1).

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

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.

We found cross-application between cell types was possible and with similarly-high levels of accuracy (Fig. 5). This gives good evidence not only that are models are not overfitting to cell-type specific noise, but also that there exist broad rules linking chromatin conformation and locus-level feature aggregation. The cross-application suggests there exists enough commonalities for compartment profile predictions to transcend the cell-type specific biology inherent to an embryonic stem cell or differentiated lymphoblast.

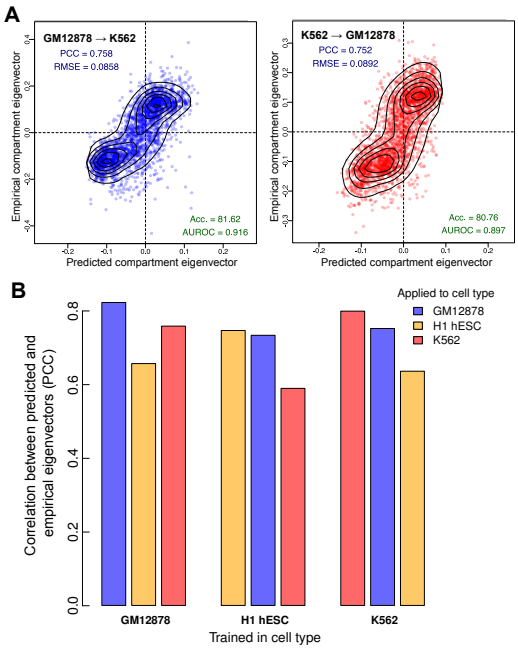

Figure 5: Models of higher order chromatin structure learned in one cell type can be cross-applied to two others Each model, trained in one cell type, was applied to the chromatin feature datasets from the other two cell types. (A) The GM12878 model achieved high accuracy when applied to K562 features (PCC = 0.76), as did the reciprocal cross (PCC = 0.75). (B) In each case, predictive accuracy decreased on cross-application but there remains significant agreement between predicted and empirical values. Acc., accuracy; AUROC, area under the receiver operating characteristic curve; PCC, Pearson correlation coefficient; RMSE, root mean-squared error.

