Costa et al. BMC Bioinformatics 2011, 12(Suppl 1):S29
http://www.biomedcentral.com/1471-2105/12/S1/S29

BMC Bioinformatics

RESEARCH Open Access

Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models

Ivan G Costa[1*], Helge G Roider[2], Thais G do Rego[1], Francisco de AT de Carvalho[1]

From The Ninth Asia Pacific Bioinformatics Conference (APBC 2011)
Inchon, Korea. 11-14 January 2011

Abstract

Background: The differentiation process from stem cells to fully differentiated cell types is controlled by the interplay of chromatin modifications and transcription factor activity. Histone modifications or transcription factors frequently act in a multi-functional manner, with a given DNA motif or histone modification conveying both transcriptional repression and activation depending on its location in the promoter and other regulatory signals surrounding it.

Results: To account for the possible multi functionality of regulatory signals, we model the observed gene expression patterns by a mixture of linear regression models. We apply the approach to identify the underlying histone modifications and transcription factors guiding gene expression of differentiated CD4+ T cells. The method improves the gene expression prediction in relation to the use of a single linear model, as often used by previous approaches. Moreover, it recovered the known role of the modifications H3K4me3 and H3K27me3 in activating cell specific genes and of some transcription factors related to CD4+ T differentiation.

Background

All cells in a multi-cellular organism arise from the same zygote and thus carry the same genetic information. However, complex regulatory programs allow stem cells to differentiate into distinct cell types. For instance, in response to different infectious agents Naive CD4+ T cells differentiate into at least four types of T helper cells—Th1, Th2, Th17, and inducible regulatory T cells (iTregs) [1]. While all of these cell types are involved in the adaptive immune response they serve distinct roles by secreting different cytokines. For example, Th1 acts against mycobacterial infections by releasing IFNγ, which activates the response of macrophages [1] while Th2 cells secrete various interleukins helping B-cells to induce humoral immunity.

On the transcriptional level, the differentiation process from stem cells to fully differentiated cell types is controlled by the interplay of chromatin modifications and transcription factor activity [2]. Chromatin structure is shaped primarily by histones. The presence or absence of these large globular protein complexes determines the accessibility of the promoter regions for the transcriptional machinery and thus performs a high-level control on gene expression [3,4]. The affinity of histones to DNA is modified by the cell via a large repertoire of post-translational protein modifications including acetylations and methylations.

The resulting epigenetic histone code appears highly intricate, with a given histone frequently carrying several different modifications at a time. Despite this complexity, it has become clear that certain modifications, such as the trimethylation of the lysine 4 residue in the tail of histone H3 (abbreviated H3K4me3) are mainly associated with

* Correspondence: igcf@cin.ufpe.br
[1]Center of Informatics, Federal University of Pernambuco, Recife, Brazil
Full list of author information is available at the end of the article

active promoters while other modifications such as H3K27me3 tend to be associated with inactive promoters [5]. The importance of histone modifications for the differentiation of Naive CD4 T-cells into Th1 cells has recently been verified at [6], which demonstrated that IFNγ expression is controlled by the histone methylation status of its promoter.

Aside from chromatin structure, transcription factors (TFs), play an essential role in controlling cell differentiation by guiding the transcriptional machinery to its target promoters and facilitating the initiation of transcription. For instance, in T-cell differentiation, in vitro studies demonstrated that either high levels of the transcription factor GATA3 or strong signalling via the transcription factor STAT5 is sufficient to determine the Th2 cell fate [1].

Particularly in the context of genome wide studies, computational biology analysis have become an essential component of elucidating the regulatory signals underlying observed gene expression patterns. Usually, the problem of identifying the promoter elements guiding differentiation and cell type specific gene expression is tackled by first selecting the genes which are most specifically expressed in the particular cell type and then performing motif over-representation analysis on their promoter sequences as in [7,8] (see [9] for a recent review). While such methods allow identifying potentially regulating transcription factors they have the intrinsic drawback of requiring a previous grouping of genes and of being able to explain only the expression of the genes with highest specificity for the condition.

In contrast, linear regression models, as first proposed by [10,11], combine all regulatory signals in order to explain the expression pattern of the genes. In their work, Bussemaker *et* al. [10] focused on explaining gene expression based on combinations of predicted TF binding sites. The coefficients of the linear model indicate the importance of a particular regulatory signal. That is, signals which obtain large positive coefficients likely correspond to putative activators while signals with large negative coefficients likely act as suppressors. Recently, Karlic *et* al. [12] also used a linear regression model in order to estimate promoter activity based on histone modification data. By design, the above approaches assume that a given regulatory signal exert the same regulatory effect on all its target genes.

However, transcription factors and thus their DNA binding motifs frequently act in a bi-functional manner, with a given DNA motif conveying both transcriptional repression and activation depending on its location with respect to the transcription start site (TSS) and the sequence motifs surrounding it. For instance, RUNX1 and RUNX3 have been shown to act both as repressors and activators in different tissues and are involved in determining T-cell fate [13].

To account for the possible multi functionality of regulatory signals, in this study, we propose to extend [10-12] by allowing the observed gene expression patterns to be explained by not just one, but by a mixture of several linear regression models [14,15]. This permits for instance to find mixture models, such that genes with high maximal expression are controlled by a different group of regulatory signals than genes with low maximal expression (see Fig. 1). That is, a regulatory signal might act as repressor when associated with lowly expressed genes while it may function as activator or neutral bystander when present in the promoter of highly expressed genes.

In order to find the regression models best explaining the expression data, our method takes as input the matrix Y of observed expression profiles from all genes as well as a matrix X, containing the regulatory signals for the corresponding promoters (i.e. predicted TF binding affinities and presence of histone modifications). For each gene, it then estimates the coefficient vector B, representing the relative importance of each regulatory signal and its effect on gene expression (activation or repression).

