Integrating high-resolution chromatin signals to accurately model gene expression over time

**Benjamin L. Moore**

**MRC Human Genetics Unit, MRC IGMM etc etc**

Supervisors: Colin A Semple, Martin S Taylor, Stuart Aitken

10 week report

03/12/2012

# Introduction

Large-scale chromatin data has recently been produced by multiple consortia, notable the ENCODE1 and NIH Roadmap Epigenomics2 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,3,4 an integrated understanding of general mechanisms underlying the causes or effects of these marks lags behind. A 2011 opinion piece asked the question “Histone modification: cause or cog? ”5 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.5 Similarly, the higher-order organisation of chromatin, while known the be an important factor in transcription, remains poorly understood mechanistically.6 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.*,7 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 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 influences 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.”7 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.8

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.9 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.9 Ultimately, 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.

These recent publications highlight the importance and relevance of advancing our understanding of chromatin biology; this is a general aim I hope to achieve through improving and extending upon the above analyses. I have extended these approaches to the examination of CAGE timecourse expression data arising from the ongoing FANTOM5 consortium (add website ref) and have begun to explore the roles of higher order structure in linear models of gene expression.

# Preliminary data

Following on from Dong *et al.*,9 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’, then predicting expression level) added little additional accuracy relative to a simple linear regression model predicting expression levels using the same set of regressors (Fig. Error: Reference source not found).