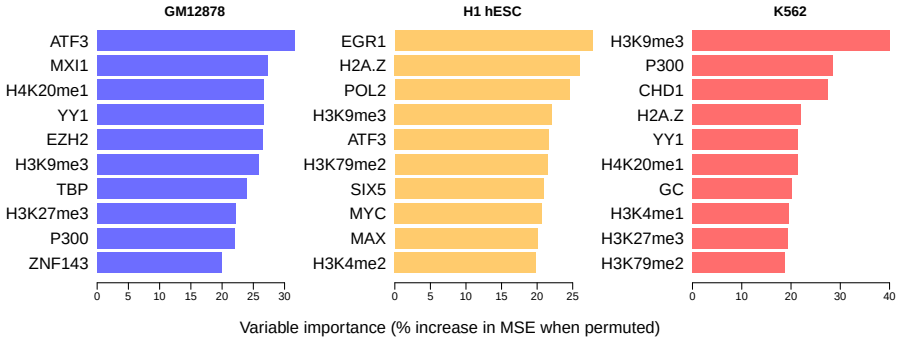

Figure 6: Variable importance per cell type specific model. Placeholder

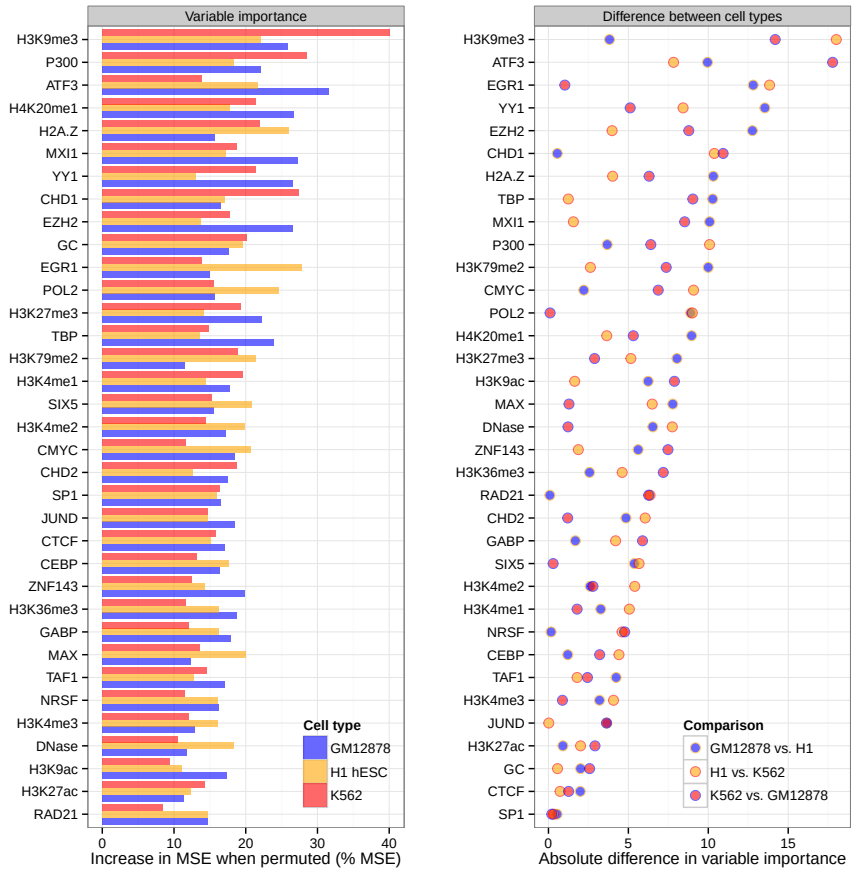

Figure 7: Variable importance per cell type specific model. Placeholder

1.4.3 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. **??**)

Only one input feature, H3k9me3, is present in the top 10 most important variables of each model. Under the assumption that variables are ranked independently, the probability of drawing the same variable in each case is low ($\frac{10}{36}^3 \approx 0.02$) however the probability of any one variable appearing in each ranking

1.4.4 Importance of resolution

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

Model accuracy when applied to higher resolution input features proved to be similarly high, with empirical PCC being 88 to 95% as high as that at 1 Mb native resolution (Fig. 8).

Note however, there is some indirect leakage between test and training set when 100 kb bins have been used in aggregate in learning

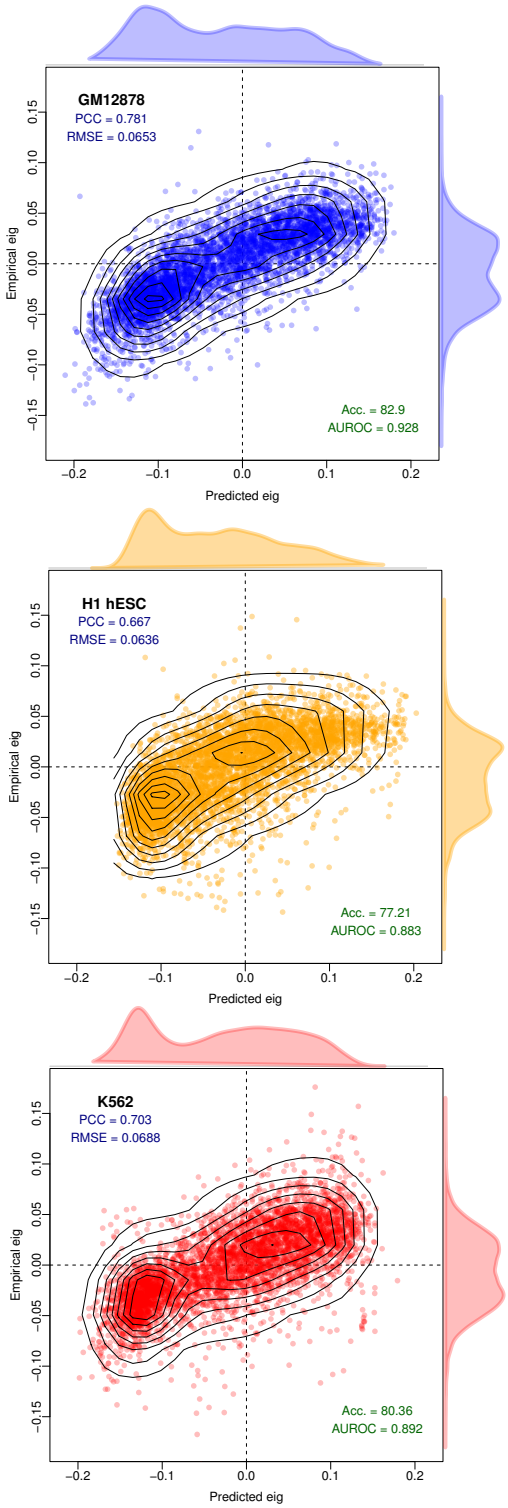

Figure 8: Models learned at 1 Mb resolution can be applied to higher resolution datasets. Despite having been trained on low resolution training sets, the Random Forest models generated can successfully predict compartment eigenvectors at higher resolution (100 kb, a 10× zoom). Eigenvectors at a higher resolution than this do not necessarily reflect A/B compartmentalisation.

Figure 9: **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.

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.

1.4.5 Other modelling approaches

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

Our results confirm RF as a suitable and powerful approach for modelling our relationships of interest in this work (Fig. 9), with both the highest cell-type specific performance (PCC between predicted and observed = 0.790) and on cross-applications (mean PCC = 0.689).

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

Partial least squares regression is another technique which used dimensionality reduction to engineer a lower-dimension orthogonal feature set. Hence this method is well-suited to multi collinear inputs, such as our feature set. As expected, PLS regression provides highly accurate cell type specific predictions (mean PCC = $0.750) and during cross-application (mean PCC=0.641$), though in both cases produces slightly inferior results to RF models (Fig. 9).

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. 9.

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

1.4.6 Non-independence

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

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

1.4.7 Correlating input features