BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 18 2012, pages 2297–2303
doi:10.1093/bioinformatics/bts362

Gene expression

Advance Access publication June 23, 2012

Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models

Thais G. do Rego[1], Helge G. Roider[2], Francisco A. T. de Carvalho[1] and Ivan G. Costa[1,*]

[1]Center of Informatics, Federal University of Pernambuco, Recife 50740-560, Brazil and [2]Discovery Bioinformatics, Merck KGaA, Merck Serono, Darmstadt, 64293, Germany

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Blood cell development is thought to be controlled by a circuit of transcription factors (TFs) and chromatin modifications that determine the cell fate through activating cell type-specific expression programs. To shed light on the interplay between histone marks and TFs during blood cell development, we model gene expression from regulatory signals by means of combinations of sparse linear regression models.

Results: The mixture of sparse linear regression models was able to improve the gene expression prediction in relation to the use of a single linear model. Moreover, it performed an efficient selection of regulatory signals even when analyzing all TFs with known motifs (>600). The method identified interesting roles for histone modifications and a selection of TFs related to blood development and chromatin remodelling.

Availability: The method and datasets are available from http://www.cin.ufpe.br/~igcf/SparseMix.

Contact: igcf@cin.ufpe.br

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

Received on June 4, 2012; revised on June 4, 2012; accepted on June 18, 2012

1 INTRODUCTION

Blood development in mammals is an ideal system to study cell differentiation and proliferation. All blood cells arise from a single multipotent hematopoietic stem cell (HSC). This stem cell can differentiate in lymphoid and myeloid progenitors, which will give rise to erythrocytes and immune cells such as monocytes, megakaryocytes, B and T cells (Matthias and Rolink, 2005; Orkin and Zon, 2008; Rothenberg and Taghon, 2005). The different cell types are distinguishable by specific cell surface proteins and can readily be obtained from the blood of adult individuals through cell sorting. For the purified cell populations, cell type-specific expression profiles and DNA–protein interactions can subsequently be obtained using microarrays and chromatin immunoprecipitation (ChIP) assays (Novershtern *et al.*, 2011; Weishaupt *et al.*, 2010). Blood cell development is thought to be controlled by a circuit of transcription factors (TFs), that determine the cell fate through activating cell type-specific expression programs (Barreda and Belosevic, 2001; Dor and Crispino, 2011; Matthias and Rolink, 2005; Orkin and Zon, 2008; Rothenberg and Taghon, 2005). In addition, there is an increasing awareness of the role of chromatin structure in regulating expression during development (Goldberg *et al.*, 2007). In particular, the presence or absence of post-transcriptional histone modifications (HMs), termed the 'histone code', modulates the affinity of histones to DNA and thus determines whether a DNA region is accessible for the transcriptional machinery (Kouzarides, 2007; Turner, 2007). For instance, the histone marks H3k4me3, H3k79me and H3kac are known to be associated with genes that are either actively transcribed or whose transcription is readily activated upon stimulus while promoters bearing the histone marks H3k27me3 and H3k9me3 tend to be inactive. Understanding the interplay between particular HMs and TF binding is crucial for uncovering cell differentiation processes.

To shed light on the interplay between histone marks and TFs during blood cell development, in this study we model gene expression by means of combinations of linear regression models. The main idea behind this approach is to combine all regulatory signals to explain the expression pattern of the genes, as TFs and HMs can act in a multi-functional manner, conveying both transcriptional repression and activation depending on their location with respect to the transcription start site (TSS) and the presence of other TFs in the surroundings (Fig. 1). The value of the regression coefficients thereby indicates not only the importance of a particular regulatory signal (Bussemaker *et al.*, 2001; Karlic *et al.*, 2010; Keles *et al.*, 2002) but also whether a signal activates or represses transcription. We have recently shown that combining more than one regression model in a mixture, where each of the regression models explains the expression of a particular group of genes, improves the expression prediction and identification of important regulatory players (Costa *et al.*, 2011).

Previous works were however based on the analysis of only a small subset of possible regulatory signals (<100) (Bussemaker *et al.*, 2001; Costa *et al.*, 2011; Karlic *et al.*, 2010; Keles *et al.*, 2002), as standard linear regression suffers from over-fitting on high-dimensional space (Hastie *et al.*, 2003). In addition, regulatory signals tend to be correlated, with a given subset of TFs and/or HMs being present on the same group of promoters. Although it is important that the model indicates all regulatory signals that are important to a particular cell type, standard regression models give arbitrary coefficients for correlated (or co-linear) variables, failing to give a proper interpretation of all important variables (Tibshirani, 1996; Zou and Hastie, 2005). Here, we propose a novel methodology, mixture of sparse linear regression models, to describe the expression of genes. Sparse linear models perform a time efficient selection of important features even in the presence of a high number of regression variables. The sparse model is determined

* To whom correspondence should be addressed.