We apply this novel approach to identify the underlying regulatory signals guiding gene expression in each of the four differentiated CD4+ T-helper cell types. As potential regulatory signals we consider both, histone modifications (HM) as measured by Chip-Seq [16] as well as predicted binding affinities [17] from a set of TFs related to lymphoiesis [1,18-20]. As we are mainly interested in cell type specific signals, we restrict the analysis to genes with low CpG content in their promoters [21] as such genes tend to be expressed in a tissue and stage specific manner while genes with high CpG promoter content tend to be broadly expressed. Using this method we expect to improve the gene expression prediction in relation to the use of single linear model, but also to reveal the regulatory roles for histone modifications and transcription factors.

Results and discussion

Regulatory signals predicts expression

As a first step, we want to determine which set of regulatory signals, X, can explain the observed gene expression data, Y , best. To this end, as a first step we supply our algorithm with a matrix X containing only predicted TF binding affinities, only histone modification data or both sets of regulatory signals and assess how well the resulting regression models can capture the data. As measure of quality for the different models we thereby compute the mean square error between the predicted

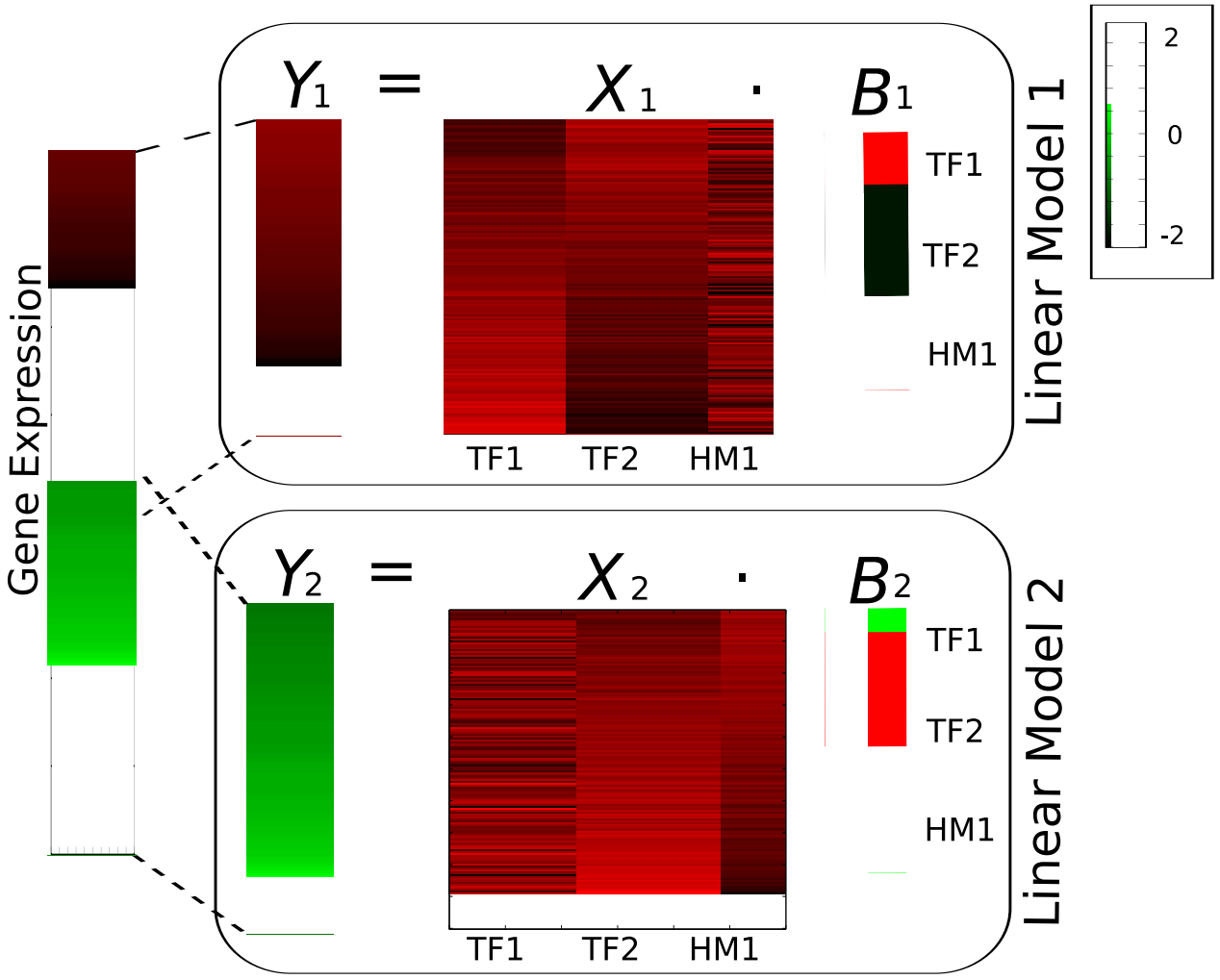

Figure 1 Example of Linear Regression Mixture Model. Illustrative example of the use of linear regression mixture models to predict gene expression from the regulatory signals. Gene expression profiles, indicated by the heat map on the left (red and green correspond to highly and lowly expressed genes, respectively), are best predicted by two different models: model 1 for genes with high to medium expression and model 2 for genes with medium to low expression. At the level of each linear model k, the gene expression vector Y_k is predicted by the multiplication of the matrix X_k (bright red values indicate higher TF binding strength or HM presence) with the vector of model coefficients B_k (red, black and green values indicate positive, near zero and negative regression coefficients, respectively). In this example, B_1 indicates that genes with high to medium expression are repressed by TF1, positively regulated by TF2 and are unrelated to the presence of HM1. In contrast, B_2 indicates that genes with medium to low expression are repressed by TF2, positively regulated by HM1, and are unrelated to the binding affinity of TF1. If we estimate a single linear model from the same data, TF2 would have a regression coefficient close to zero as the positive coefficient from model 2 and the negative coefficient from model 1 would cancel each other out.

gene expression values and the actual measurements (see Methods for details).

