BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 21 2012, pages 2789–2796
doi:10.1093/bioinformatics/bts529

Gene expression

Advance Access publication September 6, 2012

Genome-wide *in silico* prediction of gene expression

Robert C. McLeay[1], Tom Lesluyes[1,2], Gabriel Cuellar Partida[1] and Timothy L. Bailey[1,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia and [2]Département d'Informatique, Université Bordeaux 1 Sciences et Technologies, 33405 Talence, France

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Modelling the regulation of gene expression can provide insight into the regulatory roles of individual transcription factors (TFs) and histone modifications. Recently, Ouyang *et al.* in 2009 modelled gene expression levels in mouse embryonic stem (mES) cells using *in vivo* ChIP-seq measurements of TF binding. ChIP-seq TF binding data, however, are tissue-specific and relatively difficult to obtain. This limits the applicability of gene expression models that rely on ChIP-seq TF binding data.

Results: In this study, we build regression-based models that relate gene expression to the binding of 12 different TFs, 7 histone modifications and chromatin accessibility (DNase I hypersensitivity) in two different tissues. We find that expression models based on computationally predicted TF binding can achieve similar accuracy to those using *in vivo* TF binding data and that including binding at weak sites is critical for accurate prediction of gene expression. We also find that incorporating histone modification and chromatin accessibility data results in additional accuracy. Surprisingly, we find that models that use *no* TF binding data at all, but only histone modification and chromatin accessibility data, can be as (or more) accurate than those based on *in vivo* TF binding data.

Availability and implementation: All scripts, motifs and data presented in this article are available online at http://research.imb.uq.edu.au/t.bailey/supplementary_data/McLeay2011a.

Contact: t.bailey@imb.uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 19, 2011; revised on July 7, 2012; accepted on August 23, 2012

1 INTRODUCTION

A key challenge in modern genomics is understanding the regulatory interactions that occur to control gene expression. We can use predictive modelling techniques such as linear regression to both determine the accuracy of our understanding of the mechanisms of the regulation of gene expression, and to interpret and offer insight into the roles played by the various regulatory proteins, RNAs and other factors.

Perhaps the most studied and best understood mechanism of gene regulation is transcription factor (TF) binding. TFs characteristically have sequence-specific DNA-binding preferences. They typically bind sites that are short (5–15 bp), and often degenerate. Not surprisingly, naïve computational approaches that simply scan genomic sequence for matching sites are highly prone to making false-positive predictions (Wasserman and Sandelin, 2004). In an attempt to overcome this, methods incorporating evolutionary information regarding the conservation of each putative site such as BBLS (Xie *et al*., 2009) have been proposed. In reality, however, most researchers aim to make tissue-, cell- or condition-specific (hereafter 'tissue-specific') inferences regarding gene regulation, which evolutionary information alone cannot provide.

Histones are subject to many different covalent post-translational modifications such as methylation and acetylation. These modifications can alter chromatin structure and function and are known to comprise a tissue-specific histone 'code', read by proteins to effect changes in gene regulation and expression (van Ingen *et al*., 2008; Wang *et al*., 2008). Recently, global chromatin-immunoprecipitation with massively parallel sequencing (ChIP-seq) 'maps' of many histone modifications have become available through the ENCODE (Rosenbloom *et al*., 2010) and NIH Roadmap Epigenomics (Bernstein *et al*., 2010) projects.

ChIP-seq is not just limited to interrogating histone state, but also has become the gold-standard method for the genome-wide prediction of tissue-specific TF-bound regulatory regions. Unfortunately, however, TF ChIP-seq suffers from a major drawback: only one TF (in a single tissue) may be analysed at a time. It would be an extremely significant endeavour to predict the binding of the >1391 known human sequence-specific TFs (Vaquerizas *et al*., 2009) in one tissue or condition. As a result, even large-scale projects to date have focused either on a small number of TFs across several tissues (e.g. 10 TFs in modENCODE; Gerstein *et al*., 2010), or on a larger number of TFs in relatively few tissues (e.g. three primary tissues in ENCODE; Rosenbloom *et al*., 2010). Additionally, determining an appropriate cutoff for declaring putative binding sites is difficult and has been the focus of substantial research (Chen *et al*., 2008; Fejes *et al*., 2008; Valouev *et al*., 2008; Zhang *et al*., 2008).

ChIP-seq TF binding data have previously been used to construct a model of gene expression (Ouyang *et al*., 2009) in mouse embryonic stem (mES) cells. They used ChIP-seq binding data for 12 TFs that are known to contribute heavily to maintaining pluripotency (Chen *et al*., 2008) and thus are considered key factors regulating gene expression in ES cells. Ouyang *et al*. (2009) performed linear regression, correlating binding by these TFs with the expression of all genes and found that it could explain 65% of the variation in gene expression in ES cells ($r = 0.81$). This study did not consider the role of histone modifications on expression, but Karlić *et al*. (2010) built a similar regression model for human CD4 T+ cells, using histone

*To whom correspondence should be addressed.

modification data rather than TF binding data and also managed to explain a high degree of expression variance ($r = 0.74$).

In this study, we use TF binding, histone modification and chromatin accessibility data to address three major questions: (i) can we accurately predict gene expression using *in silico* predictions of TF binding rather than *in vivo* binding data? (ii) Does adding histone modification and chromatin accessibility data improve the accuracy of predicted gene expression? (iii) Are weak binding sites important to accurately predict gene expression?