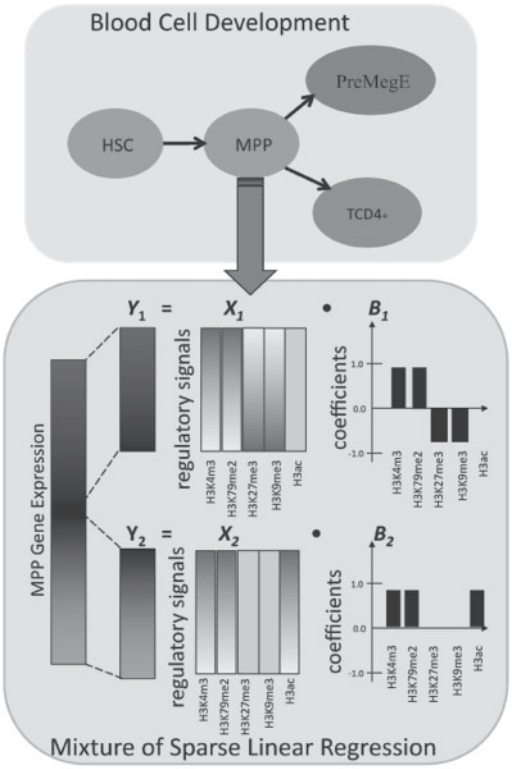

Fig. 1. Schematic blood cell developmental tree (top) and a sample mixture model inferred on the MPP cell (bottom). The mixture model predicts the gene expression of genes of a particular cell type Y—depicted as a red–green bar—by the regulatory signals of the genes X—depicted as the blue–white plot, where blue values indicate a higher presence of the histone in a gene promoter. The coefficients B indicate the roles of each regulatory signal. The mixture of sparse linear regression search for groups of genes, whose expression are determined by the same regulatory network. For example, model 1 predicts genes with high expression and indicates that H3k4me3 and H3k79me2 are activators of expression and H3k27me3 and H3k27me3 are repressors of expression. The elastic net method gives similar coefficients to co-linear signals, such as the pairs H3k79me2/H3k4me3 and H3k27me3/H3k9me3. Also, irrelevant signals, such as H3ac are removed, i.e. have the coefficient set to 0. Note that distinct models indicate distinct regulatory elements. For model 2, only the HMs H3k4me3, H3k79me2 and H3kac were selected as relevant for determining the activity of low expressed genes

with the elastic net algorithm from Zou and Hastie (2005) and Friedman *et al.* (2009). The elastic net displays the so-called grouping effect, that is, co-linear variables are simultaneously included in or excluded from the model and all grouped variables have equivalent regression coefficients (Fig. 1). Therefore, our method is not only capable of analyzing all TFs with known motifs (>600) but also allows to identify all regulatory signals (co-linear or not), which play an important role in a particular cell type.

1.1 Related work

The use of linear regression methods for predicting gene expression from TF binding sites were first proposed in Bussemaker *et al.* (2001) and Keles *et al.* (2002), and regarding HMs, in Karlic *et al.* (2010). In Keles *et al.* (2002), the problem of the dimensionality was approached with a computationally expensive backward feature selection. Ouyang *et al.* (2009) performed the prediction of gene expression on stem cells from a few TF binding sites derived from genome wide chromatin immunoprecipitation assays. Later, Park and Nakai (2011) proposed an extension of this work by integrating HM and DNA methylation data. The method was based on an initial discretization of histone marks to detect epigenetic states and indicated an advantage in combining both data types. Recently, Cheng *et al.* (2011) proposed a methodology for predicting the expression over genomic regions, as measured by RNA-Seq, for >50 regulatory signals during *Caenorhabditis Elegans* development. Regression predictions were based on a Support vector machine (SVM), which can deal with high dimensionality but did not indicated feature importance.

2 METHODS

2.1 Sparse linear regression

Here, we propose the use of a mixture of sparse linear models for modeling the expression of genes given their regulatory signals: presence of HMs or TF binding sites in the promoter region of a gene. Let y_i be the gene expression level of gene i (the dependent variable), and $x_i = (x_{i1}, ..., x_{iP})$ be a vector with the P regulatory signals (the regressor variables) of gene i, where $i = 1, .., N$. A single linear regression model can be defined as

$$y_i = x_i B + \epsilon_i, \tag{1}$$

where B is a vector $(b_1, ..., b_P)^T$ representing regression coefficients and ϵ_i is an error term (we ignore for simplicity the b_0 coefficient). We use an Elastic Net estimation for obtaining a sparse linear model (Zou and Hastie, 2005). For a given data $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ is a set on N observations x_i and $\mathbf{Y}$ a vector with N observations y_i, the elastic net is based on finding B, which minimizes the criterion (Zou and Hastie, 2005)

$$L(\lambda_1, \lambda_2) = \frac{1}{N}(\mathbf{Y} - \mathbf{X}B)^2 + \lambda_1 \cdot \lambda_2 |B| + \lambda_1 \cdot (1 - \lambda_2)|B|_2^2, \tag{2}$$

where $|B| = \sum_{j=1}^{P} |b_j|$ and $|B|_2^2 = \sqrt{(\sum_{j=1}^{P} b_j^2)}$.

The last two terms on the right are the L_1 and L_2 penalizations. The L_1 penalty, also denoted as lasso penalty, performs feature selection by shrinking some of the coefficients until reaching zero (Tibshirani, 1996). The L_2 penalty shrinks coefficients towards zero and is equivalent to the ridge regression (Hastie *et al.*, 2003). However, the L_2 penalty does not perform feature selection as it is unable to assign zero values to coefficients. The parameter λ_1 gives the stringency of the penalizations and λ_2, which varies from 0 to 1, balances between the L_1 and L_2 penalties. The main advantage of the elastic net is the so-called 'grouping effect', that is, highly correlated variables tend to be either included or excluded from the model in groups (Zou and Hastie, 2005). Such an effect is not present in the lasso penalization alone, which tends to include only one of the correlated features in the model.

