Unravelling higher order genome organisation [working title]

Benjamin L. Moore

June 11, 2015

INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

CONTENTS

1 INTRODUCTION

1.1 GENOME ORGANISATION

It's oft-stated that the DNA within each human cell would extend for two metres fully extended. Instead that same length of DNA packs into a cell nucleus with a diameter in the order of micrometers (μm). This is achieved through a complex organisation hierarchy, ranging from how chromosomes are positioned in territories down to how DNA is wrapped around nucleosomes.

1.1.1 C-methods and Hi-C

Classical studies of chromosome conformation relied on microscopy techniques to visualise nuclear architecture. These techniques led to the discovery of "chromosome territories", regions of the nucleus wherein distinct chromosomes were thought to occupy. Finer details of chromatin organisation, such as the proposed 30 nm fibre, were also introduced through microscopy-based techniques.

With the advent DNA sequencing technology, new experimental methods emerged. Chromosome conformation capture (3C), introduced by Dekker *et al.*[1] was the first sequencing-based method of measuring chromosome conformation. The method uses formaldehyde to cross-link nuclear proteins in place, trapping genomic regions that were physically co-located through bound proteins, then to apply a frequent restriction enzyme to shear the sample into fragments. Next, under dilute conditions, DNA fragments are ligated together. The dilute conditions favour ligations between fixed fragments, with the aim of generating hybrid fragments form two genomic regions which were close together in the original preparation. Cross-linking can then be reversed and, in the case of the original 3C method, measured by quantitative PCR using pre-designed primers for your fragments of interest. The end result is a relative measure of interaction frequency between any two regions of interest, in theory directly proportional to their distance in three-dimensional space.

The rapid advancement of sequencing, allowed the original 3C method to be further developed, first through microarray technology, then using high-throughput sequencing. Two protocols were proposed for a 3C-inspired one-to-many assay[2,3] (both named 4C), whereby interactions were measured for a specific viewpoint fragment against all other restriction fragments genome-wide. The same year a many-to-many assay (5C) allowed measurements for all restriction fragments within a specified region.[4]

The final step was an all-versus-all assay, capable of assaying pairwise interaction frequencies between all restriction fragments of a genome. This assay was published by Lieberman Aiden *et al.*[5] and named Hi-C. The Hi-C method added biotin tagging to pull-down only ligated fragments for sequencing. At the time of publication, resolution of Hi-C data for analysis was limited by sequencing depth, given the huge number of restriction fragments produced by a 6-cutter enzyme (HindIII and NcoI were used in[5]) but the falling costs of sequencing and proven utility of the assay meant subsequent Hi-C papers incrementally increased their sequencing depth, to a point where analysis could be performed at the level of individual restriction fragments, genome-wide.[6–9]

1.1.2 Hi-C variants

The interaction maps produced by Hi-C were noticed to exhibit several inherent biases. Fragment properties, such as their length, GC content and mappability, were confounding interaction frequency estimates and therefore needed to be normalised-away before subsequent analysis.[10] A range of statistical techniques were developed to correct for these latent variables[11–14]

Tethered chromosome capture (TCC)[15] was the first attempt to increase the signal to noise ratio of Hi-C contacts. In this method, ligations take place on a fixed surface, with the aim of preventing spurious ligations between fragments in solution which were not cross-linked.

Hi-C is a population-level assay, as the retrieved interaction counts are from a huge number of different cells. As well as building population-averaged models of genome structure, it is also of interest to probe cell-to-cell variability through single-cell approaches. The first single-cell Hi-C study[16] ... An obvious limitation of single-cell Hi-C assays is that a single restriction fragment can ligate to at most a single other fragment, meaning even if 100% yield were to be achieved, any $n \times n$ interaction matrix could at most populate $\frac{n}{2}$ cells.

Capture-C is another recent Hi-C derivative which attempts to address resolution problems associated with the genome-wide pairwise assay by enriching for promoter-enhancer interactions using *a priori* selection.[17] It could be said that Capture-C is to Hi-C as exome-capture sequencing is to a whole-genome approach.

In-site Hi-C was a recent refinement of the Hi-C method, from the published of the original method.[9] The principle difference is that fixation and ligation now happen in place, within intact cell nuclei.

