BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 4 2009, pages 435–442
doi:10.1093/bioinformatics/btn627

Genome analysis

PASTAA: identifying transcription factors associated with sets of co-regulated genes

Helge G. Roider*, Thomas Manke, Sean O'Keeffe, Martin Vingron and Stefan A. Haas
Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin

Received on September 8, 2008; revised on November 13, 2008; accepted on December 1, 2008
Advance Access publication December 9, 2008
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: A major challenge in regulatory genomics is the identification of associations between functional categories of genes (e.g. tissues, metabolic pathways) and their regulating transcription factors (TFs). While, for a limited number of categories, the regulating TFs are already known, still for many functional categories the responsible factors remain to be elucidated.

Results: We put forward a novel method (PASTAA) for detecting transcriptions factors associated with functional categories, which utilizes the prediction of binding affinities of a TF to promoters. This binding strength information is compared to the likelihood of membership of the corresponding genes in the functional category under study. Coherence between the two ranked datasets is seen as an indicator of association between a TF and the category. PASTAA is applied primarily to the determination of TFs driving tissue-specific expression. We show that PASTAA is capable of recovering many TFs acting tissue specifically and, in addition, provides novel associations so far not detected by alternative methods. The application of PASTAA to detect TFs involved in the regulation of tissue-specific gene expression revealed a remarkable number of experimentally supported associations. The validated success for various datasets implies that PASTAA can directly be applied for the detection of TFs associated with newly derived gene sets.

Availability: The PASTAA source code as well as a corresponding web interface is freely available at http://trap.molgen.mpg.de

Contact: roider@molgen.mpg.de

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

1 INTRODUCTION

The elucidation of transcriptional regulatory networks is essential for understanding how cells integrate internal as well as external signals, ultimately controlling processes like progression through the cell cycle, appropriate response to cellular stress or differentiation of stem cells into adult tissues. Transcription factors (TFs) constitute a central component of such networks by regulating the expression of housekeeping as well as cell type-specific genes. The action of one or more TFs can thereby cause the co-expression of entire cohorts of genes. Therefore, genes expressed in a certain category such as a cell type or stress condition are expected to share binding signals for the same TFs. However, uncovering binding signals of the TFs responsible for the observed expression patterns constitute a major challenge for both experimentalists as well as theoreticians.

Given a set of genes expressed in the same functional category (metabolic pathway, tissue, developmental stage, etc.), two basic strategies are traditionally applied to find regulatory signals in the sequences. The first approach is based on *de novo* identification of sequence patterns over-represented in the putative promoter regions of these genes (Bailey and Elkan, 1995; Huber and Bulyk, 2006; Smith *et al*., 2006; van Helden *et al*., 2000). While this strategy allows detecting the presence of so far uncharacterized sequence motifs, the patterns need to be well defined in order to obtain statistical significance (Frith *et al*., 2004b). Such methods are also more sensitive to the occurrence of repeat-like sequences not filtered out by standard tools (Frith *et al*., 2004b).

The alternative approach avoids many of these problems by focusing only on the occurrences of matches to predefined, experimentally derived TF binding motifs. With larger collections of experimentally derived TF binding motifs becoming available, this approach has gained wide popularity. For a manually selected set of tissue specifically expressed genes, Wasserman and Fickett (1998) were the first to use this method successfully to predict TFs involved in the regulation of muscle-specific genes. Subsequently, several studies revealed additional TF–tissue associations for a limited number of TFs (Frith *et al*., 2004a; Qian *et al*., 2005; Yu *et al*., 2006), usually by analysing the proximal promoters of tissue-specifically expressed genes derived from microarray or expressed sequence tags (EST) data. In order to be able to include distal *cis*-regulatory elements in the analysis such methods are frequently combined with phylogenetic footprinting, which limits the sequence space to likely regulatory elements (Ho Sui *et al*., 2007; Pennacchio *et al*., 2007).

An important prerequisite for any of the above methods is the adequate definition of groups of genes expected to be co-regulated by the same factor. Generally, such groups can be inferred either from databases such as Gene Ontology (GO) (Hill *et al*., 2002) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Aoki and Kanehisa, 2005) through a binary assignment of the genes to the groups or from functional genomics data such as microarrays or ESTs in which case the specificity of a given gene for a given category (e.g. vertebrate tissues) is measured quantitatively. However, also such non-binary data are usually transformed into binary assignments by introducing an arbitrary cut-off, thereby discarding the information about the relative likelihood of a gene belonging to a category.

*To whom correspondence should be addressed.

In this article, we put forward a novel method to detect TFs that are associated with particular functional categories of genes. We call our method PASTAA for predicting associated transcription factors from annotated affinities, because as a first step we rank all genes by the predicted affinity of the TF to the genes' promoters. The expectation, of course, is that target genes of the TF rank high in this list. To detect an association between TFs and a category, this ranked list is compared to another ranking of the same genes, which should reflect the likelihood of the genes belonging to the category under study. Typically, this ranking will be based on the degree of specificity of a gene for a tissue as derived from expression data.

We will show that recognizing an association between a TF and a functional category of genes can then be reduced to determining an enrichment of common genes at the top of both lists. To this end, we propose an iterated hypergeometric test applying varying cut-offs to the two lists. Repeating this procedure for all available TF binding motifs allows delineation of the most important associations of TFs with the category under study. Importantly, in this approach it is not required to set any cut-offs a priori on either binding of a TF to a promoter or membership of a gene in a category. A similar approach has been applied by (Eden *et al*., 2007) to discover TF binding motifs in ranked lists of DNA sequences.