2.2 Bayesian elastic net

We use a Bayesian interpretation of the elastic net. Therefore, we can easily plug in the models in a mixture model framework. Assuming the error ϵ in equation (1) follows a Normal distribution with variance σ^2, the linear regression model has the following distribution:

$$\mathbb{P}(y_i | x_i, B, \sigma^2) = \mathcal{N}(y_i | x_i B^{\mathbf{T}}, \sigma^2) \tag{3}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-(y_i - x_i B)^2 / 2\sigma^2}, \tag{4}$$

As pointed out by (Li and Lin, 2010), the L_1 and L_2 penalization terms are equivalent to a Laplace and Gaussian distribution, respectively, with mean 0,

that is

$$\mathbb{P}(B|\sigma^2) = \text{Laplace}(B|\mathbf{0}_P, \mathbf{I}_P\tau_1^{-1})\mathcal{N}(B|\mathbf{0}_P, \mathbf{I}_P\tau_2^{-1/2}) \quad (5)$$

$$= \frac{\tau_1}{2}^P \exp^{-|B|\tau_1} \frac{\tau_2}{2\pi}^{P/2} \exp^{-B^2\tau_2}, \quad (6)$$

where $\mathbf{0}_P$ is a P-dimensional vector with entries equal to 0 and $\mathbf{I}_P$ is a $P \times P$ identity matrix. We use a gamma distribution as prior to regularize σ^2 (Hastie *et al.*, 2003)

$$\mathbb{P}(1/\sigma^2|a,b) = \frac{b^a}{\Gamma(a)}(1/\sigma^2)^{a-1}e^{-b/\sigma^2}, \quad (7)$$

where a and b are hyper-parameters. Maximum-a-posteriori (MAP) estimates of the regression parameters can be determined by maximizing the posterior distribution

$$\mathbb{P}(B,\sigma^2|\mathbf{X},\mathbf{Y},\tau_1,\tau_2) = \prod_{i=1}^N \mathcal{N}(y_i|x_iB^{\mathbf{T}},\sigma^2)\mathbb{P}(B|\sigma_2,\tau_1,\tau_2)\mathbb{P}(\sigma^{-2}|a,b). \quad (8)$$

It is straightforward to see that the log of the posterior distribution is equivalent to equation (2) for $\tau_1 = N\lambda_1\lambda_2/\sigma^2$ and $\tau_2 = N\lambda_1(1-\lambda_2)/\sigma^2$. Therefore, the minimization of B for fixed σ^2, λ_1 and λ_2 is equivalent to the Elastic net and any algorithm for solving the Elastic Net can be applied.

2.3 Mixture of sparse linear models

A mixture of linear regression models is obtained by a convex summation of K linear distributions

$$\mathbb{P}(y_i|x_i,\Theta) = \sum_{k=1}^K \alpha_k \cdot \mathcal{N}(y_i|x_iB_k^{\mathbf{T}},\sigma_k^2), \quad (9)$$

where $\alpha = (\alpha_1,...,\alpha_K)$ are the mixture coefficients such that $\alpha_k \geq 0$, $\sum_{k=1}^K \alpha_k = 1$ and Θ are the model parameters $(\alpha, B_1,...,B_K,\sigma_1^2,...,\sigma_K^2)$. For given data $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ is a set on N observations x_i and $\mathbf{Y}$ a vector with N observations y_i, the mixture of sparse linear models is estimated with a (MAP) version of the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977; DeSarbo and Cron, 1988). The EM algorithm finding estimates Θ that maximize the posterior distribution over the data X and Y,

$$\mathbb{P}(\Theta|\mathbf{X},\mathbf{Y},\mathbf{Z}) \approx \mathbb{P}(\mathbf{Y},\mathbf{Z}|\mathbf{X},\Theta)\mathbb{P}(\Theta), \quad (10)$$

where $\mathbf{Z}$ is the vector of hidden variables with $z_i \in \{1,...,K\}$ indicating which linear model an observation i belongs to and $\mathbb{P}(Y,Z|X,\Theta)$ is the complete data likelihood. $\mathbb{P}(\Theta)$ is the prior distribution over the model parameters

$$\mathbb{P}(\Theta) = \mathbb{P}(\alpha)\prod_{k=1}^K \mathbb{P}(B_k)\mathbb{P}(\sigma_k^2), \quad (11)$$

where $\mathbb{P}(\alpha)$ follows a Dirichlet distribution and $\mathbb{P}(B_k)$ and $\mathbb{P}(\sigma_k^2)$ are equal to the Elastic Net priors defined in equation (6). The EM algorithm works by iteratively estimating the posterior probabilities that an observation belongs to a linear model and the parameters of the linear models Θ until convergence. Let r_{ik} be the posterior probability (or responsibility) (McLachlan and Peel, 2000) that observation i belongs to the linear model k:

$$r_{ik} = \mathbb{P}(z_i = k|y_i,x_i) = \frac{\alpha_k\mathcal{N}(y_i|x_iB_k,\sigma_k^2)}{\sum_{k'=1}^K \alpha_{k'}\cdot\mathcal{N}(y_i|x_iB_{k'},\sigma_{k'}^2)}. \quad (12)$$

Considering the linear model parameters, σ_k^2 is estimated as follows:

$$\sigma_k^2 = \frac{\sum_{i=1}^N r_{ik}(y_i - x_iB)^2 + 2b}{N + 3P + 2a - 2}, \quad (13)$$

where a and b are hyper-parameters. This is an approximation of σ_k^2, which is independent of the shrinkage of B_k, and yields good empirical results (Sun and Zhang, 2010). The estimator of B_k is analogous to the Elastic Net. Here we use an efficient implementation of the algorithm based on gradient descent described in Friedman *et al.* (2009), where observations are weighted by their posterior probabilities to give estimates for each mixture component.