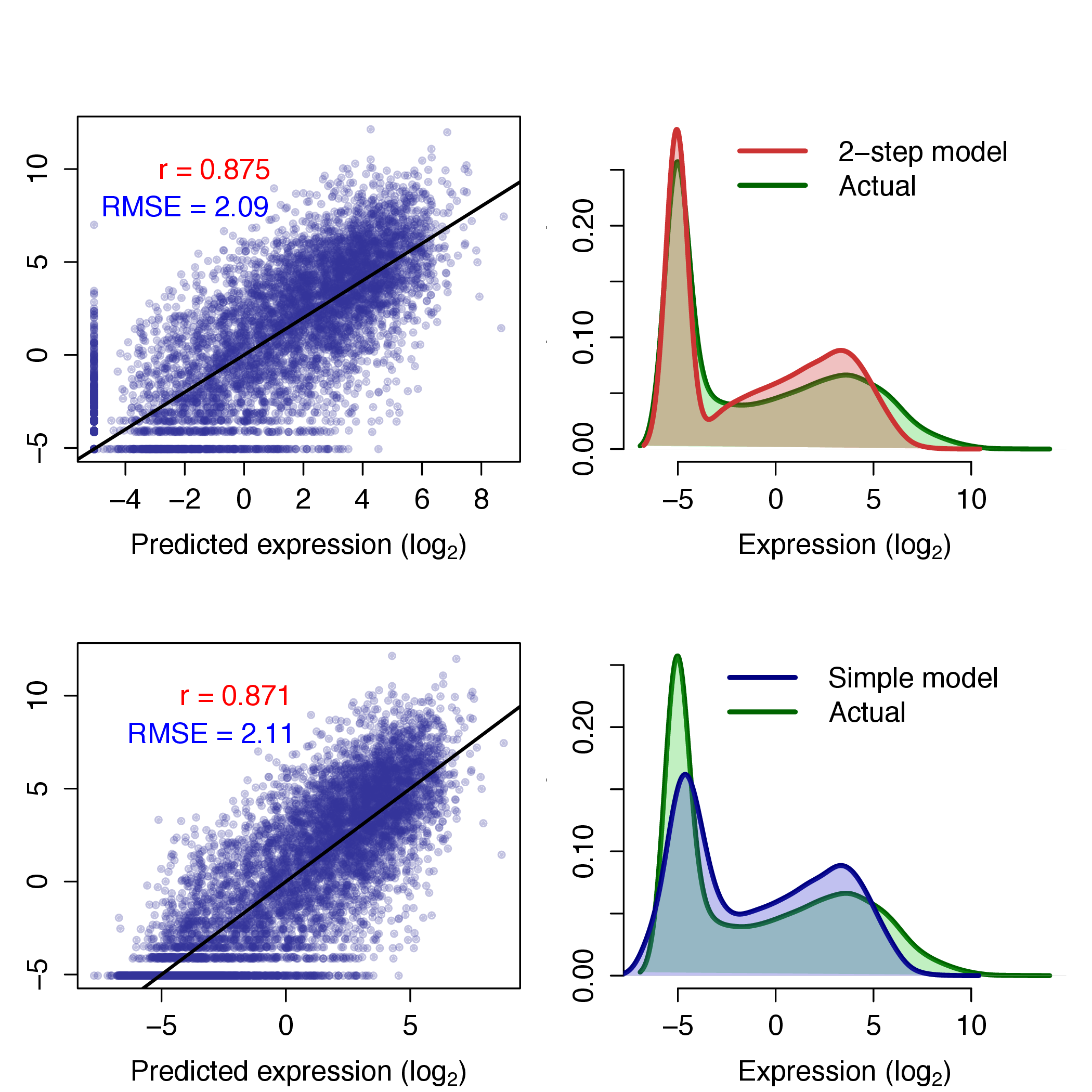


Figure 1: Comparison of classification-regression model (*upper*) with simple linear regression model (*lower*) recalculated following Dong *et al.*9 Scatterplots of predicted against empirical (‘Actual’) expression values for both methods are shown (*left*) along with frequency distributions of predicted and observed expression levels (*right*). Scatterplots are annotated with Pearson’s correlation coefficient (*r*) and the root mean squared error (RMSE); the black line describes *y*=*x*.

A 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 across an interval of X bp 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 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.9

I attempted to improve the accuracy of predicted expression values produced by Dong *et al.* through two methods: increasing the number of informative regressors and increasing the complexity of the model by adding interaction terms and/or non-linear components. While Dong *et al.* included broad coverage of different histone modifications, they did not investigate the impact of higher-order chromatin data. For this reason, I matched the TSS positions used in Dong *et al.* with previously-published genome-wide interaction probabilities measured using the Hi-C experimental technique.10 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 interaction data were relatively low-resolution (1 megabase blocks) 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 Hi-C data appeared only slightly more informative than the control input measurements when evaluated with relative importance metrics (*data not shown*), implying that large-scale chromatin domains and long range interactions do not have significant influence on the expression of the genes resident within them. This appears to be a controversial conclusion given the literature suggesting the opposite (cite recent ref eg PMID:22955621). It would be of interest to investigate this further should more detailed Hi-C data become available.

## Modelling FANTOM5 CAGE timecourse data

Using unpublished FANTOM5 data and the approach established above, I next attempted to model gene expression at timepoint zero () of a differentiation timecourse of Human H1 embryonic stem cells (H1hESC) to CD34+ hematopoietic stem cells.