To answer these questions, we derive log-linear regression-based models that relate gene expression to the binding of 12 different TFs, to 7 histone modifications and to chromatin accessibility. We find that expression models based on computationally predicted TF binding can achieve similar accuracy to those using *in vivo* TF binding data and that including binding at weak sites is critical for accurate prediction of gene expression. We also find that incorporating histone modification and chromatin accessibility data results in additional accuracy. Surprisingly, we find that models that use *no* TF binding data at all, but only histone modification and chromatin accessibility data, are essentially as accurate as those based on *in vivo* TF binding data.

2 METHODS

To predict gene expression, we build a log-linear regression model of gene expression as a function of TF binding, histone modifications and chromatin accessibility. Each gene is represented as a vector of features, where each feature represents either the predicted binding of a TF, the presence of a histone modification or the chromatin accessibility around the transcription start site (TSS) of the gene. The model is fit to the level of the expression of each gene as measured by RNA-seq.

2.1 Predicting TF binding (TF affinity scores)

For each TF–gene pair, we integrate either measured or predicted TF binding around each gene promoter into a single TF affinity score (TFAS) using a weighted sum, where the weight decreases exponentially as a function of distance from the TSS. Following the definition given by Ouyang *et al.* (2009), we define the TFAS, A_{ij}, for the ith gene and the jth TF as the weighted sum over all predicted binding sites (k) as:

$$A_{ij} = \sum_{k} g_k \mathrm{e}^{-d_k/d_0}, \qquad (1)$$

where g_k is the score of an individual predicted TF binding site. We weight the contribution of each site to our integrated TFAS by the exponential decay term e^{-d_k/d_0}, where d_k is the distance in base pairs of the putative binding site (k) from the TSS and d_0 determines the decay rate. Following Ouyang *et al.* (2009), we set $d_0 = 5000$ for all TFs except E2f1, where we set $d_0 = 500$ (Ouyang *et al.*, 2009). We only include in this weighted sum putative binding sites within 30 000 bp of the TSS, as the weighted contributions of more distant sites are negligible. We evaluate four methods for predicting individual binding sites and assigning individual site scores (g_k), which we discuss below. We then log transform and quantile normalize the TFAS matrix A (Bolstad *et al.*, 2003).

2.1.1 ChIP-seq site affinity scores We use measured ChIP-seq binding data for the 12 TFs from the Supplementary Material supplied by Ouyang *et al.* (2009), which contains pre-calculated and normalized TFASs, where g_k is the height (in mapped tags) of each declared ChIP-seq peak.

*2.1.2 Non-tissue-specific in silico (**position weight matrix**) site affinity scores* We perform a position weight matrix (PWM) scan of the entire genome using FIMO (Grant *et al.*, 2011). We use the default options, which apply a P-value threshold of 10^{-4} to predict matches and a pseudocount of 0.1. For our binding site affinity scores (g_k), we use the log-likelihood score of each predicted site reported by FIMO.

2.1.3 Tissue-specific in silico (PWM + H3K4me3) site affinity scores As gene expression is tissue-specific, our third TF binding site prediction method integrates tissue-specific epigenetic data with PWM scanning. We use FIMO with a histone-based prior (Cuellar Partida *et al.*, 2012) using H3K4me3 data, as it has previously been shown to be effective as a filter for predicting TF binding sites (Whitington *et al.*, 2009). Using the tools published in Cuellar Partida *et al.* (2012), we create a prior from the Mikkelsen *et al.* (2007) H3K4me3 data. We use parameters $\beta = 1$ and $\alpha = 4 \times 10^5$. We scan the genome using this prior and the default FIMO settings. For our site affinity scores (g_k), we use the log-posterior ratio scores reported by FIMO.

2.2 Measuring histone modifications and chromatin accessibility (histone and DNase scores)

We calculate histone modification scores (histone scores) and chromatin accessibility scores (DNase scores) for each histone modification and for DNase I hypersensitivity at the TSS of each gene. Taking a region ±2000 bases relative to the TSS, we sum the number of mapped tags for each histone modification and for DNase (see Section 2.5). We define Z_{ik} for gene i and histone modification or DNase score k as:

$$Z_{ik} = \sum_{n=-2000}^{2000} \mathrm{TC}(k, i, n), \qquad (2)$$

where $\mathrm{TC}(k, i, n)$ is the number of mapped sequence tags for histone modification or DNase I hypersensitivity k at position n relative to gene i's TSS. As we do above for the TFASs, we then log-transform and quantile normalize the matrix Z (Bolstad *et al.*, 2003).

2.3 Linear regression

We describe genome-wide gene expression by a log-linear model:

$$\log(Y_i + \sigma) = \mu + \sum_{j=1}^{M} \beta_j A_{ij} + \sum_{j=1}^{H} \gamma_j Z_{ij} + \epsilon_i, \qquad (3)$$

where Y_i is the expression level for gene i, μ is the basal expression level, A_{ij} is the jth TFAS for the ith gene and β_j is the fitted coefficient for the jth TFAS. Similarly, Z_{ij} is the jth histone or DNase score for the ith gene and γ_j the fitted coefficient for the jth histone modification or DNase score. ϵ_i is the gene-specific error term. σ is a fitted term to avoid taking log(0); we fit σ using a held-out 20% of our data. To obtain confidence intervals, we bootstrap the regression 1000 times on the remaining 80% of our data.

2.4 Linear regression with principal component analysis

We decompose a matrix, C, consisting of the column-wise concatenation of our TFAS matrix, A and our histone and DNase score matrix, Z, using the singular-value decomposition $C = U\Sigma V^T$, as described in Ouyang *et al.* (2009). We create the matrix $X = U\Sigma$, which is of the form X_{ij}, where the columns are the principal components (PCs) and the rows are genes.