2.4 Data

2.4.1 TF affinity In this work, we use the Transcription factor Affinity Prediction (TRAP) approach (Roider *et al.*, 2007) to predict the binding affinity of a given TF to a given promoter sequence. The TRAP method computes a continuous score estimating the expected number N of TFs bound to the promoter. As input, TRAP takes for each TF a position frequency matrix (PFM) suitable for computing mismatch energies and a DNA sequence of interest (see Roider *et al.* (2007) for details). PFMs represent how often a given base occurs at a given position within a set of aligned known binding sites of a TF. In our study, we use 599 PFMs from the Transfac database version 11.1 (Matys *et al.*, 2003). To minimize the number of false-binding predictions, we limit the analysis to proximal promoters covering the first 200 bp upstream of the TSSs of the genes. In the end, we obtain a matrix $\mathbf{X}$ containing the TF binding predictions, where $x_{i,j}$ corresponds to the affinity of TF j to the promoter of gene i.

2.4.2 Blood gene expression and HM data Affymetrix mRNA expression data were obtained from the Gene Expression Omnibus (GEO) database for HSCs, multipotent progenitors (MPPs), megakaryocyte/erythrocyte progenitors (PreMegE) and CD4+ T cells (TCD4) in *Mus musculus* (GEO accession number GSE18669). We use the MAS5 normalized data provided by the authors (Weishaupt *et al.*, 2010). Final expression values are computed by taking the median of replicates followed by a log transformation. In addition, for each of the above cell types, we also obtained the binding location of the histone marks H3K4me3, H3K79me2, H3ac, H3K9me3 and H3K27me3 (Weishaupt *et al.*, 2010). These HM data were measured with mini ChIP-chip experiments and are available from GEO (accession number GSE18734).

For computing the histone mark profiles, we use the MA2C program (Song *et al.*, 2007) using a window-size of 1000 bp at P-value cutoffs of 0.5 and the minimum number of probes required in the sliding window to be 5. The choice of a high P-value is based on the fact that we want continuous location measurements over the most possible genes regardless of the presence of a peak. Bound regions are annotated with the CEAS program (Shin *et al.*, 2009). All analysis are based on the genome version NCBI36/MM8. As 'regulatory signal', for each gene and histone mark, we compute the sum of the binding signals in the region $\pm$ 1000 bps around the respective TSS. HM values are incremented by a small value (0.0001) to avoid zeros and then log-transformed. For subsequent analysis, we excluded genes for which less than two histone were measured. Previous studies have shown that CpG rich promoters tend to be active in many tissues and contain few cell type-specific TF binding signals while CpG depleted promoters are often active in a cell type-specific manner and have cell type-specific TF binding sites close to their TSS (Roider *et al.*, 2009). We therefore restrict our analysis to those 4089 genes which have CpG-depleted promoters (normalized CpG content < 0.5). Finally, each binding signal is normalized to have mean equal 0 and standard deviation equal 1 across all genes.

2.4.3 ES gene expression, HM and TF data As a second test case, we obtained a gene expression dataset from murine embryonic stem (ES) cells (Ouyang *et al.*, 2009). In addition, we downloaded ChIP-Seq binding location data for the histone marks H3K4me3, H3K36me3, H3K9me3, H4K20me3 and H3K27me3 from (Mikkelsen *et al.*, 2007) as well as for the TFs E2f1, n-Myc, Zfx, c-Myc, Klf4, Tcfcp2l1, Esrrb, Nanog, Oct4, Sox2, Stat3 and Smad1 from Park and Nakai (2011) (GEO accessions GSE18734, GSE12241 and GSE11431). For normalization of the data, we applied the same pipeline as describe above.

2.5 Experimental design

We perform mixture estimation for modeling expression in four hematopoietic cell types (HSC, MPP, PreMegE and Tcd4) and one ES cell. We use in each scenario either TF affinities, HM and their combination (HM/TF). We vary the number of linear models from 1 to 10, the parameter λ_2 is set to 0.5 and λ_1 is varied within $(0, 0.01, 0.05, 0.1, 0.5, 1.0)$. The choice

of λ_2 represents the most parsimonious parametrization: L_1 and L_2 penalty are equal. The choice of λ_1, which is the main parameter controlling model sparsity, varies from no sparsity (0) to very high sparsity (>1).

We use the Bayesian Information Criterion (BIC) as defined in (Zou and Hastie, 2005) to indicate the best λ_1 parameterization. BIC is a model selection procedure that indicates which models present the best tradeoff between fit of the data and model complexity. A mixture with several models and using all regulatory signals can yield a good fit to the training data, but due to its complexity will most likely over-fit the data. To evaluate the performance of the best λ_1 parameterization as indicated by the BIC criteria, we perform a costly but more powerful 10-fold cross-validation procedure and measure the normalized mean squared regression error (NMSE).

$$\text{NMSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \quad (14)$$

Lastly, we use a statistical test proposed by Cule *et al.* (2011) to identify the regression coefficients which are significantly distinct from 0. This test takes into consideration the penalization imposed by the Elastic Net and sample size.