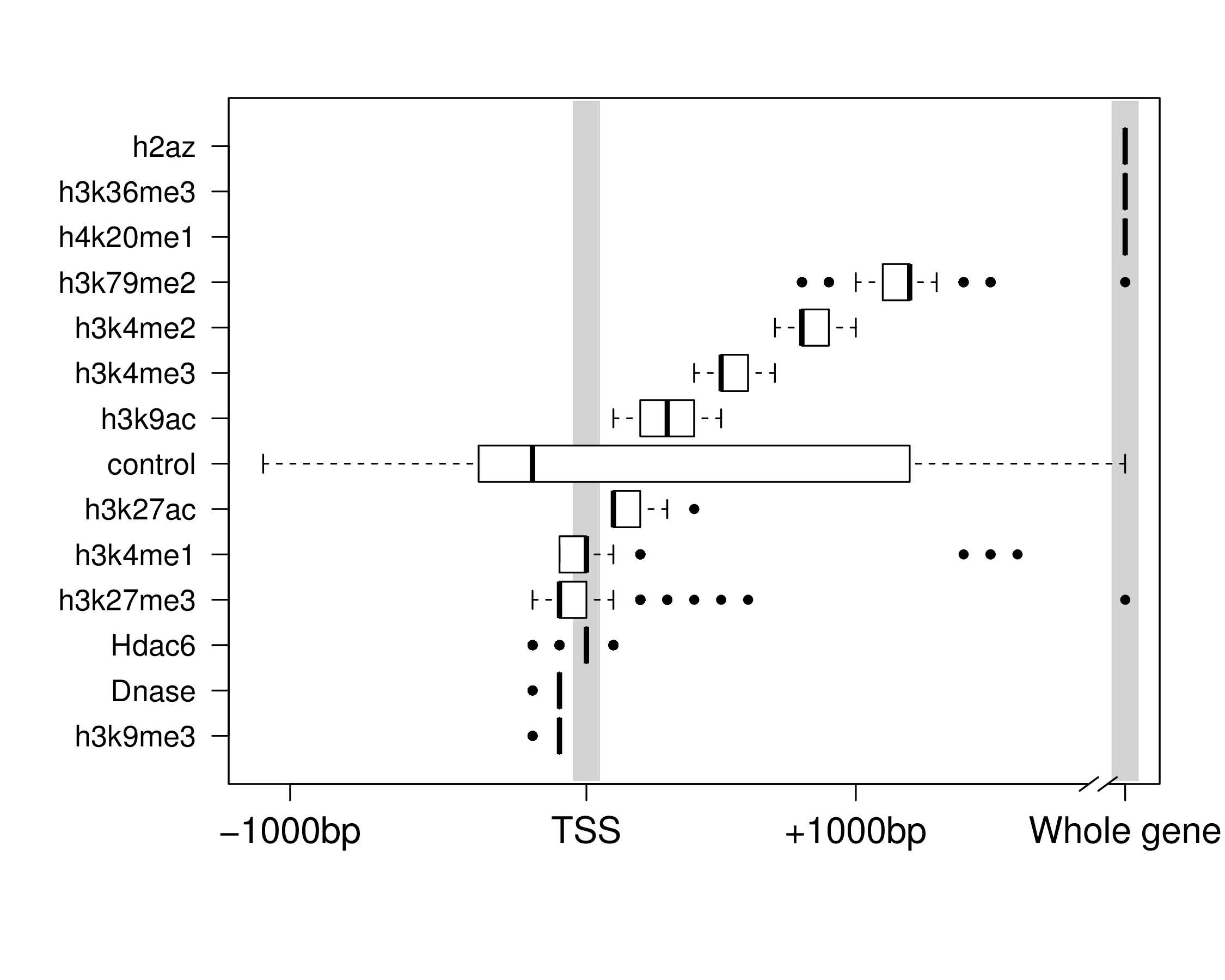
Having defined a representative TSS for each robustly mapped CAGE cluster, I retrieved a number of genome-wide histone modification datasets from the ENCODE and NIH Roadmap consortia which were taken from H1hESC cells, taking these to be reflections of the chromatin state . I implemented a modified version of the previously-described ‘bestbin’ strategy9 to objectively select the most-correlated binned signal for each chromatin mark. Specifically, I analysed the stability of chosen bestbins by calculating them on 200 subsets of 1000 randomly selected TSS samples (with each sample representing approximately 8% of the dataset) and the result is shown in Figure Error: Reference source not found.

Figure 2: 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 NCBI RefSeq.

This result shows that bestbin selections are often consistent, indicating there are predictably informative regions relative to a TSS for each chromatin factor (Figure Error: Reference source not found). Furthermore, the selected bestbins match known biological mechanisms; for example the H3K36me3 mark’s bestbin is consistently the whole gene measurement and this mark is known to be enriched in actively transcribed exons.7,11,12

Having matched a variety of genome-wide H1hESC chromatin datasets to the FANTOM5 timecourse expression data, I then built a regression model using an RF approach.13 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.14 RF works by building an ensemble of decision trees which each perform a succession of binary partitions aimed at maximising the separation between elements of the child nodes.15 Crucially, each branchpoint uses a random subset (size *mtry*) of the available predictors and each tree is fitted to a bootstrapped sample of the training set, allowing the model to be fitted without variable selection or risk of overfitting.14,16 Having grown a large number (*ntrees*) of these trees, continuous dependant variables (gene expression in this case) can be predicted by averaging the leaf-node classifications over all trees. A criticism of RF is that the model itself is less interpretable than a simple linear model; while this may be the case, it also seems that an additive linear model (e.g. 7,9) is unlikely to reflect the complex reality of transcriptional machinery.