Using this transformed data, we perform the regression,

$$\log(Y_i + \sigma) = \mu + \sum_{j=1}^{M} \beta_j X_{ij} + \epsilon_i, \qquad (4)$$

where Y_i is the gene expression level, μ represents basal expression, M is the number of PCs, ϵ_i is a gene-specific error term and B_j is the fitted coefficient for the jth PC. We fit σ as previously described. We bootstrap the PC analysis (PCA) regression as previously described.

2.5 Data

2.5.1 Mouse embryonic stem cells The TSS of genes is taken from the gene annotation dataset for the mouse genome (mm8/NCBIM36.46) from the Ensembl FTP server (ftp://ftp.ensembl.org/). To exclude the possibility that we include multiple transcripts or isoforms of the same gene in our regression, we use the most 5′-located TSS from the Ensembl gene annotation data as each gene's TSS. We also exclude all genes from haplotype variants and unmapped contig regions. This, combined with the removal of any genes with low confidence RNA-seq mapping, leaves a set of 18 008 genes for analysis.

Histone modification ChIP-seq data for H3K4me1, H3K4me2, H3K4me3, H3K9me3, H4K20me3, H3K27me3 and H3K36me3 are from Mikkelsen *et al.* (2007) and Meissner *et al.* (2008) pre-compiled 'density' files (similar to 'wig' files).

Raw DNase I hypersensitivity sequence tags are from Schnetz *et al.* (2010), which we mapped using Bowtie (Langmead *et al.*, 2009).

We use TF ChIP-seq data for E2f1, Esrrb, Klf4, c-Myc, n-Myc, Nanog, Oct4, Smad1, Sox2 and Stat3 from Chen *et al.* (2008). Position weight matrices (PWMs) describing the DNA-binding preferences (motifs) for Esrrb, Klf4, c-Myc, n-Myc and Stat3 are taken from Chen *et al.* (2008). For E2f1, Nanog and Smad1, we use PWMs from UniProbe (Berger and Bulyk, 2009), JASPAR (Portales-Casamar *et al.*, 2010) and TRANSFAC (Matys *et al.*, 2003), respectively. For Oct4 and Sox2, for which Chen *et al.* (2008) reported a composite Oct-Sox motif, we trimmed off the non-Oct and non-Sox portion from the Oct4 and Sox2 motifs, respectively. For full details, please see the Supplementary Material.

We use an extremely deep RNA-seq mouse expression dataset (Cloonan *et al.*, 2008), which was re-normalized in reads per kilobase of exon per million tags (RKPM) by Ouyang *et al.* (2009). RKPM has been previously shown to be proportional to mRNA abundance (Mortazavi *et al.*, 2008).

2.5.2 GM12878 cells The TSS of genes is taken from the gene annotation dataset for the human genome (hg19/GRCh37) from the Ensembl FTP server (ftp://ftp.ensembl.org/). To exclude the possibility that we include multiple transcripts or isoforms of the same gene in our regression, we use the most 5′-located TSS from the Ensembl gene annotation data as each gene's TSS. We also exclude all genes from haplotype variants and unmapped contig regions. This, combined with the removal of any genes with low confidence RNA-seq mapping, leaves a set of 37 458 genes for analysis.

Histone modification ChIP-seq data for H3K4me1, H3K4me2, H4K20me1, H3K27me3 and H3K36me3 are from the ENCODE project 'bigWig' files produced by the Broad laboratory (ENCODE, 2011). H3K4me3 was remapped onto hg19 from the 'fastq' file produced by the Broad laboratory using BowTie (Langmead *et al.*, 2009). Raw DNase I hypersensitivity data are from the 'signal' bigWig file produced by the Duke laboratory as part of the ENCODE project (ENCODE, 2011).

We use TF ChIP-seq data for c-Fos, Ctcf, Egr1, Nrf1 Nrsf, Pou2f2, Sp1, Srf, Stat3, Usf1 and Yy1.

We also use an ENCODE-provided RNA-seq dataset for our GM12878 gene expression data, which we mapped onto hg19 and normalized into RKPM using TopHat (Trapnell *et al.*, 2009). Full details on the sources of the RNA-seq data, ChIP-seq data and the PWMs are provided in the Supplementary Material.

2.6 Implementation

We implement the log-linear regression models in R (R Development Core Team, 2008). All R scripts, motifs and data (including expression data, TFASs, histone and DNase scores) are available online at http://research.imb.uq.edu.au/t.bailey/supplementary_data/McLeay2011a/.

3 RESULTS

3.1 Using *in silico* TF binding predictions to model gene expression

We first explore whether *in silico* TF binding predictions can replace ChIP-seq binding measurements. To do this, we build models using only TF binding features and compare models of gene expression based on TF ChIP-seq binding data with models based on *in silico* TF binding estimates (Fig. 1a). As expected, models using *in vivo* measurements of TF binding (ChIP-seq TFAS) explain substantially more variance (adjusted $R^2 = 0.644$; Fig. 1a, bar 3) than models using either the non-tissue-specific *in silico* (PWM) or tissue-specific *in silico* (PWM + H3K4me3) affinity scores. We find that a model using tissue-specific *in silico* TF binding predictions (PWM scanning with a H3K4me3 prior) is much more predictive of expression than the non-tissue-specific *in silico* affinity score model (adjusted $R^2 = 0.521$ and adjusted $R^2 = 0.279$; Fig. 1a, bars 1 and 2, respectively), nearly doubling the explained variance.