The method implementation is based on Pymix (Georgi *et al.*, 2010) and is freely available at `http://www.cin.ufpe.br/~igcf/SparseMix`.

3 RESULTS AND DISCUSSION

In the following, we apply our method for predicting gene expression using HM data, TF or HM data in combination with TF (HM/TF) binding predictions as input. The quality of the expression predictions made by our algorithm relies on identifying the appropriate number of regression models to be used as well as on the optimal model parameter λ_1 which determines how many regulatory signals will be used in the predictions. The larger the λ_1, the smaller is the number of regulatory signals used in the model.

3.1 Predicting gene expression in blood cells

We apply our regression model to four mouse blood cell types (Weishaupt *et al.*, 2010) for which both mRNA expression data as well as histone binding data are available (Fig. 1). To identify the optimal settings for our algorithm, we apply the model selection criteria BIC (see Section 2.5 for details).

BIC indicates that using two regression models and $\lambda_1 = 0.05$ for HM data, $\lambda_1 = 0.1$ for TF and $\lambda_1 = 0.01$ for HM/TF data are optimal to predict gene expression over all cell types. We base the further analysis on these model selections.

For predicting gene expression in HSC cells and HM data alone using two linear models, the algorithm selects four histone marks out of five available while for the highly dimensional scenario with TF data only 67 regulatory signals out of 599 were selected. For HM/TF data, 39 regulatory signals out of 604 are selected. As desired the feature selection is more stringent when providing also TF data. This indicates that HM signals are more predictive of gene expression and as expected not all TFs are relevant for the cell types analyzed.

Another interesting aspect is the robustness with which specific regulatory signals are selected by the mixture with different number of linear models. For HSC cells with HM and TF combined, the number of selected regulatory signals for 1 to 10 models are 4, 29, 51, 54, 55, 55, 55, 55, 55 and 55, respectively. We observe that more regulatory signals are retained when more linear models are added. Importantly, the signals selected by the simple mixtures are thereby retained in the more complex ones, demonstrating a high degree of robustness (see Supplementary Material for complete data).

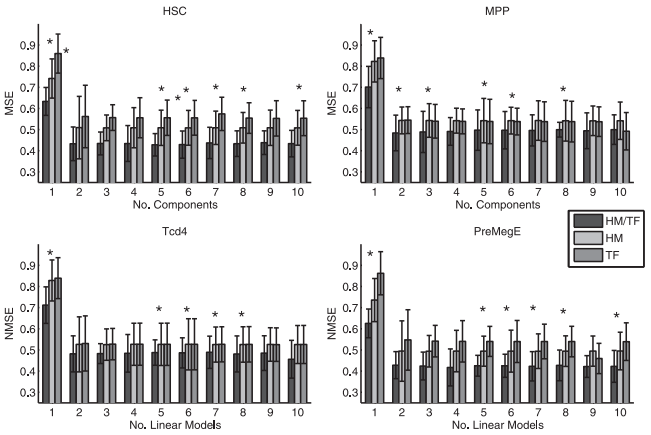

Fig. 2. NMSE of mixture models using HM, TF and HM/TF signals on HSC, MPP, Tcd4 and PreMegE cells. Asterisks indicate cases where HM/TF data have significantly smaller NMSE values than HM and TF alone (paired t-test P-value < 0.05)

To further evaluate the appropriate number of regression models, we perform a cross-validation procedure using the optimal values of λ_1 as indicated by BIC. As shown in Figure 2, compared with the single linear model NMSE improves significantly when two models are used (paired t-test P-value <0.01). However, no significant difference was found when comparing the NMSE from two linear models with that of three or more linear models.

Interestingly, the two linear models always separate the data into high and low expression genes on all blood cells. For instance, on HSC and HM data, the high expression module has a mean expression of 4.9 (1.6 std) over 790 genes and the low expression a mean of 2.5 (0.85 std) over 3297 genes (see supplement for a histogram with expression distributions). Similar results are found for all cell types or with TF data. Note that some genes belonging to the low expression models also display some level of expression and we cannot make assumptions about their activity.

A close look at the relation between the main regulatory signals controlling HSC expression reveals that mixture models improve the prediction by capturing non-linearities. As exemplified in Figure 3, in HSCs the repressive histone mark H3k27me3 is found on all low expression genes with similar frequency while for high expression genes, the amount of H3k27me3 is inversely proportional to expression level.

As shown in Figure 2, the use of combined HM/TF data results in the smallest NMSE in all cell types. A paired t-test (P-value <0.05) indicates a significant NMSE improvement for HM/TF data in several cases (see $*$ marks in Fig. 2). In Sections 3.3 and 3.4, we will look at predictions for individual histone marks and TFs based on the HM/TF data.

An alternative to use NMSE for evaluating the gene expression prediction is to compute Pearson correlation coefficients. For HM data on HSC, we obtain a correlation coefficient of 0.54 for one model and 0.75 for two models. This correlation coefficient is smaller than reported in Karlic *et al.* (2010), which was 0.72 for low CpG genes for a single linear model based on H3K4me3 and H3K79me2 histone markers on T CD4 cells. We stress however that the T CD4 data used in Karlic *et al.* (2010) were based on ChIP-Seq, which yield more precise histone location signals than the mini-ChIP-Chip protocol used in the blood cells.