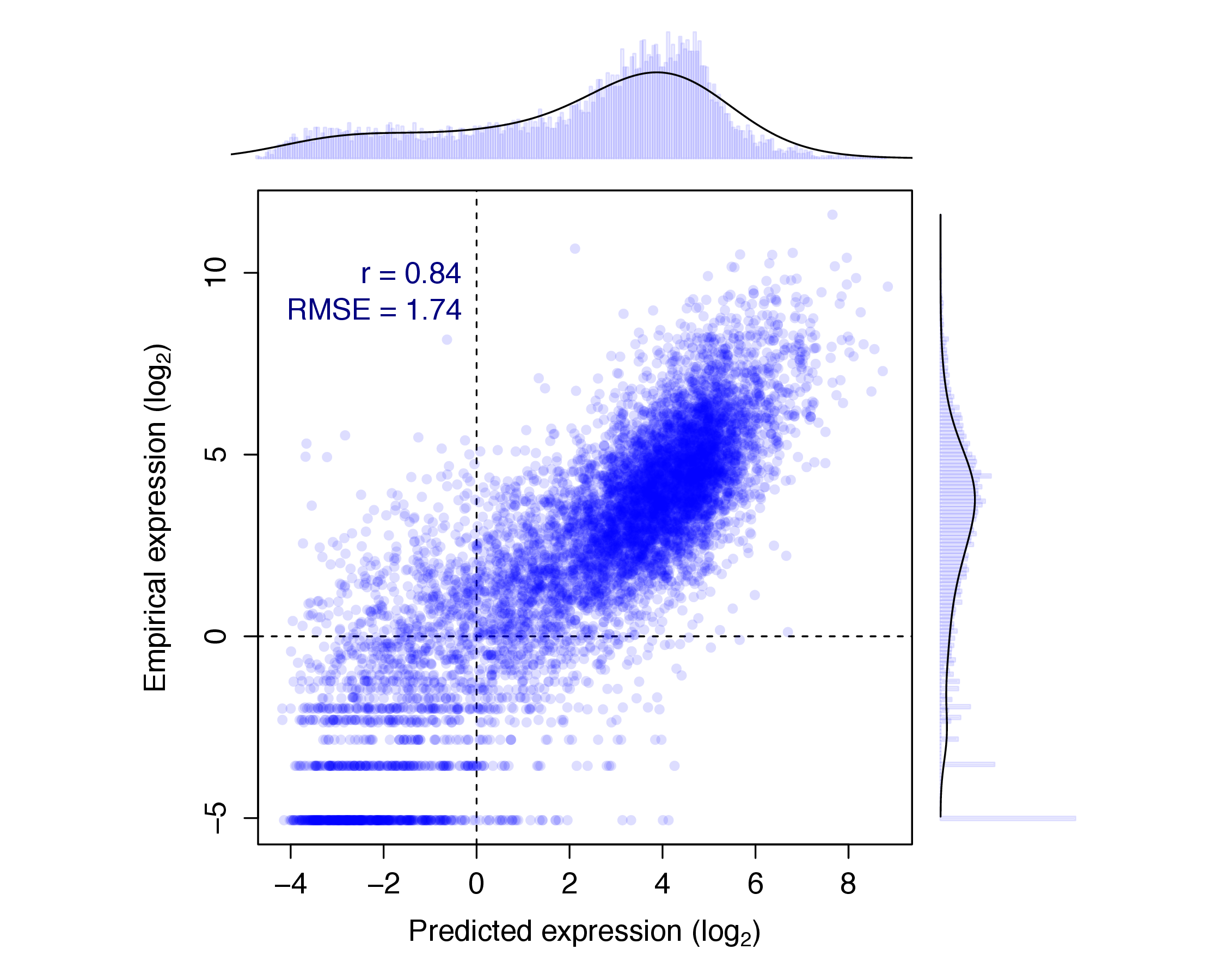


Figure 3: Evaluation of RF model predictions (x-axis) against an independent test set (*y*-axis). The distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson’s correlation coefficient (*r*) and the root mean-squared error (RMSE) are also shown (*inset*).

Figure Error: Reference source not found shows the resulting predictions of a preliminary RF model against the actual recorded expression over a test set of approximately 8400 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 show a significant correlation with measured CAGE levels (*r*=0.838±0.006; , ), and the model is able to explain around 70% of the variance in the expression response.

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 0.83.9 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 H1hESC cells, the experiments took place at different institutes and likely using differing culture protocols. 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 0.77.17 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 pseudocount 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 (Figure Error: Reference source not found) this approach gave no obvious advantage (for example, *r*=0.834±0.007, *RMSE*=1.77 when applied to the same test and training data used in Fig. Error: Reference source not found).

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

# Future work

The FANTOM5 consortium has produced numerous high-resolution timecourse CAGE datasets and many of these cell lines have matched ENCODE or NIH Roadmap chromatin data. Of particular interest are the cancer cell lines K562 (ENCODE Tier 1) and MCF-7 (ENCODE Tier 2.5) which both have detailed timecourse expression data. By combining the timecourse and chromatin data, I can compare the predictive power of my chromatin variables with expression levels at subsequent timepoints. This could lend insight to the questions: when do histone modification datas lose their predictive power? And how much variability of chromatin state can we infer is occurring over the timecourse? One might assume that the predictive potential of these chromatin data will generally degrade as the timecourse proceeds. On the other hand, if chromatin modifications are directing changes in transcriptional states, we might even expect that measurements would have an increased predictive power of gene expression at subsequent timepoints.

