Unravelling higher order genome organisation [working title]

Introduction

Benjamin L. Moore

June 10, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

MRC Medical Research Council

THE UNIVERSITY OF EDINBURGH

CANCER RESEARCH UK

1 PREDICTIVE MODELLING OF TRANSCRIPTION

1.1 INTRODUCTION

Large-scale chromatin data has recently been produced by multiple consortia, most notably the ENCODE[1] and NIH Roadmap Epigenomics[2] 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,[3–5] 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?"[6] 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.[6] 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.[7] 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.*,[8] 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."[8] 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.[9]

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.[10] 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.[10] 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.[11,12]

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[13] 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 PRELIMINARY DATA

Following on from Dong *et al.*,[10] 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).

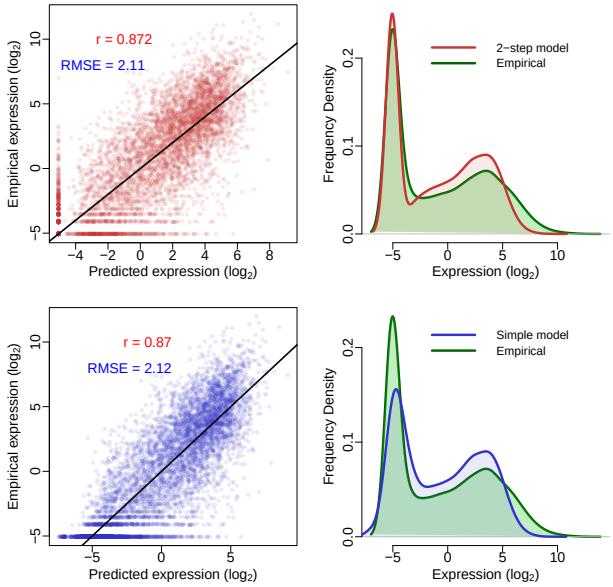

Figure 1: Comparison of classification-regression model (*upper*) with simple linear regression model (*lower*) recalculated following Dong *et al.*[10] Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for both methods are shown (*left*) along with frequency 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 $d.f. = 7998$.

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.[10]

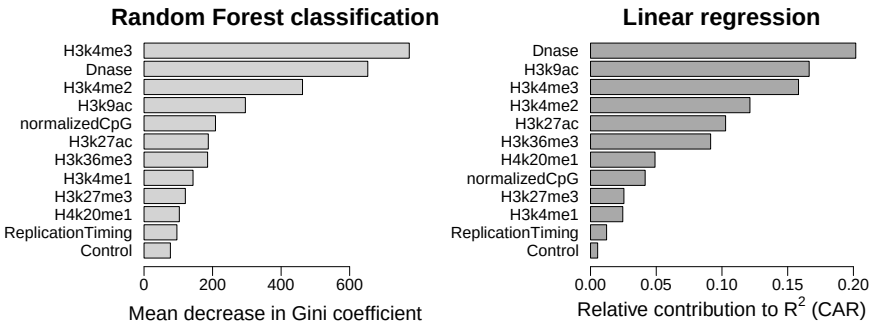

Figure 2: Relative importance metrics for variables in both the classification (*left*) and regression (*right*) stages of my reimplementation of Dong *et al.*'s two-step model.[10] The additional variable 'ReplicationTiming' shows the influence of $\log_2(early/late)$ replication timing ratio measured in the BG02 ESC cell type;[14] 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).[15]

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 replication timing ratios measured in BG02 ESCs.[14] 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[16] 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 (H1hESC) 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[13] 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 H1hESC cells, taking these to be reflections of the chromatin state t_0. I implemented the previously-described 'bestbin' strategy[10] to objectively select the most-correlated binned signal for each chromatin 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.

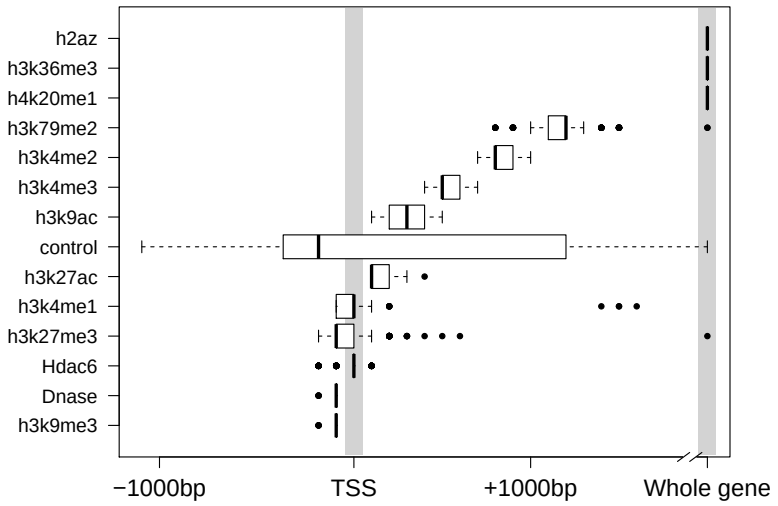

Figure 3: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, Dnase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 Kb flanking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