We hypothesize that possibly one or more of the TFs are poorly described by their PWM, hampering the performance of our models that use *in silico* binding predictions. To explore this, we take the ChIP-seq TFAS model and replace a single TF with the equivalent tissue-specific *in silico* data. Replacing the single E2f1 tissue-specific *in silico* TFAS with ChIP-seq E2f1 data yields a model that is statistically indistinguishable from one using all 12 ChIP-seq datasets (Fig. 1b, top two bars. $R^2 = 0.630$ and $R^2 = 0.644$, respectively, error bars overlap). Replacing any other tissue-specific *in silico* TFAS with a ChIP-seq TFAS has either minimal or no effect (Fig. 1b, compare bottom bar with others), with the exception of Sox2, where its inclusion surprisingly decreases R^2 (see Section 4). Thus, with our *in silico* method and ChIP-seq data for one TF, we can achieve indistinguishable accuracy to using 12 ChIP-seq datasets.

Having observed that E2f1 is the sole TF, where replacing tissue-specific *in silico* E2f1 data with ChIP-seq TF data is sufficient to increase explained gene expression variance to be indistinguishable from using 12 TF ChIP-seq datasets, we also examine whether in an all ChIP-seq TFAS model, replacing the E2f1 ChIP-seq data with *in silico* data decreases the accuracy of the model. In our model using all ChIP-seq TFASs, we find that replacing any ChIP-seq TFAS except E2f1 with PWM + H3K4me3 data has no statistically significant effect on model accuracy (Fig. 1c, compare bottom bar with others). Replacing E2f1 ChIP-seq TFAS with the E2f1 PWM + H3K4me3 TFAS, however, causes a large decrease in R^2, suggesting that the tissue-specific *in silico* binding predictions fail to completely capture the binding of E2f1. For any other TF in the model, replacing ChIP-seq data with tissue-specific *in silico* binding data causes no significant decrease in accuracy.

When we regress each ChIP-seq TFAS alone against expression ('Results' section in the Supplementary Material; Fig. 3), we find that E2f1 is the most individually informative TF. It has been suggested that E2f1 binds <20% of its binding sites directly in human ES cells (Bieda *et al.*, 2006). If this is the case, PWM-based methods would therefore be unable to predict the majority of E2f1 binding sites. Given the importance of E2f1 in this model, it is not surprising that models including *in silico* E2f1 data are less predictive of gene expression.

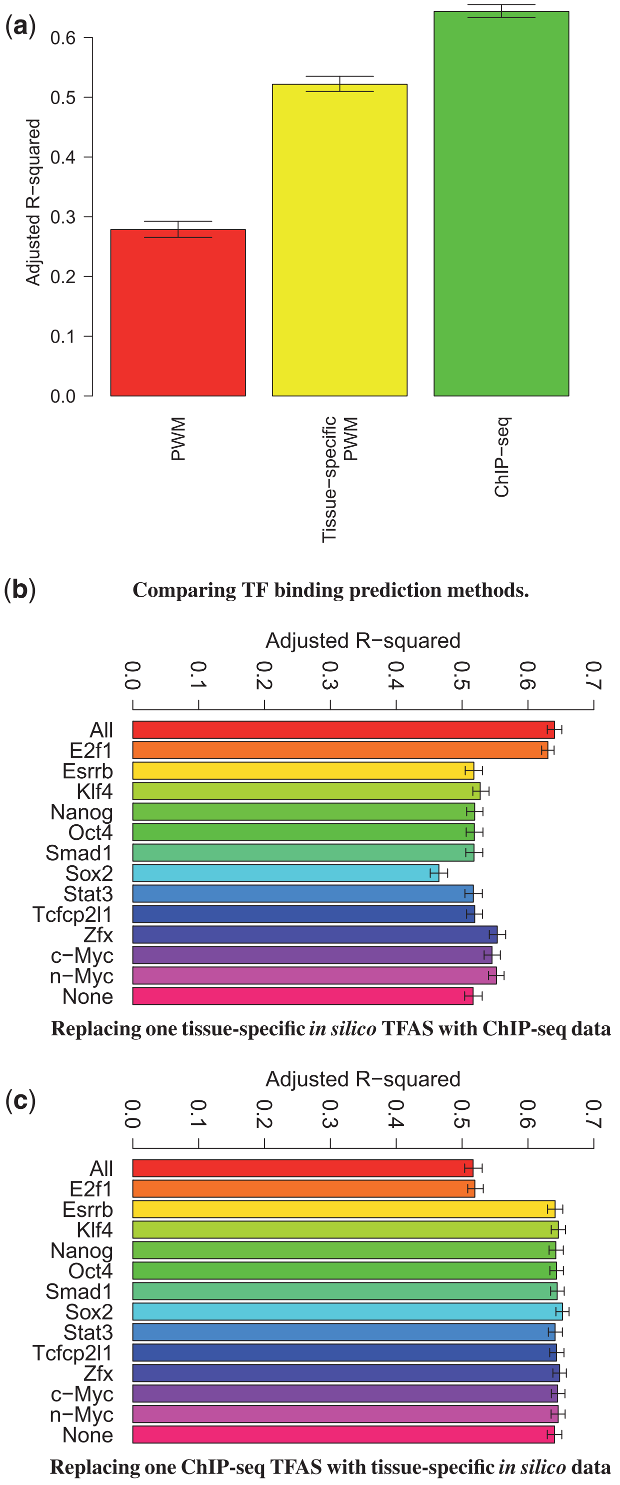
3.2 A combination of histone modification and DNase I hypersensitivity data is as predictive of expression as TF ChIP-seq data

Next, we examine whether integrating histone modification data and chromatin accessibility information (DNase) will improve the predictive accuracy of both tissue-specific *in silico* TFAS and ChIP-seq TFAS models of gene expression. We perform log-linear regression using our 12 TF affinity scores, both alone and combined with histone modification and chromatin accessibility data. In Figure 2, we compare the descriptive power (expressed as adjusted R^2) of four models: ChIP-seq TFAS only, tissue-specific *in silico* TFAS with histone modification and DNase data, ChIP-seq TFAS with histone modification and DNase data, and histone modification and DNase data alone (no TF binding data).