Predicting the gene expression from the four T-cell types based on only histone modification data and by means of only a single regression model yields MSEs of about 0.5 for HM and HM+TF on all data sets (see red bars in Fig. 2). A mixture of two regression models further reduces the MSEs to an average value 0.25 across all cell types. In all scenarios, the difference of MSE between one and two models were statistically relevant (t-test p-value < 0.01) indicating the advantage of using mixtures to predict expression. The model selection procedure (see Methods) indicates that the data is optimally explained by the combination of 2-4 regression models (see Fig. 2) and that gene expression data can be well predicted based on histone modification data alone.

In contrast, using a single regression model to predict gene expression data based on TF binding affinities alone yields considerably larger MSEs across all cell types (average MSE = 2, see blue bars in Fig. 2). Interestingly, supplying our algorithm with the combined

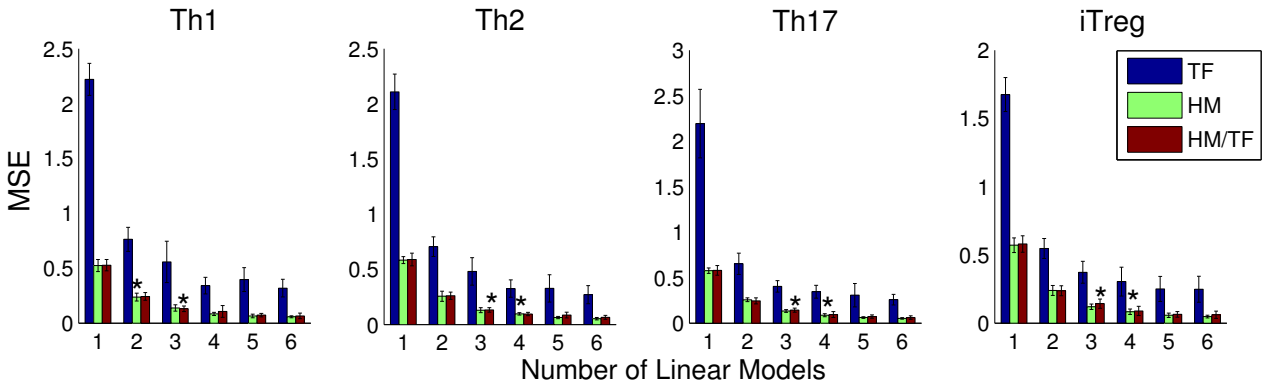

Figure 2 Regression Prediction Error. We depict the MSE error for 1 to 6 models for the prediction of expression on Th1, Th2, Th17 and iTreg. Bars marked with * indicate number of linear models indicated by the model selection. The MSE with TF is higher than the use of either HM/TF or HM on all combinations of expression data and number of models. For all combination of expression data and number of models, the was no significant difference between the MSE from HM or HM/TF.

data from both histone modifications and TF binding affinities yields MSEs similar to the ones obtained with only histone modification data alone (see Fig. 2). This indicates that the utilized histone modification and TF binding data cover rather redundant than complementary information about gene expression. As histone modifications and HM+TF affinities yield the solutions with the lowest MSEs, in the following we will continue the analysis with the models based on these data sets.

Control of Th1 gene expression

Having established that histone modification data together with a mixture of two regression models yields the most significant results we now investigate which type of modifications contribute most strongly to these models. We thereby restrict our analysis to the data from Th1 cells and the corresponding regulatory signals which obtain the largest absolute regression coefficients (results from other cell types closely resemble those from Th1 cells, see Additional File 1 for details).

For model 1, which explains the expression pattern of the most highly as well as moderately expressed genes, the histone modifications with largest influence are H3K4me3 and H3K27me3, with regression coefficients of +0.7975 and -0.4533, respectively. As shown in the top part of Fig. 3 these two modification form a gradient with H3K4me3 being most frequently found in the highly expressed genes while being absent in moderately to lowly expressed genes. In contrast, H3K27me3 is consistently detected in promoters of lowly expressed genes but appears weaker or even absent in promoters of highly expressed genes.

For model 2, which explains the transcriptional activity of a small subset of highly expressed as well as most of the lowly expressed genes, we again find H3K4me3 to have the strongest positive regression coefficient (b = 0.71). This is reflected by a strong association of this modification with the most highly expressed genes of this set (see Fig. 3). In contrast, H3K27me3 obtains a regression coefficient of close to zero (b = 0.03) in this model as this modification appears with the same intensity in nearly all genes assigned to model 2.

An alternative view of this results is presented at Fig. 4. There, we have the interpolated values of the histone modifications against the gene expression, the linear models for each component and the resulting mixture model. Clearly, dependence of H3K27me3 on gene expression is not linear, as low expressed genes all present a high presence of histone modifications. This non-linearity is captured by the mixture model (red line), and explains the lowest MSE errors obtained when more than one linear models is applied. To see whether the influence of TFs may contribute to this effect we next look in detail at the results obtained from combined histone and TF data together with a mixture of three correlation models. As shown in Fig. 5), we see similar results in respect to the histone markers: H3K4me3 as enhancer and H3K27me3 as inhibitor of expression for genes with high expression and H3K4me3 as enhancer for genes with low expression. Moreover, only for the genes with high expression, there are some TFs (Pax5, Stat5, Meis/Hox, Iscbp) promoting gene expression and a TF (MyB) inhibiting expression. For all TFs, regression coefficients were in the range of 0.1 to 0.15 (see Additional File 1 for additional results). For genes with low expression, we found no relation between the TFs binding affinities and expression. Th1 cells are known to be regulated by T-bet and Stat4 [1]. While our study lacked the PFM of T-bet, it listed the closely related Stat5 as a positive regulator of the genes