1.1.3 Chromosome compartments

In the paper describing the Hi-C technique,[5] Lieberman-Aiden *et al.* described low-resolution structures they name "A" and "B" nuclear compartments. These are regions with a median size of around 5 megabases which showed properties typical of euchromatin and heterochromatin, respectively. A compartments were observed through 3D-FISH to be centrally-positioned in the nucleus and ChIP-seq data showed several hallmarks of transcriptional activity. B compartments, conversely, were heterochromatic and lamina-associated regions, with little transcription and repressive histone modifications such as H3k9me3.[5] As expected from positioning data, the co-location of compartment types is also visible in their contact maps.

Figure: correlation matrix of contacts with eigenvector profile

These compartments were identified through a continuous eigenvector profile, derived from a normalised Hi-C contact matrix.[5] Importantly, this measure holds more information than a simple two-state classification, rather the continuous values can be interpreted as relative levels of compaction or activity.[11,12]

1.1.4 Topological domains

The falling cost of high-throughput sequencing enabled increasingly deep sequencing of Hi-C experiments. Sequencing is the main resolution-limiting resource for this assay, as to increase the analysis resolution and maintain the level of coverage requires an exponential increase in the total amount of sequencing required.[5]

In experiments totalling around two billion total sequencing reads, Dixon *et al.*[6] produced Hi-C contact maps in human and mouse cell lines at 40 kb resolution. The authors noticed smaller domains they designated "topologi-

cal associative domains" (or TADs) which were observable as self-interacting, off-diagonal blocks of higher-than-expected self-interaction frequency. They defined a domain calling algorithm based on the directional bias of a genomic region's contacts, and used a Hidden Markov Model to infer blocks of strongly up- or downstream-biased, reasoning that domain boundaries are present when a strongly upstream biased region is adjacent to a region of opposite bias. These boundaries themselves were investigated and were found to display suggestive functional enrichments for DNA binding proteins including CTCF, long thought to act as an insulator of chromatin state.

Dixon *et al.*[6] also performed some comparative analysis, reporting large and significant overlap of domain boundary positions both within species and between human and mouse cell lines.

1.1.5 Other proposed structures

Filippova *et al.*[18] developed a tuneable algorithm which identifies "alternative topological domains".

Recent high-resolution studies have been able to resolve ever-smaller levels of sub-structure. Rao *et al.*[9] refined the concept of chromosome compartments to "sub-compartments", dividing simple A/B divisions into a total of 5 subtypes. The authors were also able to identify "contact domains" of median size 185 kb, many of which were associated with identifiable individual looping events.[9] The authors also suggest that previously-observed large TADs may be the result of insufficient sequencing; that is, not all boundaries could be detected using 40 kb binned contact maps thus multiple contact domains were unintentionally combined into large domains.

1.2 MODELS OF CHROMATIN ORGANISATION

Theoretical mechanistic models of chromatin folding such as the "strings and binders switch" model[19] and the "fractal globule" model[20**?** ,21] have both produced simulated data that reflects empirical C-method observations and potentially describe the polymer dynamics of chromatin folding.

1.2.1 Fractal globule

1.2.2 Strings, binders and switches

1.3 CRITICISMS OF C-METHODS

Compare / contrast C-methods and FISH, Bickmore lab papers.

1.4 MACHINE LEARNING IN GENOMICS

The link between epigenomic features and local chromatin state has been analysed computationally in a number of publications, notably in developing the Hidden Markov Model-based ChromHMM[22] algorithm which predicts states such as active promoters and enhancers, using a range of histone marks and other underlying features.[23] Similarly a Random Forest-based algorithm was developed to predict enhancers from histone modification data.[24] However few studies have spanned all of these levels of chromatin structure and nuclear organisation, and it is not yet known how locus-level chromatin features may be related to higher order genome organisation.

1.4.1 ENCODE

The recent comprehensive ChIP-seq datasets produced by the ENCODE consortium [25] combined with Hi-C genome-wide contact maps in a number of human cell types [6,15?] present a remarkable opportunity to investigate the relationships between local chromatin features and higher order structure. In this work, a machine-learning approach was employed to model the compartmental characteristics of large genomic regions based on their aggregate levels of various histone marks and DNA binding proteins. Dissection of the resulting models was then used as a means of gleaning biological insights into the basis of higher order structure and of highlighting important differences between cell types.