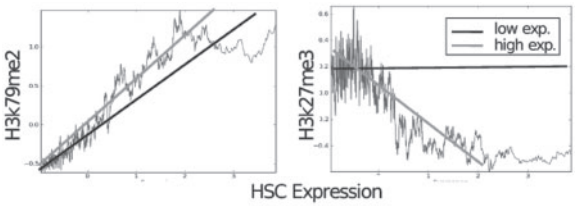

Fig. 3. Association between the presence of histone markers (H3k79me2 and H3k27me3) and gene expression in HSC cells. Lines indicate the regression models for the lowly and highly expressed genes

3.2 Predicting gene expression in ES cells

As a second and independent test case for validating the expression prediction power of the proposed method, we applied the mixture of sparse linear models to an embryonic stem ES cell dataset, which contains gene expression and ChIP-Seq measurements for five HMs and 12 TFs (Mikkelsen *et al.*, 2007; Ouyang *et al.*, 2009). For this dataset, BIC indicates that two models with $\lambda_1 = 0.01$ are optimal, while NMSE obtained after cross-validation indicates three models to be optimal. Pearson correlation coefficients for 1, 2 and 3 models are thereby 0.68, 0.76 and 0.89, respectively, demonstrating a clear improvement of the mixture models over a single linear regression model.

Similarly to what was obtained for blood cells, the mixture consists of a low expression model capturing 3312 genes and, in this case, two high expression models capturing the expression of 451 and 476 genes. This again suggests that there exist distinct modes of regulatory control for lowly and highly expressed genes (for detailed results see supplement).

3.3 Inferred role of histone marks in blood cells

In this section, we analyze the specific effects of the HMs H3K4me3, H3K79me2, H3ac, H3K9me3 and H3K27me3 on gene expression in blood cells as indicated by the optimal model obtained in Section 3.1 (two linear models and $\lambda_1 = 0.05$). The relevance of a particular modification can be estimated from the regression coefficients of the optimal linear model. That is, we apply a statistical test (Cule *et al.*, 2011) that indicates how significantly a particular coefficient deviates from zero for each of the regression models. In Figure 4, we display the significant coefficients of the HMs as obtained for the models for lowly and highly expressed genes across all blood cell types.

For highly expressed genes, H3k79me2 shows significant positive association with gene expression levels across all cell types while H3ac and H3K4me3 have significant positive association in MPP and T cells. For genes with low expression, H3K4me3 is positively associated with expression in MPP and T cells and H3k79me2 in HSC and PreMegE cells. These findings are mostly in accordance with previous works (Barski *et al.*, 2007; Ernst and Kellis, 2010; Weishaupt *et al.*, 2010), where these modifications are related to active promoters and expressed genes. Considering repressive marks, we observe that for genes with high expression both H3k9me3 and H3k27me3 have negative coefficients in HSC, MPP and T cells. For PreMegE, only H3k27me3 has a significant negative coefficient. For the group of lowly expressed genes, only H3k9me3 in MPP cell obtained a significant negative coefficient.

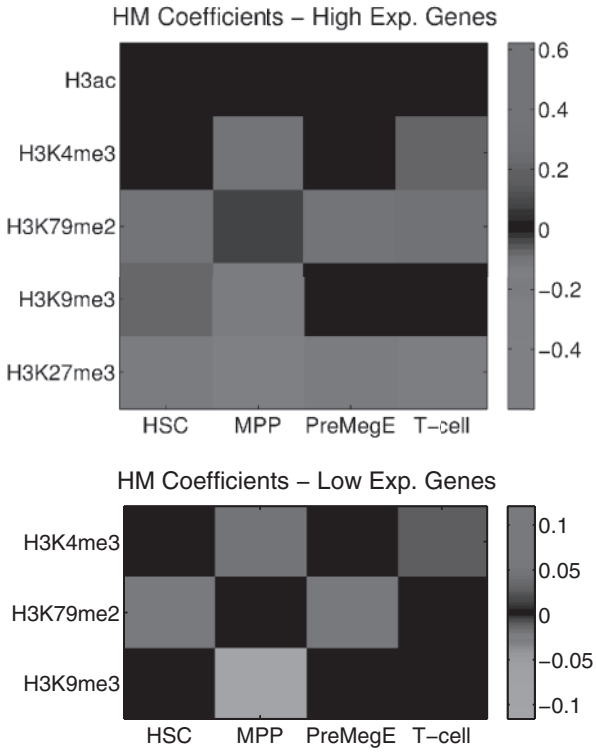

Fig. 4. Heatmaps of HM coefficients that significantly affect the expression of distinct groups of genes (high or low expression) and cell types (P-value <0.05). Red and green indicate positive and negative coefficients, respectively

Importantly, the selection of histone marks and their regulatory role indicated by the models is highly robust against changes in the complexity of the mixture model. The only exceptions are the positive correlation between H3ac and H3k79me2 with expression in MPP cells which was only detected with the mixture model as well as the association of H3k4me3 in HSC and PreMegE which was only detected with the single model. Therefore, the simple linear model and the mixture model appear largely equivalent with respect to correctly inferring the role of histone marks on expression, despite the clear advantage of the mixture model for predicting gene expression levels.