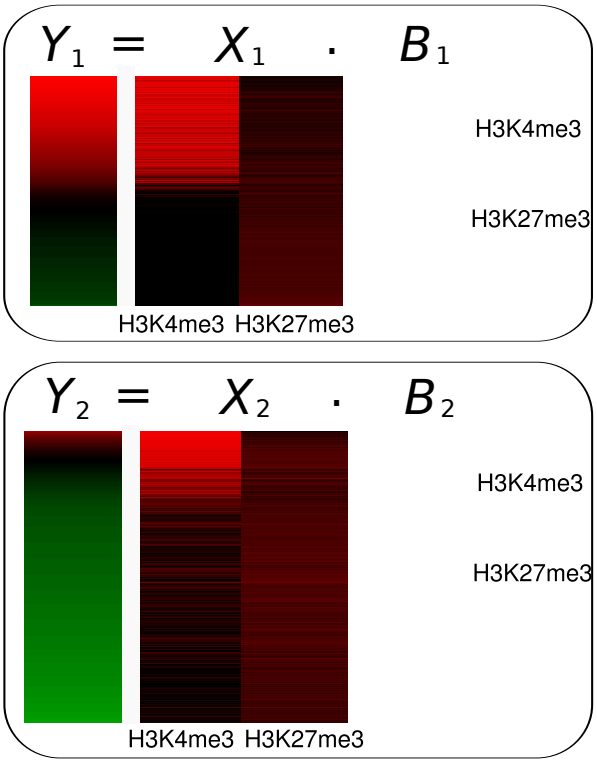

Figure 3 Regression Results Th1. Results for mixtures of two regression models utilizing only histone modification data on Th1. Model 1 captures the expression of 2231 genes and model 2 of 3923 genes. Corresponding gene expression levels are shown by the vectors Y on the left (red and green indicate high and low expression, respectively). Association strengths of the regulatory signals for the corresponding promotes are indicated by matrices X (red and black indicate high and low association, respectively). Finally, the magnitude of the corresponding regression coefficients for each of the regulatory signals in each of the models is indicated by vector B (bright red for large positive and bright green for large negative values). In model 1, the amount of H3K4me3 correlates positively while the level of H3K27me3 correlates negatively with gene expression. In model 2, only H3K4me3 levels correlate with gene expression as H3K27me3 levels remain constant over the whole expression range.

with high expression. In relation to factor related to inhibition, there has been a recent implication of c-MyB to bind to H3 histone tails and to promote histone acetylations in Humans [22]. These results indicate a putative role of the MyB in down-regulating the expression of genes during CD4+ T differentiation by promotion of epigenetic changes. However, further acytilation modification data would be required for a better characterizaion of the role of this factor.

Comparison with previous studies

Several computational biology methods have been previously proposed for the use of linear models for predicting gene expression in the context transcription factor binding [10,11] or histone modifications [12]. In all cases, distinct datasets were used and results are not directly comparable. In relation to [12], the analysis were based on Human naive CD4+ T cells and included 38 histone modifications. Their predicted model obtained a correlation coefficient of 0.72 on genes with low CpG content with HM H3K4me3 and H3K79me1, while our method had a coefficient at the range 0.64 – 0.68 for one model and 0.85 – 0.87 for two models for H3K4me3 and H3K27me3 data. The increase of the correlation coefficient from single linear models to two linear models is an indication that all these approaches would profit from the use of the mixture of linear regressions framework.

Conclusion

Predicting gene expression from regulatory signals is an important but unmet goal in bioinformatics. In this study, we propose a novel approach which uses mixtures of linear regression models together with transcription factor binding and histone modification data for estimating transcriptional activity of CpG depleted promoters. In addition the approach allows to determine the functional activity of the various regulatory signals. We show that our approach obtains significantly smaller errors in predicting the expression of genes in comparison to simple linear regression models as used in previous approaches. For gene expression data from CD4+ T helper cells we find that both, histone modification data alone and histone modifications together with predicted TF binding affinities, yields the best expression predictors. In accordance with previous dedicated studies we recover the well known regulatory roles of H3K4me3 as an enhancing and H3K27me3 as a repressive regulatory signal for gene expression. Moreover, our predictions suggest that histone modifications act not in a binary on/off fashion but rather in a continuous way with levels of H3K4me3 and H3K27me3 steadily rising or falling over a large range of expression values in a non-linear way. With the use of TF binding affinities, we also partially recover the main factors such as the Stat family involved in T helper cell type specific gene expression. Interestingly, we observe a negative effect of cMyb on expression in all T helper cell types. This raises the question whether MyB, which has been recently showed to promote histone acetylation marks in hematopoiesis [22], could play a role in the down-regulation of genes in T helper cells types.

The advent of next generation sequencing provides an ever growing stock of high quality data for the full range of histone modifications, DNA methylation state and transcription factor occupancy across the entire genome from various cell types and differentiation stages. Several methodological improvements will be required to