1.5 AIMS

2 METHODS

2.1 INPUT DATA

Some more stuff here.

2.1.1 Hi-C data

Raw Hi-C reads were downloaded from three published datasets through GEO[26] or the SRA[27] with identifiers: GSE35156 (H1 hESC), GSE18199 (K562) and SRX030113 (GM12878). These paired reads were mapped independently to theand mapped to the genome (hg19/GRCh37). Mapping was performed using the `hiclib` software package[11] and `bowtie2`[28] with the `--very-sensitive` flag. Mapped reads were then binned into contact maps and iteratively corrected[11]. The `hiclib` software was also used for eigenvector expansion of each intrachromosomal contact map, performed independently on each chromosome arm.

Table 1: Public Hi-C data used in this work.

Cell line	Total reads	Accession	Citation
Gm12878	31×10^6	SRX030113	15
H1 hESC	331×10^6	GSE35156	6
K562	36×10^6	GSE18199	5
Cortex	373×10^6	GSE35156	6
mESC	476×10^6	GSE35156	6
IMR90	355×10^6	GSE35156	6

2.1.2 Locus-level features

Genome-wide ChIP-seq datasets for: 22 DNA binding proteins and 10 histone marks were made available by the ENCODE consortium[25,29] along with DNase I hypersensitivity and H2A.z occupancy, for each of the Tier 1 ENCODE cell lines used in this work: H1 hESC, K562 and GM12878. These data were pre-processed using MACSv2[30] to produce fold-change relative to input chromatin. GC content was also calculated and used in the featureset.

2.1.3 Clustering input features

To quantify collinearity of input features, correlation matrices built from genome-wide vectors of input feature measures were build and hierarchicaly clustered. The ”significance” of observed clustering was assessed using sub- and super-sampled bootstrapping, with stable clusters deemed significant. The `pvclust` R package

2.2 MODELLING

2.2.1 Random Forest

Random Forest (RF) regression,[31] was used as implemented in the R package `randomForest`.[32] The RF algorithm makes use of a collective of regres-

sion trees (size *ntrees*), each built from a bootstrapped sample of the training set. In growing each tree, a small number of variables (*mtry*) is tested at each bifurcation node, and that which minimises the variance in child node subsets is selected at a specific threshold. Having trained a group of trees, these can then be used as predictive tools by inputting a vector of features to each tree and averaging the output leaf node value across the forest. RF regression was used as it is known to be one of the most powerful regression methods developed to date,[33,34] typically providing low bias and low variance predictions without the need for variable selection.[35,36] Additionally the RF method represents an example of "algorithmic modelling"[37] in that it makes no assumptions about the underlying data model. Parameters of $mtry = \frac{n}{3}$ (where n is the number of input features) and $ntrees = 200$ were assumed as they are known to be largely insensitive;[36,38] this was verified with the dataset used in this work (Fig. **??**).

Variable importance within Random Forest regression models was measured using mean decrease in accuracy in the out-of-bag (OOB) sample. This represents the average difference (over the forest) between the accuracy of a tree with permuted and unpermuted versions of a given variable, in units of mean squared error (MSE).[34,36]

2.2.2 Model performance

The effectiveness of the modelling approach was measured by four different metrics. Prediction accuracy was assessed by the Pearson correlation coefficient between the predicted and observed eigenvectors (determined by 10-fold cross-validation), and the root mean-squared error (RMSE) of the same data. Classification error, when predictions where thresholded into $A \geq 0; B < 0$, was also calculated using accuracy (% correct classifications or True Positives) and area under the receiver operating characteristic (AUROC) curve. Together these give a comprehensive overview of the model performance, both in terms of regression accuracy of the continuous eigenvector, and in how that same model could be used to label discrete chromatin compartments.

For cross-application of cell type specific models, a single Random Forest regression model was learned from all 1 Mb bins for a given cell type. This was then used to predict all bins from each of the other two cell types.