3.4 Inferred role of TFs in blood cells

In addition to the modifications of histones, we also observe the effects of TFs in the four blood cell types (two linear models, $\lambda_1 = 0.01$ and HM/TF data). Out of the 600 TFs, 31 had a statistical significant coefficient (P-value <0.05) in at least one condition (see Fig. 5 for the list of TFs). Out of the 31 TFs, 15 were related to development on hematopoietic system (E2F1, Foxj2, Nf-E2, Nrf2, Cp2, Foxd3, Hif1, Evi1, Mzf1, Gzf1, Gfi1, Arp1, Hox9-Meis1B, Pax1 and Lrf), 5 with chromatin structure remodeling (Fac1, Hsf, Kaiso, Sp1 and Egr-1), 7 embryonic development (Pax6, Axl4, Pbx, Pax4, Rbpjk, Rfx1 and ChCh) and only 3 had no relation to hematopoiesis (Cap1, Lxr and Cdx-1). See supplement for an detailed discussion of these factors. This indicates a clear enrichment of recovering TFs related to chromatin reorganization,

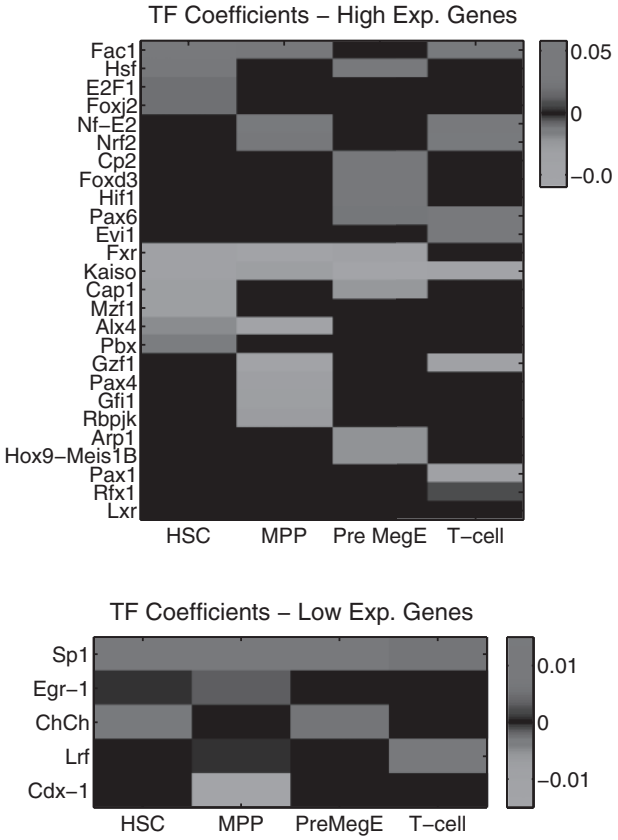

Fig. 5. Heatmaps of TF coefficients that significantly affect the expression of highly expressed genes (top) and lowly expressed genes (bottom) in different cell types (P-value < 0.05)

hematopoiesis and development. Moreover, the results with a single linear model had none TF with a significant regression coefficient value. These indicates the power of the mixture of sparse linear models in recovering interesting TF candidates from hundreds of candidates.

We concentrate our discussion on TFs with potential chromatin remodeling function. Of the selected TFs, 11 are involved in the activation process of expression in genes associated with high expression (Fig. 5 top). The TF FAC1—also known as Bptf—is present in the process of gene activation in HSC, MPP and T-cell. This gene is a component of the NURF complex, which is known to promote trimethylation of the H3 lysine 4 and gene activation in mammals (Wysocka *et al.*, 2006). Another protein that has been related to immune cell development is HSF (Morange, 2006), which acts as a activator of expression in HSC and PreMegE cells. This protein has been recently implicated with histone acetylation and gene activation in mammals (Fritah *et al.*, 2009). On the other hand, 15 TFs were involved in the repression of expression in genes associated with high expression (Fig. 5 top). Of those, Kaiso is a known chromatin remodeling factor. Kaiso is known to bind to methylated DNA and the recruitment of H3 lysine 9 methylation and gene repression (Yoon *et al.*, 2003).

Regarding the five TFs related to group of low expressed genes, the method detects two chromatin-related proteins Sp1 and Egr-1 (see Fig. 5 bottom). Sp1, which is indicated to be active in all cell types, has a known role in chromatin modeling by interacting with HDAC enzymes or p300 for either repressing or promoting gene expression (Doetzlhofer *et al.*, 1999; Sun *et al.*, 2006). Another gene with a putative chromatin remodeling role is Egr-1 (Krox). This gene was detected as an activator of gene expression in HSC and MPP cells. It interacts with EP300 and CBP, which are known to promote histone acetylation and activation of expression (Silverman *et al.*, 1998). These examples demonstrate the recovery of a TFs with chromatin remodeling roles and the accurate prediction of their functional role (repression or activation). Note that most of these predictions were not previously characterized in hematopoiesis.

4 CONCLUSION

We have developed a novel method for predicting gene expression that combines the use of linear mixture models with an efficient way to select the relevant predictor variables from a large set of regulatory signals. The approach is ideally suited to integrate high-dimensional data normally not applicable in standard linear regression analysis. For instance, when provided the full set of binding affinities from vertebrate TRANSFAC matrices together with HM data to model gene expression in blood cells, the algorithm performed a sparse selection retaining only 29 out of over 600 input variables. The resulting sparse linear mixture models not only significantly improved gene expression predictions by capturing non-linear relations with the retained regulatory signals but also allowed to readily identify and characterize the relevant regulatory signals. In particular, the model predicted known roles of HMs and could select a small set of TFs related to development and hematopoiesis to particular developmental stages.