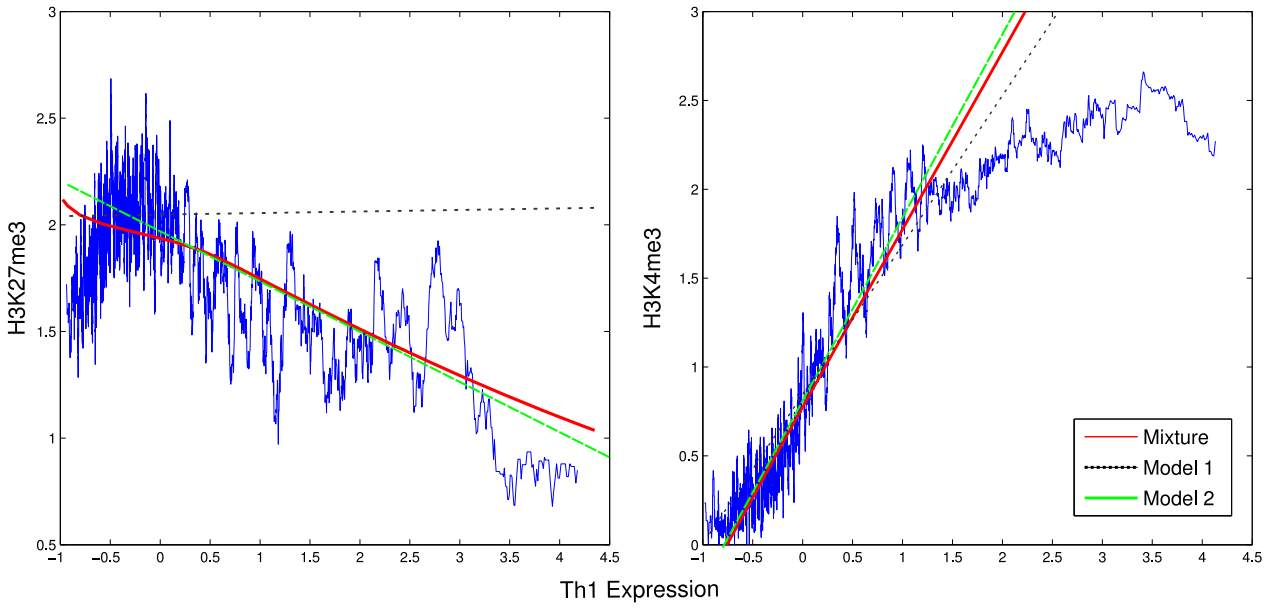

Figure 4 Histone Modification against Th1 Gene Expression. We depict the values of Th1 gene expression against H3K27me3 modification (left) and H3K4me3 modification (right). The blue line represents a nearest-neighbor interpolation (30 samples) of the histone modification signal, the dashed black line represents the linear model for the 1 component, the green line for the 2 component and the red curve represents the mean regression value of the mixture of linear regression with 2 components. The interpolation indicates that the H3K4me3 is close to linear, but not for H3K27me3 modification, which is negatively related to high expressed genes, but has little effect on low expressed genes.

integrate this wealth of data in order to shed light on the complex interplay between the different regulatory signals acting in eukaryotics. Moreover, in an ideal case where all possible regulatory signals have been measured, advanced feature selection procedures such as postulated by [23], will be vital for the detection of all the players involved in determining gene expression.

Methods

Mixture of linear regressions

In the following we want to model the observed expression level of all N genes, using different linear combinations of the M different regulatory signals associated with the promoters (i.e. binding affinities for various TFs and different histone modifications). To this end let y_i be the gene expression level of gene i (the dependent variable) and x_i be a corresponding vector of M regulatory signals (the regressor variables). The single linear regression model is then defined as

$$y_i = b_0 + x_i B^{\mathbf{T}} + \epsilon_i, \quad (1)$$

where B is a vector $(b_1, \ldots, b_M)$ representing regression coefficients and ϵ_i is an error term. For mathematical convenience, we redefine the vector with the regressor variables to be $x_i = (1, x_{i1}, \ldots, x_{iM})$ and include the bias parameter b_0 in the beginning of B, that is $B = (b_0, b_1, \ldots, b_M)$. Assuming the error ϵ follows a Normal distribution with standard deviation σ^2, the linear regression model has the following distribution

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

A mixture of linear regression models is defined as a convex summation of K distributions

$$\mathbb{P}(y_i \mid x_i, \Theta) = \sum_{k=1}^{K} \pi_k \cdot \mathbf{N}(y_i \mid x_i B_k^T, \sigma_k^2) \quad (3)$$

where $\Pi = (\pi_1, \ldots, \pi_K)$ are the mixture coefficients, which respect $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and Θ are the model parameters $(\Pi, B_1, \ldots, B_K, \sigma_1^2, \ldots, \sigma_K^2)$.

For a given data X and Y, where X is a set on N observations x_i and Y a vector with N observations y_i, the mixture of linear regression models can be estimated with the Expectation-Maximization algorithm [14,24]. We resort to Maximum-a-posteriori (MAP) estimates of the parameters, as described in the next section, to avoid over-fitting [25]. The EM works by finding estimates Θ maximizing the posterior distribution over the data X and Y

$$\mathbb{P}(\Theta | X, Y, Z) \approx \mathbb{P}(Y, Z | X, \Theta) \mathbb{P}(\Theta) \quad (4)$$

where Z is the vector of hidden variables with $z_i \in \{1, \ldots, K\}$ indicating which linear model an observation i