2.2.3 Other modelling approaches

Linear regression was used as a baseline for comparison with more complicated approaches such as Random Forest. If the same modelling accuracy could be achieved with simple multiple linear regression, this would be a faster and more interpretable modelling framework.

Partial least squares regression was also used to model compartment profiles. This method is well-suited to highly correlated inputs.

2.2.4 Graphical lasso

Regularised models made use of the Graphical LASSO[39] (least absolute shrinkage and selection operator) as a method of L_1-norm based regularisation, implemented via the `glasso` R package. The graphical lasso provides tuneable regularisation which is capable of feature selection via minimising regression parameters to 0. It was chosen in this case due to the multicollinearity of the featureset, the algorithm's fast speed of execution and the intuitiveness a graphical model presents.[39]

More specifically, the graphical lasso regulates the number of os in the inverse covariance matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, also known as the precision matrix. Then if element $\theta_{ij} = 0$, the variables X_i and X_j can be said to be conditionally independent, given the remaining variables.[40] The algorithm minimises a negative log-likelihood (Eqn. 1 [40]) given the tuning parameter λ, which was tuned in this case to leave a small number of variables (< 10) directly dependent on the eigenvector data.

$$\underset{\mathbf{\Theta} \prec \mathbf{0}}{\text{minimise}}\ f(\mathbf{\Theta}) := -\log\det(\mathbf{\Theta}) + \operatorname{tr}(\mathbf{S\Theta}) + \lambda\|\mathbf{\Theta}\|_1 \tag{1}$$

2.3 VARIABLE REGIONS

2.3.1 Stratification by variability

Median absolute deviation (MAD) was chosen as a robust measure of the variability in a given 1 Mb block between the three primary cell types used in this work: H1, K562 and GM12878. Blocks were ranked by this measure and split into thirds that represented "low" variability (the third of blocks with the lowest MAD), "mid" and "high" variability. Each subgroup was then independently modelled using the previously-described Random Forest approach.

"Flipped" regions are those whose compartment state differs in one cell type relative to the other two. For example, if a 1 Mb bin was classified as "open" in H1 hESC and "closed" in both K562 and GM12878, this is said to be a "flipped" compartment (to open).

2.3.2 Enhancer enrichment

Enhancer annotations were collected from the ChromHMM / SegWay combined annotations in each cell type.[41] Enhancers were considered "shared" if there was an overlapping enhancer annotation in either of the two other cell types, and labelled as "tissue-specific" otherwise.

This was repeated for other chromatin states.

2.4 BOUNDARIES

2.4.1 TADs

TAD boundaries were called using the software provided in Dixon *et al.*[6] using their recommended parameters. For the generation of boundary profiles, the same parameters were used: input features were averaged into 40 kb bins spanning ±500 kb from the boundary centre.

To align boundaries between cells ...

2.4.2 Compartments

Compartment boundaries were called by first training a two-state hidden Markov model (HMM) on the compartment eigenvector and then using the Viterbi algorithm to predict the most likely state sequence that produced the observed values. The point at which transitions occurred between states was taken as a boundary which was then extended ±1.5 Mb to give a 3 Mb window in which a boundary was though to occur.

To test for the enrichment or depletion of a chromatin feature over a given boundary, a two tailed Mann-Whitney test was used to compare the boundary bin with the ten outermost bins of the window (5 from either side).

The significance level at $\alpha = 0.01$ was then Bonferonni-adjusted for multiple testing correction, and results with p-values exceeding this threshold were deemed significantly enriched or depleted at a given boundary.

2.4.3 MetaTADs

2.5 GIEMSA BAND COMPARISON

Cytogenic band data and Giemsa stain results were downloaded from the UCSC genome browser (table `cytoBandIdeo`). The genomic co-ordinates are an approximation of cytogenic band data inferred from a large number of FISH experiments.[42]

To compare G-band boundaries with our compartment data, we allowed for a $\pm$500 kb inaccuracy in G-band boundary. For each G-band boundary, the minimum absolute distance to any compartment or TAD boundary was calculated for each cell type. To generate a null model, ...