We find that models that use *no* TF binding data at all are surprisingly accurate. A model using just seven histone scores and the DNase score (bar 2, Fig. 2) is as descriptive of gene expression as the model using the ChIP-seq TF binding data for all 12 TFs (bar 1, Fig. 2). There is no statistically significant difference in the accuracy of these two models ($R^2 = 0.624$ *cf.* $R^2 = 0.644$, 95% confidence intervals overlap). While it is well known that histone modifications play key roles in mediating gene expression (Karlić *et al.*, 2010), it is still surprising that a model based on only these seven histone modifications and chromatin accessibility, and not on TF binding data, predicts expression almost as well as a model using the 12 TFs known to be key regulators of gene expression in pluripotent mES cells (Chen *et al.*, 2008).

Next, we determine whether combining histone modification and chromatin accessibility data with our tissue-specific *in silico* TF binding predictions improves predictive accuracy. We find that including incorporating histone modification and chromatin accessibility data with our tissue-specific *in silico* (PWM + H3K4me3) method to be as effective as using ChIP-seq TFAS data ($R^2 = 0.649$ and $R^2 = 0.644$, Fig. 2, bars 3 and 1, respectively, error bars overlap). We also find that our PWM + H3K4me3 method with histone and DNase scores to be more effective than only using histone modification and DNase data ($R^2 = 0.649$ *cf.* $R^2 = 0.624$, Fig. 2, bars 3 and 2, respectively, error bars do not overlap).

Finally, combining histone modification and chromatin accessibility data with ChIP-seq TFAS (right-most bar, Fig. 2)



Comparing TF binding prediction methods.

Replacing one tissue-specific *in silico* TFAS with ChIP-seq data

Replacing one ChIP-seq TFAS with tissue-specific *in silico* data

Fig. 1. TF-only models of gene expression in mES cells. Each bar shows the proportion of gene expression variance explained by each TF-only model (adjusted R^2). Error bars correspond to 95% confidence intervals estimated via bootstrapping. (**a**) The accuracy of models based on non--tissue-specific *in silico* scores (PWM), tissue-specific *in silico* scores (PWM + H3K4me3) or TF ChIP-seq data (ChIP-seq). (**b**) The accuracy of models where the TFASs for all but the named TF are based on tissue-specific *in silico* scores, and the TFAS for the named TF is from ChIP-seq data. The bar marked 'All' ('None') is for a model where all (no) TFASs are based on ChIP-seq data. (**c**) The accuracy of models where the TFASs for all but the named TF are based on ChIP-seq data and the TFAS for the named TF is based on tissue-specific *in silico* scores. The bar marked 'All' ('None') is for a model where all (no) TFASs are based on tissue-specific *in silico* data

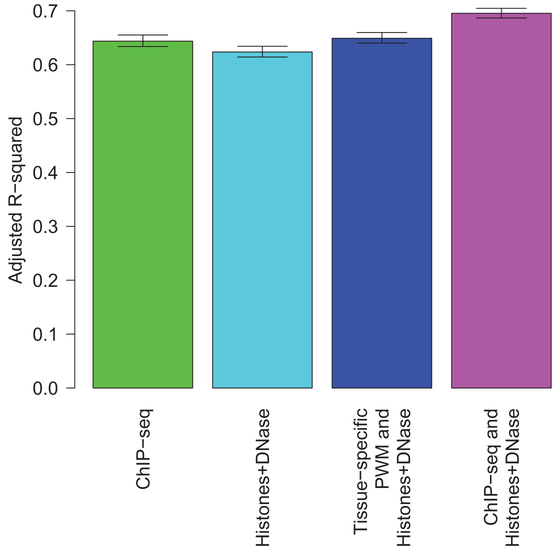


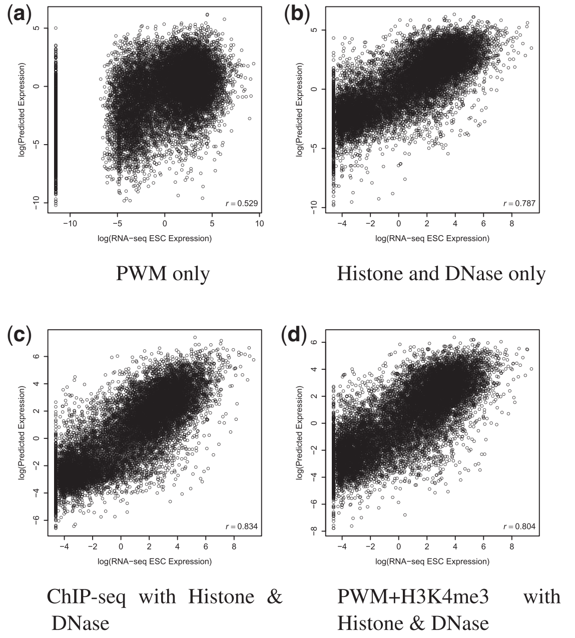
Fig. 2. Proportion of gene expression variance explained by each model (adjusted R^2). Each bar corresponds to the proportion of the variance in RNA-seq gene expression levels explained by each model. Error bars correspond to 95% confidence intervals estimated via bootstrapping. The ChIP-seq model consists of 12 TFASs, while the remaining three models are histone and DNase scores only, tissue-specific PWM (PWM + H3K4me3) TFAS with histone and DNase scores, and ChIP-seq TFAS with histone and DNase scores, respectively

provides a significant improvement to the regression fit ($R^2 = 0.695$) over ChIP-seq TFAS alone. Both the histone and DNase score only model, and the ChIP-seq TFAS only model, are quite descriptive of gene expression. The improvement gained by combining histone, DNase and TF binding data suggests that there is additional information available in the histone and DNase data that is not present in the TF ChIP-seq data.