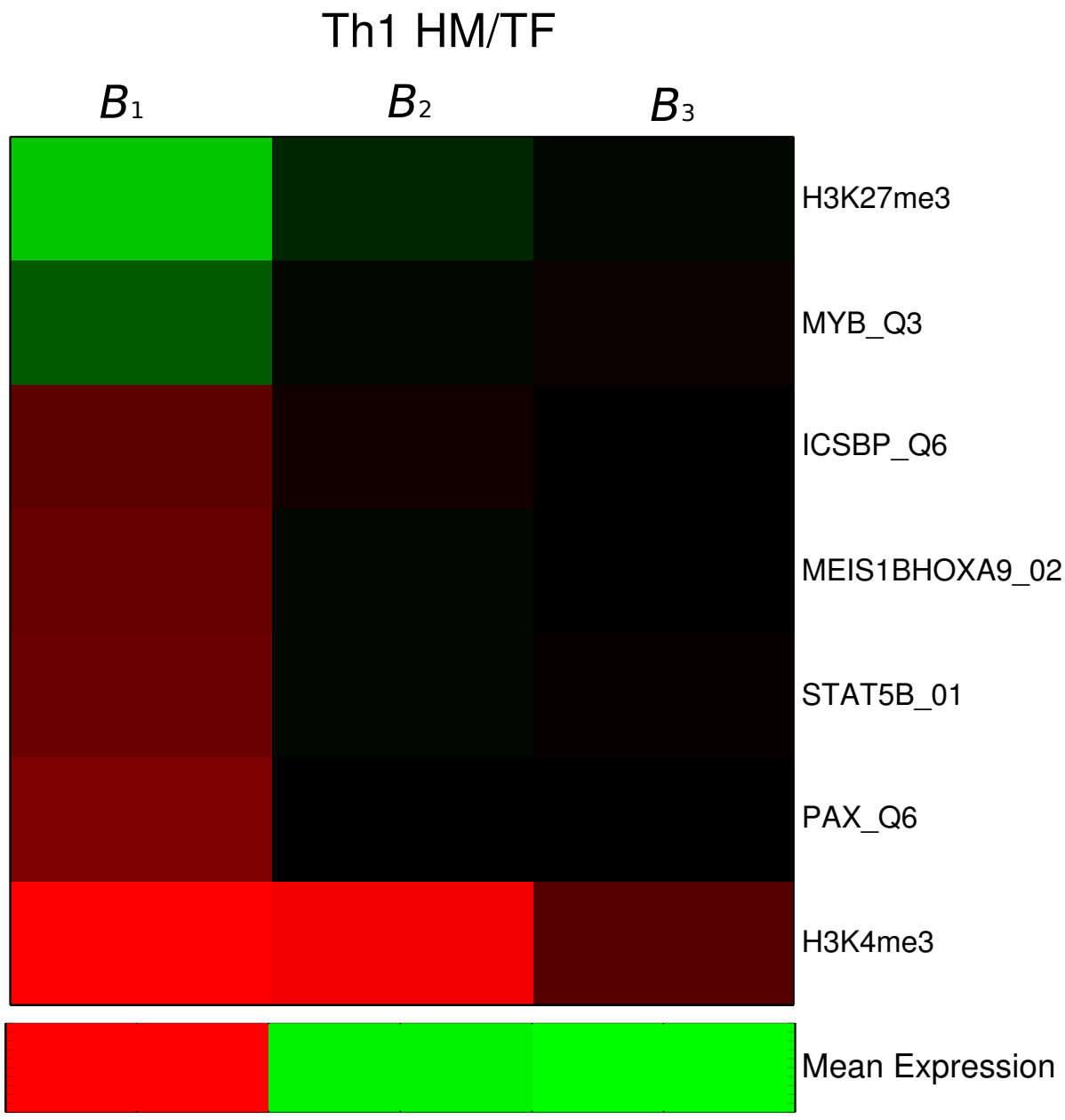

Figure 5 Regression Coefficients on Th1 with HM/TF data. We depict the regression coefficients of the most relevant regulatory signals and mean expression values for the mixture with three linear models on Th1 with HM/TF data.

belongs to. $\mathbb{P}(\Theta)$ is the prior distribution over the model parameters (see next section for the definition of the prior). $\mathbb{P}(Y, Z|X, \Theta)$ is the complete data likelihood and is given by:

$$\mathbb{P}(Y, Z \mid X, \Theta) = \prod_{k=1}^{K} \mathbb{P}(\Theta) \prod_{i=1}^{N} \left(\pi_k \cdot \boldsymbol{N}\left(y_i \mid x_i B_k^T, \sigma_k^2 \right) \right)^{r_{ik}} \quad (5)$$

where r_{ik} is the posterior probability (or responsibility) [25] that observation i belongs to the linear model k and is given by:

$$r_{ik} = \mathbb{P}(z_i = k \mid y_i, x_i) = \frac{\pi_k \boldsymbol{N}\left(y_i \mid x_i B_k, \sigma_k^2 \right)}{\sum_{k'=1}^{K} \pi_{k'} \cdot \boldsymbol{N}\left(y_i \mid x_i B_{k'}, \sigma_{k'}^2 \right)}. \quad (6)$$

For further details on mixture models we refer the reader to [25].

The EM algorithm works by iteratively estimating the model assignments (r_{ik}) and the model parameters Θ until some convergence criteria is reached. In the context of the mixture of linear regression models, we need estimates of the linear regression parameters (B_k, σ_K^2)

for a particular model k, and all other parameters (r_{ik},Π) follow the usual EM algorithm [25].

Once the mixture model is estimated, the predicted value $\hat{y}_i$ for a particular regressor observation x_i is given by

$$\hat{\gamma}_i = \sum_{k=1}^{K} \mathbb{P}\left(z_i = k \mid x_i\right) \cdot x_i B_k^T. \quad (7)$$

That is, the linear regression prediction is a mixture of the predictions of each individual component times the posterior probability of the observation i to belong to the model k. In our particular application problem, we are interested in estimating the models which corresponds to an unsupervised learning problem, that is, the coefficients indicating whether a regulatory signal plays an important repressive or activating role. The predictions $\hat{y}$ can thereby be used for evaluating the fit of our model. In cases where one wants estimate the expression level of genes, that is, estimation of $\hat{y}$ (supervised learning problem), the above equation should not be used, as the posterior probabilities are based on the response variable y, which is usually unknown in a predictive scenario. In such a context, methods for combinations of predictors, such as [26], are required.

Bayesian linear regression estimates

We resort to Bayesian approach for obtaining MAP estimates of the linear regression models as proposed in [27]. Therefore, we avoid problems related to overfitting which usually occur with the EM algorithm and mixture models [25]. More formally, the prior distribution in Eq. 4 can be decomposed as

$$\mathbb{P}\left(\Theta\right) = \mathbb{P}\left(\Pi\right) \prod_{k=1}^{K} \mathbb{P}\left(B_k\right). \quad (8)$$

We use the following conjugate prior for the regression coefficient B_k

$$\mathbb{P}(B_k) = \boldsymbol{N}(B_k|0, \beta_k\mathbf{I}), \quad (9)$$

where 0 is a vector with M zeros, $\mathbf{I}$ is a M x M identity matrix and β_k is the hyper-parameter.

Let r_k be an N dimensional vector (r_{1k}, ...,r_{Nk}) containing the posterior probabilities of the observations belonging to model k and let W_k = diag(r_k), then the estimates from model k maximizing Eq. 4 are defined as

$$B_k = \frac{\frac{X^T\left(W_k Y\right)}{\sigma_k^2}}{\frac{I}{\beta_k} + \frac{X^T\left(W_k X\right)}{\sigma_k^2}} \quad (10)$$

with

$$\sigma_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} r_{ik} \left(\gamma_i - x_i B_k^T\right)^2. \quad (11)$$

From Eq. 10, we can see that β_k works by shrinking the regression coefficients. Small β_k imposes a higher shrinkage on the regression coefficients. Furthermore, for $\beta_k \to \infty$ we have a non-informative prior and the regression coefficients are the maximum likelihood estimates.