2.6 NUCLEAR POSITIONING

Previously published data on chromosome positioning preference within the nucleus was used to label each chromosome as "inner", "middle" or "outer".[43] Chromosomes whose DAPI hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by Boyle *et al.*[43], made up the "inner" group and included chromosomes 1 and 16. Similarly the "outer" group had enriched signals ($p \leq 5 \times 10^{-3}$) in the outer shell relative to the inner nuclear shell and included chromosomes 2, 3, 11-13 and 18. The remaining chromosomes in our filtered dataset, 6, 14 and 15, were assigned to the "middle" group and showed no significant to either inner or outer nuclear shells ($p \geq 0.1$).[43] The significance of the difference in distribution of eigenvectors in the inner versus outer shell was determined by a one-sided Kolmogorov-Smirnov (K-S) test, with the alternative hypothesis that the empirical cumulative density function of the inner chromosome eigenvectors F_{inner} is greater-than or equal-to F_{outer}. This chromosomal positioning data was measured in lymphoblastoid cells though nuclear architecture is though to be largely conserved between cell types[44,45] and even higher primates.[46]

3 PREDICTIVE MODELLING OF TRANSCRIPTION

3.1 PRELIMINARY DATA

4 INTEGRATIVE MODELLING AS A TOOL TO EXPLORE BIOLOGICAL SYSTEMS

4.1 INTRODUCTION

Large-scale chromatin data has recently been produced by multiple consortia, most notably the ENCODE[47] and NIH Roadmap Epigenomics[48] 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,[23,49,50] 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?"[51] 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.[51] 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.[52] 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.*,[53] 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."[53] 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.[54]

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.[55] 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.[55] Addition-

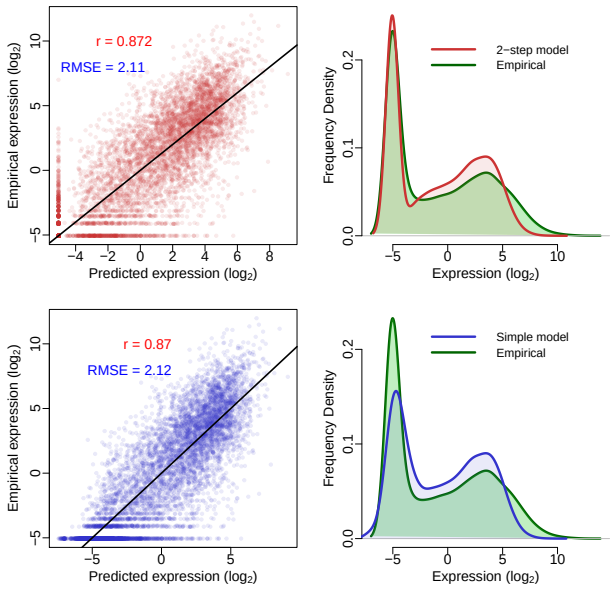

Figure 1: Comparison of classification-regression model (*upper*) with simple linear regression model (*lower*) recalculated following Dong *et al.*[55] Scatterplots of predicted against empirical $\log_2$ reads per million (RPM) expression values for both methods are shown (*left*) along with frequency 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 $d.f. = 7998$.

ally, 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.[56,57]

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[58] 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.

4.2 REPRODUCING DONG *et al.*

Following on from Dong *et al.*,[55] 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).

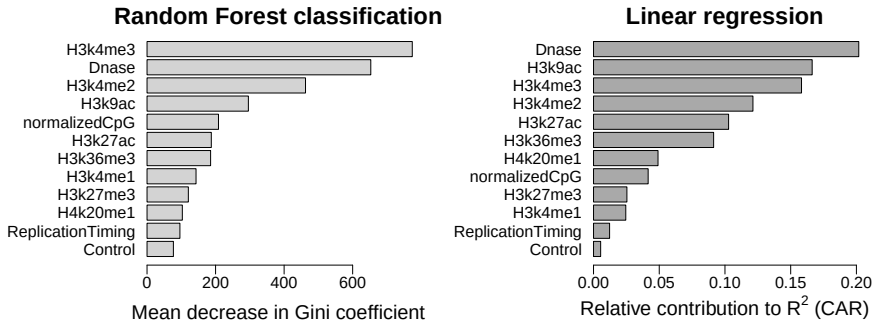

Figure 2: Relative importance metrics for variables in both the classification (*left*) and regression (*right*) stages of my reimplementation of Dong *et al.*'s two-step model.[55] The additional variable 'ReplicationTiming' shows the influence of $\log_2(early/late)$ replication timing ratio measured in the BG02 ESC cell type;[59] 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).[60]

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.[55]

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.[59] 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[6] and these could be compressed to principle component eigenvectors as has been done with other cell lines.[?]