## Specific aims (first year of PhD studies)

1. Evaluate the predictive power of chromatin state across expression levels in a fine-grained timecourse analysis.

(a) When is predictive power lost or significantly decreased?

(b) Is there evidence for key chromatin transitions during the timecourse?

(c) Could a model accurately predict a succession of timecourse points?

2. Analyse how higher order chromatin data, such as contact probabilities, relate to locus-level chromatin measurements.

(a) Can higher order structure be predicted from local histone modifications?

3. Examine how different promoter classes respond to local chromatin state.

(a) Is there evidence of different transcriptional regulation of different (e.g. coding and non-coding, CpG island and non-CpG island, etc) TSS classes and is this predicted by their respective chromatin environments?

(b) How do enhancers, promoters, insulators and other genomic regions differ in their chromatin status?

Longer term work themes (2nd PhD year and beyond)

1. Investigate the application of non-linear models that have the potential to provide novel biological understanding of how individual chromatin marks can influence gene expression. Additionally, testing explicit interaction terms may highlight cooperativity or redundancy of chromatin features.

2. Explore methods of moving beyond correlative relationships to address the ‘cause or cog’ functional importance of histone modifications.

3. Explore conservation of models across species for orthologous TSSs using matched mouse and human ENCODE and FANTOM5 datasets. Are the behaviours of TSSs conserved where their genomic sequences are different?

Bibliography

[1] Gerstein, M. B, Kundaje, A, Hariharan, M, Landt, S. G, Yan, K.-k, Cheng, C, Mu, X. J, Khurana, E, Rozowsky, J, Alexander, R, Min, R, Alves, P, Abyzov, A, Addleman, N, Bhardwaj, N, Boyle, A. P, Cayting, P, Charos, A, Chen, D. Z, Cheng, Y, Clarke, D, Eastman, C, Euskirchen, G, Frietze, S, Fu, Y, Gertz, J, Grubert, F, Harmanci, A, Jain, P, Kasowski, M, Lacroute, P, Leng, J, Lian, J, Monahan, H, O’Geen, H, Ouyang, Z, Partridge, E. C, Patacsil, D, Pauli, F, Raha, D, Ramirez, L, Reddy, T. E, Reed, B, Shi, M, Slifer, T, Wang, J, Wu, L, Yang, X, Yip, K. Y, Zilberman-Schapira, G, Batzoglou, S, Sidow, A, Farnham, P. J, Myers, R. M, Weissman, S. M, & Snyder, M. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature **489**, 91–100.

[2] Bernstein, B. E, Stamatoyannopoulos, J. a, Costello, J. F, Ren, B, Milosavljevic, A, Meissner, A, Kellis, M, Marra, M. a, Beaudet, A. L, Ecker, J. R, Farnham, P. J, Hirst, M, Lander, E. S, Mikkelsen, T. S, & Thomson, J. a. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology **28**, 1045–8.

[3] Sajan, S. A & Hawkins, R. D. (2012) Methods for identifying higher-order chromatin structure. Annual review of genomics and human genetics **13**, 59–82.

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

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

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

[7] Tippmann, S. C, Ivanek, R, Gaidatzis, D, Schöler, A, Hoerner, L, van Nimwegen, E, Stadler, P. F, Stadler, M. B, & Schübeler, D. (2012) Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. Molecular systems biology **8**, 593.

[8] McLeay, R. C, Lesluyes, T, Cuellar Partida, G, & Bailey, T. L. (2012) Genome-wide in silico prediction of gene expression. Bioinformatics (Oxford, England) **28**, 2789–96.

[9] Dong, X, Greven, M, & Kundaje, A. (2012) Modeling gene expression using chromatin features in various cellular contexts. Genome … pp. 1–17.

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

[11] Kolasinska-Zwierz, P, Down, T, Latorre, I, Liu, T, Liu, X. S, & Ahringer, J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nature genetics **41**, 376–81.

[12] Schaft, D. (2003) The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation. Nucleic Acids Research **31**, 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] Svetnik, V, Liaw, A, & Tong, C. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical … pp. 1947–1958.

[16] Dasgupta, A, Sun, Y. V, König, I. R, Bailey-Wilson, J. E, & Malley, J. D. (2011) Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genetic epidemiology **35 Suppl 1**, S5–11.

[17] Karlić, R, Chung, H.-r, 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**, 2926–31.