The proposed method has three main parameters to be optimized (K, λ_1 and λ_2) and at the moment cannot make use of the regularization paths (Friedman *et al.*, 2009) together with the EM algorithm. Further work will thus be required to develop efficient methods for optimizing the parameters. In addition, TF binding affinities can accurately be predicted only for proximal promoters and well-conserved upstream elements. The addition of epigenetic information, which can further characterize distal enhancers and active promoters (Ernst and Kellis, 2010), is likely to improve the TF binding predictions and is current work in progress.

Large consortia such as the Epigenetic Roadmap, ENCODE and Blueprint Epigenome are releasing expression, epigenetic and binding data on an unprecedented scale. Although our approach offers a new way to integration such large-scale datasets it remains a big challenge to further unravel the regulatory mechanisms underlying the developmental processes shaping the human body.

Funding: Brazilian funding agencies dont require more information than that. This figure could also be made wider.

Conflict of interest: none declared.

REFERENCES

Barreda,D.R. and Belosevic,M. (2001) Transcriptional regulation of hemopoiesis. *Dev. Comp. Immunol.*, **25**, 763–789.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **5**, 823–837.

Bussemaker,H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Cheng,C. *et al.* (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.*, **12**, R15+.

Costa,I. *et al.* (2011) Predicting gene expression in t cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics*, **12(Suppl. 1)**, S29.

Cule,E. *et al.* (2011) Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, **12**, 372.

Dempster,A. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.

DeSarbo,W. and Cron,W. (1988) A maximum likelihood methodology for clusterwise linear regression. *J. Classif.*, **5**, 249–282.

Doetzlhofer,A. *et al.* (1999) Histone deacetylase 1 can repress transcription by binding to sp1. *Mol. and Cell. Biol.*, **19**, 5504–5511.

Dor,L.C. and Crispino,J.D. (2011) Transcription factor networks in erythroid cell and megakaryocyte development. *Blood*, **118**, 231–239.

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **8**, 817–825.

Friedman,J. *et al.* (2009) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**.

Fritah,S. *et al.* (2009) Heat-shock factor 1 controls genome-wide acetylation in heat-shocked cells. *Mol. Biol. Cell*, **20**, 4976–4984.

Georgi,B. *et al.* (2010) Pymix—the python mixture package—a tool for clustering of heterogeneous biological data. *BMC Bioinformatics*, **11**, 9.

Goldberg,A.D. *et al.* (2007) Epigenetics: a landscape takes shape. *Cell*, **128**, 635–638.

Hastie,T. *et al.* (2003) *The Elements of Statistical Learning*. Springer, Berlin, corrected edition.

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

Keles,S. *et al.* (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

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

Li,Q. and Lin,N. (2010) The bayesian elastic net. *Bayesian Anal.*, **5**, 151–170.

Matthias,P. and Rolink,A.G. (2005) Transcriptional networks in developing and mature b cells. *Nat. Rev. Immunol.*, **5**, 497–508.

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

McLachlan,G.J. and Peel,D. (2000) *Finite Mixture Models.* Wiley Series in Probability and Statistics. Wiley, New York.

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

Morange,M. (2006) Hsfs in development. *Handb. Exp. Pharmacol.*, **172**, 153–169.

Novershtern,N. *et al.* (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.

Orkin,S.H. and Zon,L.I. (2008) Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, **132**, 631–644.

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

Park,S.-J.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.

Roider,H.G. *et al.* (2007) Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, **23**, 134–141.

Roider,H.G. *et al.* (2009) Cpg-depleted promoters harbor tissue-specific transcription factor binding signals–implications for motif overrepresentation analyses. *Nucleic Acids Res.*, **37**, 6305–6315.

Rothenberg,E.V. and Taghon,T. (2005) Molecular genetics of T cell development. *Annu. Rev. Immunol.*, **23**, 601–649.

Shin,H. *et al.* (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics*, **25**, 2605–2606.

Silverman,E. *et al.* (1998) camp-response-element-binding-protein-binding protein (cbp) and p300 are transcriptional co-activators of early growth response factor-1 (egr-1). *Biochem. J.*, **336**, 183–189.

Song,J. *et al.* (2007) Model-based analysis of two-color arrays (ma2c). *Genome Biol.*, **8**, R178.

Sun,H.-J. *et al.* (2006) Transcription factors ets2 and sp1 act synergistically with histone acetyltransferase p300 in activating human interleukin-12 p40 promoter. *Acta Biochimica et Biophysica Sinica*, **38**, 194–200.

Sun,T. and Zhang,C.-H. (2010) Comments on: ℓ_1-penalization for mixture regression models. *TEST Official J. Spanish Soc. Stat. Oper. Res.*, **19**, 270–275.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. (Ser. B)*, **58**, 267–288.

Turner,B.M. (2007) Defining an epigenetic code. *Nat. Cell Biol.*, **9**, 2–6.

Weishaupt,H. *et al.* (2010) Epigenetic chromatin states uniquely define the developmental plasticity of murine hematopoietic stem cells. *Blood*, **2**, 247–256.

Wysocka,J. *et al.* (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, **442**, 86–90.

Yoon,H.-G. *et al.* (2003) N-cor mediates DNA methylation-dependent repression through a methyl cpg binding protein kaiso. *Mol. Cell*, **12**, 723–734.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.