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 mark's bestbin is consistently the

whole gene measurement and this mark is known to be enriched in actively transcribed exons.[8,17,18]

Having matched a variety of genome-wide H1hESC chromatin datasets to the FANTOM5 timecourse expression data, I then built a regression model using a Random Forest (RF) approach.[19] 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.[20] 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.[21] 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.[20,22] 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, so for this reason linear models as well as other regression techniques will continue to be considered in future applications.

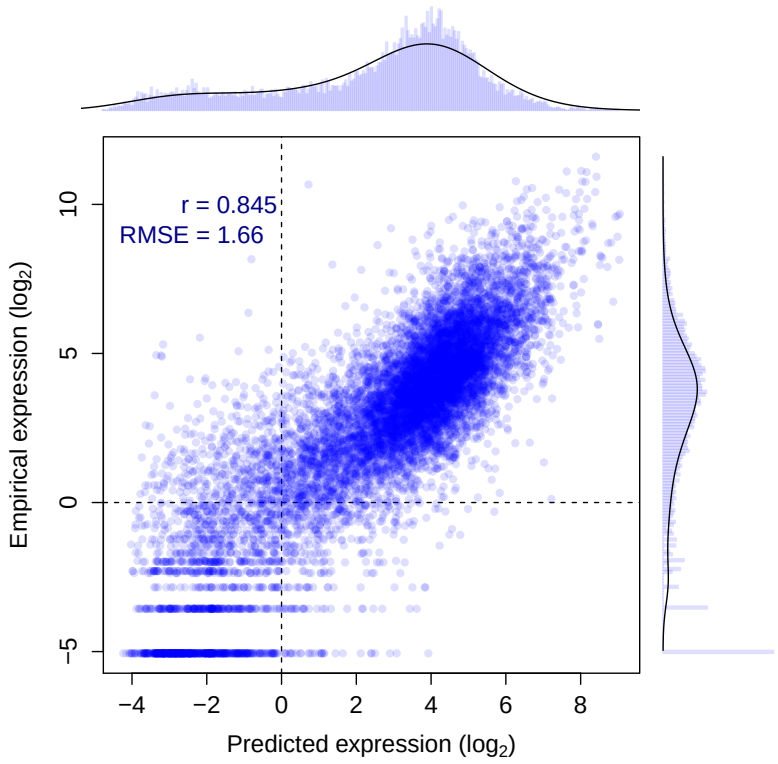

Figure 4: Evaluation of RF model predictions (x-axis) against an independent test set (y-axis). The distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson's correlation coefficient (r) and the root mean-squared error (RMSE) are also shown (*inset*).

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 vari-

ance 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.[10] 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 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.[23] 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.

REFERENCES

[1] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan Kk, Cheng C, Mu XJ, Khurana E, Rozowsky J, *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**(7414): 91–100.

[2] 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.

[3] Nikolov M, Fischle W (2012) Systematic analysis of histone modification readout. *Molecular bioSystems*, **Advance Ac**.

[4] Sajan SA, Hawkins RD (2012) Methods for identifying higher-order chromatin structure. *Annual review of genomics and human genetics*, **13**: 59–82.

[5] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345): 43–9.

[6] Henikoff S, Shilatifard A (2011) Histone modification: cause or cog? *Trends in genetics : TIG*, **27**(10): 389–96.

[7] Li G, Reinberg D (2011) Chromatin higher-order structures and gene regulation. *Current opinion in genetics & development*, **21**(2): 175–86.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] RIKEN Omics Science Center (2012) FANTOM5. *http://fantom.gsc.riken.jp/*.

[14] 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.

[15] 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.

[16] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398): 376–80.

[17] 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.

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

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

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

[21] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, **43**(6): 1947–58.

[22] Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD (2011) Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic epidemiology*, **35 Suppl 1**(Suppl 1): S5–11.

[23] 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.