3.3 Predicted expression correlates well with measured expression

It is also informative to plot the predicted expression against the actual RNA-seq expression data for our models (Fig. 3a–d) and to calculate the Pearson correlation coefficients (r). For several of our methods (tissue-specific *in silico* TF predictions with histone and DNase data; ChIP-seq TF data only; ChIP-seq TF data with histone modification and DNase data; histone modification and DNase data alone), we actually achieve higher correlation between our predicted expression values and actual (RNA-seq) expression values than that observed between RNA-seq and microarray expression experiments in the same tissue ($r = 0.734$; see the Supplementary Material to Ouyang *et al.*, 2009).

The predicted and actual expression values are fairly well correlated in the model that is based on ChIP-seq TF, histone and DNase scores (Fig. 3c, $r = 0.834$) and in the model t based on tissue-specific *in silico* TF, histone and DNase scores (Fig. 3d, $r = 0.806$). The correlation in the non-tissue-specific *in silico* model (PWM, Fig. 3a). We find only a relatively small difference in accuracy between these models, both having predicted expression values that correlate well with actual expression.



PWM only

Histone and DNase only

ChIP-seq with Histone & DNase

PWM+H3K4me3 with Histone & DNase

(e)

TFAS Method	r (TFAS only)	r (Histone & DNase)
None	-	0.790 (0.007)
PWM	0.530 (0.012)	-
PWM+H3K4me3	0.724 (0.008)	0.806 (0.006)
ChIP-seq	0.804 (0.007)	**0.835** (0.005)

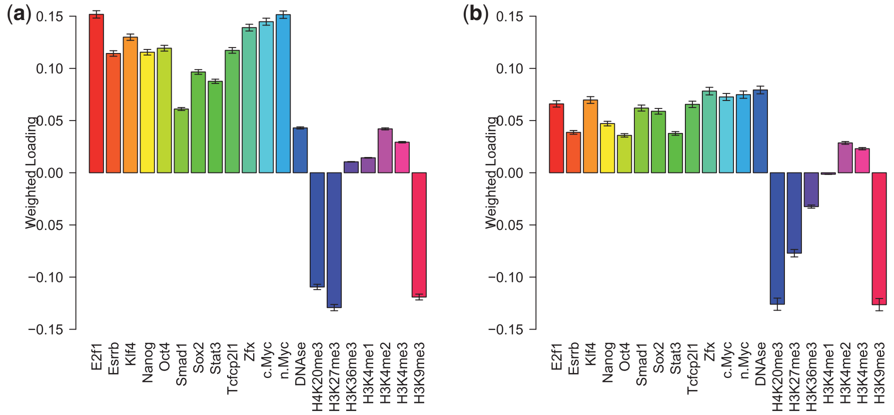
Table of r values for all methods

Fig. 3. (**a–d**) Correlation of actual RNA-seq expression versus predicted expression in models of mES gene expression. Plots show measured (X) versus predicted expression (Y) of each gene. PWM is the non-tissue-specific *in silico* method, and PWM + H3K4me3 refers to the tissue-specific *in silico* method. (**e**) Table of r (Pearson correlation coefficient, bootstrapped) comparing predicted expression versus actual expression for each model tested. 95% confidence intervals are reported in parentheses

3.4 Principal components provide information on TF and histone roles

We then perform PCA using singular value decomposition of TF, histone modification and chromatin accessibility (DNase) data, decomposing a joint matrix (with TF, histone modification and DNase data as columns) into PCs (see Section 2.4). Using these components rather than TFASs, and histone and DNase scores directly, we perform log-linear regression. PCA allows the extraction of a small number of non-correlated PCs responsible for the majority of gene variance. We examine the role of each TF, histone modification and chromatin accessibility in the regression. We plot the relative loading of each feature within the most informative PC (PC2), weighted by the proportion of variance explained (adjusted R^2) by that PC (Fig. 4).

We find that the tissue-specific *in silico* TF PCA model is very similar to the ChIP-seq TF PCA model at providing biological information on the regulatory roles of the features, especially in the most important PCs. Visually, a high degree of similarity between the PCs is present, with only minor differences (comparing Fig. 4a and b). Both models show all TFs as positively associated with transcription, with E2F1, Klf4, Zfx, c-Myc and



ChIP-seq PC2

Tissue-specific *in silico* PC2

Fig. 4. The relative loading of each feature in the second principal component for two models. (**a**) The relatively loading of each feature of PC2 for a model containing ChIP-seq TFAS with histone and DNase scores. (**b**) The relative loading of each feature of PC2 in a model containing tissue-specific *in silico* TFAS (PWM + H3K4me3) with histone and DNase scores. Each bar corresponds to the weighted loading of a TF or histone mark in the regression. We weight the loadings of each TFAS or histone and DNase score by the variance explained by the principal component to which they belong. See 'Results' section in the Supplementary Material for all principal components