We estimate the hyper-parameter β_k in an Empirical Bayes approach with

$$\beta_k = \frac{B_k^T B_k}{\gamma_k}, \quad (12)$$

where

$$\gamma_k = \sum_{j=1}^{M} \frac{\lambda_j}{\beta_k^{-1} + \lambda_j} \quad (13)$$

and λ_j is the *jth* eigenvalue of the PCA decomposition of matrix $\frac{X^T\left(W_k X\right)}{\sigma_k^2}$ (see [27] for details).

Note that β_k requires the definition of B_k, which in our context is taken from the previous iteration of the EM algorithm.

For the mixture mixing coefficients, we use a symmetric Dirichlet distribution as prior

$$\mathbb{P}(\Pi) = \text{Dirichlet}(\Pi|\alpha), \quad (14)$$

where α is the hyper-parameter. Hence, the mixing coefficients estimates used by the EM algorithm are

$$\pi_k = \frac{\sum_{i=1}^{N} r_{ik} + \alpha - 1}{N - N\left(\alpha - 1\right)}. \quad (15)$$

We use a prior of α = 2, which avoids models with a low number of observations assigned to it.

Transcription factor affinity

TF binding motifs are traditionally described in the form of position frequency matrices (PFMs). PFMs show how often a certain base occurs at a given position in the alignments of known binding sites of the TF. To predict the binding strength of a given TF to a promoter sequences we utilize the TRAP method [17]. In contrast to motif matching algorithms which make a binary distinction between binding sites and non-binding sites, TRAP avoids this artificial separation and instead

computes the probability of a TF to bind site i in the sequence using the following equation

$$a_i = \frac{R_o e^{-\delta E_i(\lambda)}}{1 - R_o e^{-\delta E_i(\lambda)}}, \quad (16)$$

where $\delta E_i(\lambda)$ is the energy difference between the state in which the factor is bound to site i and the state in which the factor is bound to its consensus site. This so called mismatch energy is scaled by a parameter λ which was previously determined to have an optimal value of 0.7 [17]. The second transcription factor dependent parameter R_0 determines both, the binding energy between the factor and its consensus site as well as the TF concentration. R_0 is derived for each PFM individually as

$$R_0 = \exp(0.6 \cdot W - 6), \quad (17)$$

where W is the number of columns in the PFM with information content exceeding 0.1 bits. Matrix positions which fall below this entropy cutoff also do not contribute to the mismatch energy in Eq. 16. The nucleotide dependent mismatch energies for each site in the promoter sequence are computed by

$$\delta E_i(\lambda) = -\frac{1}{\lambda} \sum_{\alpha \in \{A,C,G,T\}} \ln \frac{v_{i,\max}}{v_{i,j}}, \quad (18)$$

where $v_{i,max}$ is the frequency of the consensus base at position i in the PFM and $v_{i,\alpha}$, is the frequency of the observed base α at position i in the PFM. Eventually, TRAP obtains the expected number N of TFs bound to the promoter by summing over the individual probabilities from all L sites in the sequence:

$$N = \sum_{i=1}^{L} a_i. \quad (19)$$

As input, TRAP requires for each TF a PFM suitable for computing the mismatch energies and a DNA sequence of interest (see [17] for details).

For our study we use a selection of 102 PFMs from the Transfac database version 11.1 [28], which correspond to TFs involved in lymphoid development (see Additional File 1 for TF list). As we are mainly interested in binding sites near the promoter, the analysis was based on the 200 base pairs upstream of the transcription start site (TSS) of the genes. We restrict the analysis to genes with normalized CpG content < 0.5 in their promoter sequence [21], as such genes tends to be expressed in a tissue and stage specific way. In the end, we calculate the affinity (Eq. 19) for all the selected genes and PFMs. This yields the matrix X containing the TF binding data, where $x_{i,j}$ corresponds to the affinity of TF j to the promoter of gene i.

T-cell gene expression and histone modification data

We use the gene expression and histone modification data from Th1, Th2, Th17 and iTreg cells published by [16]. The histone modification data was measure with the Chip-Seq Illumina platform. We used the Cisgenome tool [29] to align sequence data and to detect peaks. As we are only interested in the modifications near the promoter, we consider the region of 8000 bps upstream and 2000 bps downstream of the TSS and kept the tag counts of the highest peak. Finally, we added a pseudo count to avoid zero values and applied a log transform. This yields the matrix X containing the histone modification data, where $x_{i,j}$ corresponds to the number of ChIP-seq tags derived from a particular histone modification j that are being mapped to the promoter of gene i.

The expression data was measured with Affymetrix 430 chips. The raw data has been normalized using the variance stabilization method of [30] and normalized the tissues to have mean expression equal to zero.Microarray probes were mapped to ENSEMBL gene identifiers with the help of the biomart tool [31]. We thereby kept all genes that had their expression measured by multiple probe sets. In the following, we restrict our analysis to those 6154 genes with low CpG content for which both, gene expression as well as histone modification data is available. The final data sets used in this analysis can be found at http://www.cin.ufpe.br/~igcf/MixLin.

Experimental design

We model gene expression from four different T helper cell types (Th1, Th2, Th17 and iTreg) with the use of either transcription factor affinities (TF), histone modifications (HM) or both regulatory signals combined (HM +TF). As parameter of our method, we vary the number of linear models, K, from 1 to 6. In order to select the optimal model for each cell type, we first perform 10 fold cross-validation on each parameter setting and then estimate the Mean Square Errors (MSE) from the validation sets. As the MSE tends to decrease with higher K[32], we use a model selection procedure, the Bayesian Information Criteriun (BIC) [25], to indicate the optimal number of models. The method has been implemented with Pymix [33] and is freely available at http://www.pymix.org.

Additional material

Additional file 1: Supplementary Figures and Tables This file contains additional Figures and Tables.

Acknowledgements and funding
This work has been partially supported by Brazilian research agencies: FACEPE, CNPq and CAPES.
This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 1, 2011: Selected articles from the Ninth Asia Pacific Bioinformatics Conference (APBC 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S1.