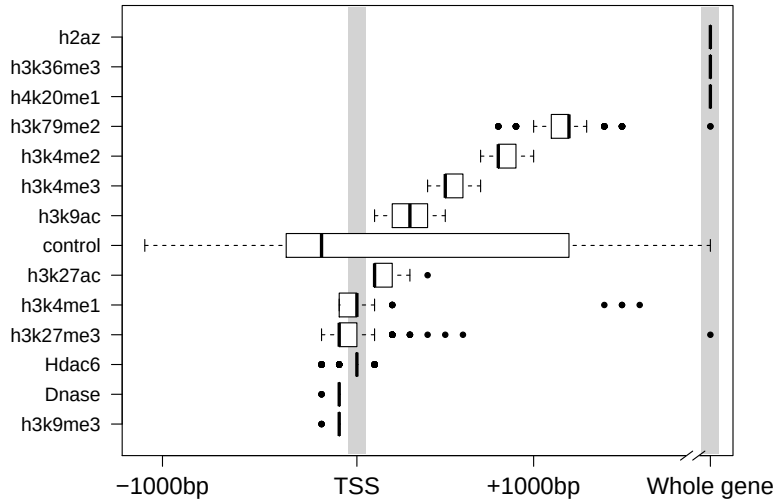

Figure 3: Distributions of bestbin locations relative to the TSS. Bestbins were selected for normalised ChIP-seq signal intensities for 10 histone marks, the H2A.Z histone variant, Hdac6 histone deacetylase, Dnase hypersensitivity and a ChIP-seq input chromatin control. Bins analysed extended 2 Kb flanking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

4.3 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 [58] 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 [55] 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 mark's bestbin is consistently the whole gene measurement and this mark is known to be enriched in actively

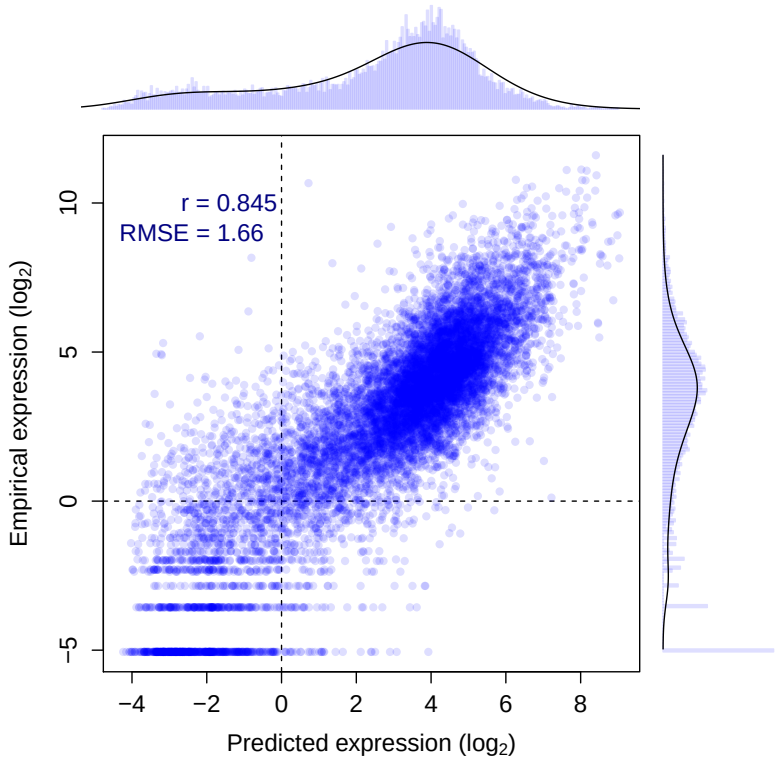

Figure 4: Evaluation of RF model predictions (x-axis) against an independent test set (y-axis). The distributions of predicted and empirical expression values are shown opposite their respective axes. Pearson's correlation coefficient (r) and the root mean-squared error (RMSE) are also shown (*inset*).

transcribed exons.[53,61,62]

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.[63] 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.[35]

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.[55] 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 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.[64]

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.