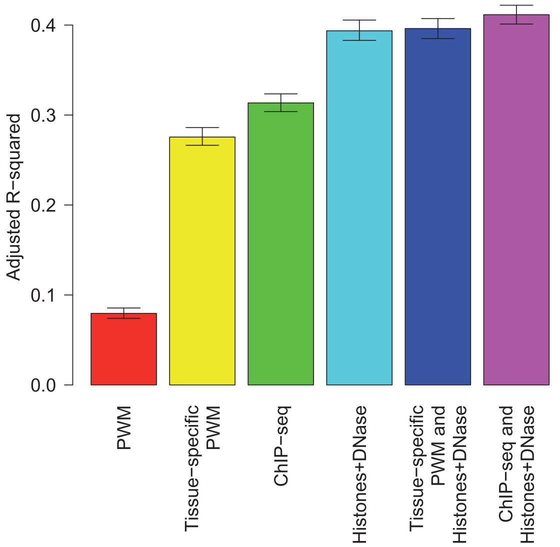


Fig. 5. Proportion of GM12878 gene expression variance explained by each model (adjusted R^2). Each bar corresponds to the proportion of the variance in RNA-seq gene expression levels explained by each model. Error bars correspond to 95% confidence intervals estimated via bootstrapping

n-Myc as the most weighted TFs. Likewise, the weighted loading on each histone modification is qualitatively similar between models. Figure 4 shows the second PC, which alone is the most informative PC in the regression. When a model with PC2 as the only feature is fit to expression, it explains almost half of gene expression variance ($R^2 = 0.475$). The remaining PCs and a short discussion are provided in 'Results' section in the Supplementary Material.

Figure 4 shows the relative loading of each component within PC2 (which accounts for ≈48% of explained variance) successfully recapitulating known knowledge of histone modification effects. Both models suggest that H4K20me3, H3K27me3 and H3K9me3 repress gene transcription. Both H3K9me3 and H4K20me3 are known to be involved in gene silencing, associating with heterochromatin (Schotta *et al.*, 2004). H3K27me3 is associated with facilitating binding of the Polycomb repressor complex 1 (Cao *et al.*, 2002). In PC2, both the loadings of H3K4me2 and H4K4me3 suggest activating effects, which are well reported throughout the literature (e.g. Barski *et al.*, 2009; van Ingen *et al.*, 2008). A difference between the ChIP-seq and tissue-specific *in silico* (PWM + H3K4me3) PCA models is noted for H3K36me3, which is known to maintain genes in a 'poised' state for either activation or repression. The similarity between the predicted roles by our tissue-specific *in silico* method and the roles predicted by the ChIP-seq data strongly suggests that our *in silico* method is capable of making biologically sound predictions of TF and histone roles.

3.5 Confirmation in GM12878 cells

To insure that our results were not specific to mES cells, we repeated our experiments using equivalent data from GM12878 cells. The pattern of the results are extremely similar to those using mES cell data, with models based on TF ChIP-seq, histone ChIP-seq and DNase I hypersensitivity explaining more gene expression variance than models based on TF ChIP-seq alone (Fig. 5, bars 3 and 6). Compared with the mES cell models, the relative improvement in accuracy of the ChIP-seq-based gene expression model when chromatin modification and accessibility data are added is much larger (compare Fig. 2 and Fig. 5). This is consistent with the fact that the model based on histone and DNase I data alone (without any TF-specific data) is much more accurate than model based on TF ChIP-seq data. This may be due to the set of TFs we use with the mES cell data being more complete (with respect to the complete set of key regulators) than the set we use in the GM12878 experiments. We note as well that the model using PWMs plus histone and DNase data is far more predictive of gene expression than the model based solely on TF ChIP-seq data (Fig. 5, bars 3 and 5).

We note that compared with our previous results in mES cells, each model explains comparatively less expression variance in GM12878 cells (compare Figs 1a and 2 with Fig. 5). For example, the best performing model in both tissues—ChIP-seq binding measurements combined with histone modification and chromatin accessibility data—explains ≈30% less of the total gene expression variance in GM12878 cells than in mES cells ($R^2 = 0.412$ and $R^2 = 0.695$, respectively). As noted earlier, this may be partially due to the TFs used in the model. Additionally, one of the most informative histone modifications in mES cells, H4K20me3 (Fig. 4) was unavailable for GM12878 cells.

We also measure correlation between actual and predicted expression in GM12878 cells for each regression model and perform the singular value decomposition regressions in GM12878 cells and these results are included in the Supplementary Material.

4 DISCUSSION

We have successfully explained the majority of variance in gene expression in mES cells by building an integrated model of 12 TFs, seven histone modifications and DNase hypersensitivity data. In fact, we find that gene expression levels predicted by both the tissue-specific *in silico* models and the histone modification and DNase-only model are more highly correlated with mESC RNA-seq gene expression data than microarray data from the same tissue type. The wide availability of TF binding site motifs (Berger and Bulyk, 2009; Matys *et al*., 2003; Portales-Casamar *et al*., 2010) gives our *in silico* method the advantage of allowing the evaluation of many different TFs without the cost and time required to perform TF ChIP-seq experiments, instead requiring only a PWM and H3K4me3 data in the tissue of interest.

Surprisingly, we find that using the seven histone modifications and chromatin accessibility (DNase) data alone produces a model as descriptive of expression as one using 12 TF ChIP-seq datasets. By using histone modifications and DNase data alone, we explain more variation (up to ≈70%) in gene expression than other studies that use more complex *in silico* models of TF binding, which vary from ≈11% (Das *et al*., 2006) to ≈33% (Das *et al*., 2004). Several of these models incorporate TF–TF interactions and other higher order effects compared with our simpler log-linear regression models. We also find that models combining histone data with TF ChIP-seq data outperform models using TF ChIP-seq data alone. This suggests that the histone modifications perhaps capture an ‘expression state’, containing information not captured by our 12 TFs.