Author details
[1]Center of Informatics, Federal University of Pernambuco, Recife, Brazil. [2]Dept. of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany.

Authors contributions
IGC, RH, TGR implemented the approach and performed the experiments. IGC, RH and FATC designed the study and evaluated the results. All authors wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 15 February 2011

References

1. Zhu J, Paul WE: **CD4 T cells: fates, functions, and faults.** *Blood* 2008, **112(5)**:1557-1569.
2. Goldberg AD, Allis CD, Bernstein E: **Epigenetics: a landscape takes shape.** *Cell* 2007, **128(4)**:635-638.
3. Kouzarides T: **Chromatin modifications and their function.** *Cell* 2007, **128(4)**:693-705.
4. Turner BM: **Defining an epigenetic code.** *Nat Cell Biol* 2007, **9**:2-6.
5. Bibikova M, Laurent LC, Ren B, Loring JF, Fan JB: **Unraveling epigenetic regulation in embryonic stem cells.** *Cell Stem Cell* 2008, **2(2)**:123-134.
6. Schoenborn JR, Dorschner MO, Sekimata M, Santer DM, Shnyreva M, Fitzpatrick DR, Stamatoyannopoulos JA, Stamatoyonnapoulos JA, Wilson CB: **Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding interferon-gamma.** *Nat Immunol* 2007, **8(7)**:732-742.
7. Costa IG, Roepcke S, Schliep A: **Gene expression trees in lymphoid development.** *BMC Immunol* 2007, **8**:25.
8. Costa IG, Roepcke S, Hafemeister C, Schliep A: **Inferring differentiation pathways from gene expression.** *Bioinformatics* 2008, **24(13)**:i156-i164.
9. Bussemaker HJ, Foat BC, Ward LD: **Predictive modeling of genome-wide mRNA expression: from modules to molecules.** *Annu Rev Biophys Biomol Struct* 2007, **36**:329-347.
10. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27(2)**:167-171.
11. Keles S, van der Laan M, Eisen MB: **Identification of regulatory elements using a feature selection method.** *Bioinformatics* 2002, **18(9)**:1167-1175.
12. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M: **Histone modification levels are predictive for gene expression.** *Proc Natl Acad Sci U S A* 2010, **107(7)**:2926-2931.
13. Woolf E, Xiao C, Fainaru O, Lotem J, Rosen D, Negreanu V, Bernstein Y, Goldenberg D, Brenner O, Berke G, Levanon D, Groner Y: **Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis.** *Proc Natl Acad Sci U S A* 2003, **100(13)**:7731-7736.
14. DeSarbo W, Cron W: **A maximum likelihood methodology for clusterwise linear regression.** *Journal of Classification* 1988, **5(2)**:249-282.
15. Hinton GE, Revow M, Dayan P: **Recognizing Handwritten Digits Using Mixtures of Linear Models.** In *NIPS*. MIT Press;Tesauro G, Touretzky DS, Leen TK 1994:1015-1022.
16. Wei G, Wei L, Zhu J, Zang C, Hu-Li J, Yao Z, Cui K, Kanno Y, Roh TY, Watford WT, Schones DE, Peng W, Sun HW, Paul WE, O'Shea JJ, Zhao K: **Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells.** *Immunity* 2009, **30**:155-167.
17. Roider HG, Kanhere A, Manke T, Vingron M: **Predicting transcription factor affinities to DNA from a biophysical model.** *Bioinformatics* 2007, **23(2)**:134-141.
18. Barreda DR, Belosevic M: **Transcriptional regulation of hemopoiesis.** *Dev Comp Immunol* 2001, **25(8-9)**:763-789.
19. Matthias P, Rolink AG: **Transcriptional networks in developing and mature B cells.** *Nat Rev Immunol* 2005, **5(6)**:497-508.
20. Rothenberg EV, Moore JE, Yui MA: **Launching the T-cell-lineage developmental programme.** *Nat Rev Immunol* 2008, **8**:9-21.
21. Roider HG, Lenhard B, Kanhere A, Haas SA, Vingron M: **CpG-depleted promoters harbor tissue-specific transcription factor binding signals-implications for motif overrepresentation analyses.** *Nucleic Acids Res* 2009, **37(19)**:6305-6315.
22. Mo X, Kowenz-Leutz E, Laumonnier Y, Xu H, Leutz A: **Histone H3 tail positioning and acetylation by the c-Myb but not the v-Myb DNA-binding SANT domain.** *Genes Dev* 2005, **19(20)**:2447-2457.
23. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *Journal of the Royal Statistical Society Series B* 2005, **67(2)**:301-320.
24. Dempster A, Laird N, Rubin D: **Maximum likelihood from incomplete data via the EM algorithm.** *Journal of the Royal Statistical Society, Series B* 1977, **39**:1-38.
25. McLachlan GJ, Peel D: *Finite Mixture Models* Wiley Series in Probability and Statistics., Wiley, New York; 2000.
26. Breiman L: **Bagging Predictors.** *Machine Learning* 1996, 123-140.
27. MacKay DJC: **Bayesian Interpolation.** *Neural Computation* 1992, **4(3)**:415-447.
28. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DUU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic acids research* 2003, **31**:374-378.
29. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nat Biotechnol* 2008, **26(11)**:1293-1300.
30. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl 1)**: S96-104.
31. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart-biological queries made easy.** *BMC Genomics* 2009, **10**:22.
32. Brusco MJ, Cradit JD, Steinley D, Fox GL: **Cautionary Remarks on the Use of Clusterwise Regression.** *Multivariate Behavioral Research* 2008, **43**:29-49.
33. Georgi B, Costa IG, Schliep A: **PyMix - The Python mixture package - a tool for clustering of heterogeneous biological data.** *BMC Bioinformatics* 2010, **11**:9.

doi:10.1186/1471-2105-12-S1-S29
Cite this article as: Costa *et al.*: Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics* 2011 **12**(Suppl 1):S29.