4.3.1 Dissecting the *best bin* approach

4.4 MODELLING HIGHER ORDER CHROMATIN

4.4.1 Cross-application

4.4.2 Variable importance

4.4.3 Correlating input features

5 PREDICTIVE MODELLING OF TRANSCRIPTION AND CHROMATIN ORGANISATION

Hello

6 PREDICTIVE MODELLING OF TRANSCRIPTION AND CHROMATIN ORGANISATION

Hello

7 PREDICTIVE MODELLING OF TRANSCRIPTION AND CHROMATIN ORGANISATION

Hello

REFERENCES

[1] Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science (New York, N.Y.)*, **295**(February): 1306–1311.

[2] Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, **38**(11): 1341–1347.

[3] Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, **38**(11): 1348–1354.

[4] Dostie J, Richmond Ta, Arnaout Ra, Selzer RR, Lee WL, Honan Ta, Rubio ED, Krumm A, Lamb J, *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**(10): 1299–1309.

[5] 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.

[6] 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.

[7] Selvaraj S, R Dixon J, Bansal V, Ren B (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology*, **31**(12): 1111–8.

[8] Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen Ca, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**(7475): 290–4.

[9] Rao S, Huntley M, Durand N, Stamenova E, Bochkov I, Robinson J, Sanborn A, Machol I, Omer A, *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**(7): 1665–1680.

[10] Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**(11): 1059–65.

[11] 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.

[12] Dekker J, Marti-Renom Ma, Mirny La (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, **14**(6): 390–403.

[13] Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics (Oxford, England)*, **28**(23): 3131–3.

[14] Li W, Gong K, Li Q, Alber F, Zhou XJ (2014) Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics (Oxford, England)*, (November): 1–3.

[15] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, **30**(1): 90–8.

[16] Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**(7469): 59–64.

[17] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, (April): 1–12.

[18] Filippova D, Patro R, Duggal G, Kingsford C (2014) Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, **9**: 14.

[19] Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, Pombo A, Nicodemi M (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(40): 16173–8.

[20] Mirny La (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, **19**(1): 37–51.

[21] Grosberg AY, Nechaev S, Shakhnovich E (1988) The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique*, **49**(12): 2095–2100.

[22] Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3): 215–6.

[23] 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.

[24] Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS computational biology*, **9**(3): e1002968.

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

[26] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall Ka, Phillippy KH, Sherman PM, *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, **41**(Database issue): D991–5.

[27] Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. *Nucleic acids research*, **39**(Database issue): D19–21.

[28] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4): 357–9.

[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] Zhang Y, Liu T, Meyer Ca, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**(9): R137.

[31] Breiman L (2001) Random Forests. *Machine learning*, **45**(1): 5–32.

[32] Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News*, **2**(December): 18–22.

[33] 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.

[34] Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology*, **88**(11): 2783–92.

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

[36] 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.

[37] Breiman L (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3): 199–231.

[38] 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.

[39] Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, **9**(3): 432–41.

[40] Mazumder R, Hastie T (2012) The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, pp. 1–21.

[41] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes Ja, *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2): 827–41.

[42] Furey TS (2003) Integration of the cytogenetic map with the draft human genome sequence. *Human Molecular Genetics*, **12**(9): 1037–1044.

[43] Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis Ja, Bickmore Wa (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, **10**(3): 211–9.

[44] 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.

[45] de Wit E, Bouwman BaM, Zhu Y, Klous P, Splinter E, Verstegen MJaM, Krijger PHL, Festuccia N, Nora EP, *et al.* (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, pp. 1–7.

[46] Tanabe H, Müller S, Neusser M, von Hase J, Calcagno E, Cremer M, Solovei I, Cremer C, Cremer T (2002) Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(7): 4424–9.

[47] 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.

[48] 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.

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

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

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

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

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

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

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

[64] 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.