Although using tissue-specific *in silico* (PWM + H3K4me3) predictions of our 12 TFs alone does achieve a less-accurate fit to RNA-seq expression data than using 12 ChIP-seq TFAS, we find that replacing a single TFAS (E2f1) with ChIP-seq data improves the accuracy of our tissue-specific *in silico* method to be indistinguishable from using all 12 ChIP-seq TFASs. We hypothesize that this is due to both E2f1 being the most informative of these 12 TFs (see Supplementary Material) for predicting expression, and the majority of its binding being indirect (Bieda *et al*., 2006) and thus not directly measurable by sequence-based methods.

Replacing Sox2 *in silico* TFAS data with Sox2 ChIP-seq TFAS data, however, actually decreases the fit of the regression (Fig. 1a–c). We are unsure of why this is the case, although it may be due to noise in the ChIP-seq dataset for Sox2. Re-analyzing the Sox2 ChIP-seq data with more accurate ChIP-seq peak calling algorithms may shed light on why the *in silico* Sox2 TFAS appears to give an improvement in fit over the ChIP-seq Sox2 TFAS. A recent article shows that models using this TF ChIP-seq data fit expression more accurately when the ChIP-seq data are re-analyzed with more accurate peak calling algorithms (Park and Nakai, 2011).

Using PCA, we extended the study of the roles of TFs (Ouyang *et al*., 2009) to also include the roles of histone modifications. The results we obtain, accurately assigning (known) roles to histone modifications, suggests that this method is of value for exploring regulation. Applying this method to the full set of histone modification data from the NIH Roadmap Epigenomics Project (Bernstein *et al*., 2010) would be of particular interest, especially for the examination of less well-studied histone modifications, due to the availability of data across a broad spectrum of tissues.

Some histone modifications, for example, H4K20me3 are known to act combinatorially with other marks such as H3K4me1 (Balakrishnan and Milavetz, 2010). Similarly, some marks are mutually exclusive with other marks (e.g. H3K27me3 and H3K27ac) (Pasini *et al*., 2010). These marks have also been suggested to potentially play different regulatory roles at different genomic loci (Kouzarides, 2007). It has been suggested that the combination of marks around a TSS or enhancer form a ‘histone state’ (Ernst *et al*., 2011). Using the histone states defined by Ernst *et al*. (2011) rather than raw tag counts may prove to be more effective in our linear model at capturing the combinatorial effects of histone modifications.

Funding: Australian Postgraduate Award and the Queensland Government Department of Tourism, Regional Development and Industry (to R.C.M.) and National Institutes of Health (RO-1 RR021692 to T.L.B.).

Conflict of Interest: none declared.

REFERENCES

Balakrishnan,L. and Milavetz,B. (2010) Decoding the histone h4 lysine 20 methylation mark. *Crit. Rev. Biochem. Mol. Biol*., **45**, 440–452.

Barski,A. *et al*. (2009) Chromatin poises mirna- and protein-coding genes for expression. *Genome Res*., **19**, 1742–1751.

Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc*., **4**, 393–411.

Bernstein,B.E. *et al*. (2010) The nih roadmap epigenomics mapping consortium. *Nat. Biotechnol*., **28**, 1045–1048.

Bieda,M. *et al*. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res*., **16**, 595–605.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Cao,R. *et al*. (2002) Role of histone h3 lysine 27 methylation in polycomb-group silencing. *Science*, **298**, 1039–1043.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mrna sequencing. *Nat. Methods*, **5**, 613–619.

Cuellar Partida,G. *et al.* (2012) Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**, 56–62.

Das,D. *et al.* (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.

Das,D. *et al.* (2006) Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.*, **2**, 2006.0029.

ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Fejes,A.P. *et al.* (2008) Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.

Gerstein,M.B. *et al.* (2010) Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, **330**, 1775–1787.

Grant,C.E. *et al.* (2011) Fimo: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Karlić,R. *et al.* (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.

Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.*, **10**, R25.

Matys,V. *et al.* (2003) Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Meissner,A. *et al.* (2008) Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.

Mikkelsen,T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Ouyang,Z. *et al.* (2009) Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **106**, 21521–21526.

Park,S.-J. and Nakai,K. (2011) A regression analysis of gene expression in es cells reveals two gene classes that are significantly different in epigenetic patterns. *BMC Bioinformatics*, **12** (Suppl. 1), S50.

Pasini,D. *et al.* (2010) Characterization of an antagonistic switch between histone h3 lysine 27 methylation and acetylation in the transcriptional regulation of polycomb group target genes. *Nucleic Acids Res.*, **38**, 4958–4969.

Portales-Casamar,E. *et al.* (2010) Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

R Development Core Team. (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rosenbloom,K.R. *et al.* (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **38**, D620–D625.

Schnetz,M.P. *et al.* (2010) Chd7 targets active gene enhancer elements to modulate es cell-specific gene expression. *PLoS Genet.*, **6**, e1001023.

Schotta,G. *et al.* (2004) A silencing pathway to induce h3-k9 and h4-k20 trimethylation at constitutive heterochromatin. *Genes Dev.*, **18**, 1251–1262.

Trapnell,C. *et al.* (2009) Tophat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**, 1105–1111.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.

van Ingen,H. *et al.* (2008) Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3. *Structure*, **16**, 1245–1256.

Vaquerizas,J.M. *et al.* (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Whitington,T. *et al.* (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.

Xie,X. *et al.* (2009) MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, **25**, 167–174.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.