We validate the method by attempting to rediscover the TFs that were used in different Chromatin ImmunoPrecipitation on chip (ChIP)–chip experiments, utilizing the binding *P*-values from the experiments for the ranking of the genes. For gene lists derived from tissue-specific expression data, we show that PASTAA yields a more comprehensive number of functional TF–tissue associations than alternative methods.

2 METHODS

2.1 TF binding data (ChIP–chip, ChIP–PET)

As a first set of validation categories, we utilized the yeast genome-wide datasets on *in vitro* TF–DNA interactions available for the three TFs (Rap1, Mig1 and Abf1) from Mukherjee *et al*. (2004) and the *in vivo* ChIP–chip data from Harbison *et al.* (2004) for more than 200 TFs in various cell conditions. In both studies, the authors provide binding measurements for each factor to all approximately 6000 yeast intergenic regions. Here, we analyse the datasets corresponding to those 25 TFs for which position specific frequency matrix (PFMs) are available in TRANSFAC. Our TF binding affinity predictions are computed for each of the 25 matrices to all approximately 6000 intergenic regions.

For validation on vertebrate ChIP–chip data, we refer to the study by Odom *et al.* (2004) where the binding of the three factors HNF1, HNF4 and HNF6 to approximately 13 000 human promoters was measured. PASTAA thereby uses the provided *in vivo* binding *P*-values to rank all promoters for a given TF, while the sequences spotted on the microarrays are used to compute the binding affinities for each of the 589 vertebrate PFMs contained in TRANSFAC.

For validation on ChIP–PET data, we utilize the cMYC dataset by Zeller *et al.* (2006). In contrast to ChIP–chip binding values, the size of paired end tags (PET) clusters does not allow for an unambiguous ranking of the target sequences, i.e. there are some 2×10^5 PET singletons, 12×10^3 PET clusters of size 2 and about 10^3 clusters of size ≥ 3. We thus ranked the sequences according to cluster size but followed the proposal of Zeller *et al.* (2006) and used only clusters of size ≥ 3 as input to PASTAA. The sequences spanning the clusters (average length: 2121 bp) were used to compute binding affinities. As background set, 10 000 sequences of length 2121 bp with random genomic start positions were used.

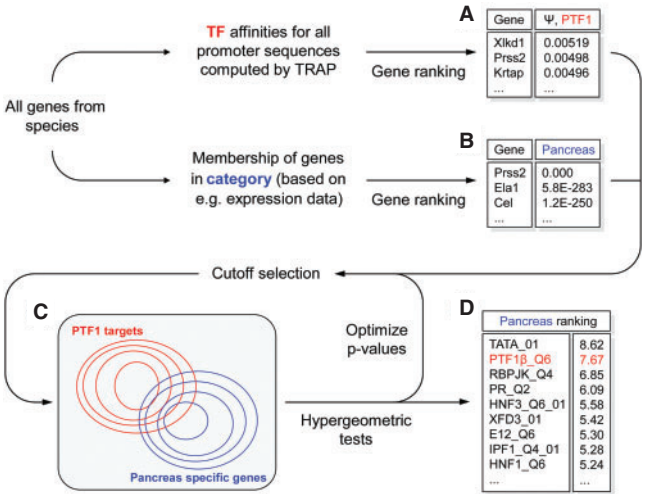

Fig. 1. The PASTAA workflow.

2.2 EST expression data

The expression of a given gene in a given tissue from human and mouse is determined by analysing corresponding EST clusters from the GeneNest database (Haas *et al*., 2000), which includes the annotation of the originating tissue for each EST. Tissue specificity of a given gene is thereby evaluated by computing a *P*-value reflecting the overrepresentation of ESTs from a tissue among all ESTs of a given cluster (see Supplementary Material and Gupta *et al*., 2005 for details). To make results comparable between the different tested methods, only EST clusters with *P*-value $<10^{-6}$ in at least one of the tissue categories are utilized. For the PASTAA analysis, TF affinities are computed for all 589 vertebrate TRANSFAC matrices and for 200 bp upstream of the transcription start sites of all 26 000 mouse genes in Ensembl (Birney *et al.*, 2006).

2.3 Methods overview

An overview of PASTAA's workflow is shown in Figure 1. All genes are ranked according to their predicted affinity for a given TF such as the pancreas-specific TF PTF1 (A). At the same time, the genes are also ranked according to their association with a given category such as pancreas (B). After applying a cut-off to the lists in (A) and (B), a hypergeometric test is used to determine the significance of the overlap between the top target genes of the TF and the top ranking genes in the category [illustrated by the Venn diagram in (C)]. Cut-offs are thereby chosen iteratively in such a way that the obtained hypergeometric *P*-values (ovals indicate the corresponding changes in set sizes) are minimized. The negative logarithms of these optimal *P*-values are used as scores to subsequently rank all PFMs for the category under investigation (D).

2.4 TF binding predictions

For the analysis of the vertebrate datasets, we use the 589 PFMs for vertebrates provided by the TRANSFAC database version 11.1 (Matys *et al.*, 2006). For the yeast analysis, we use the set of 56 fungi PFMs in TRANSFAC. For each PFM the binding affinity, $\langle N \rangle$, between the corresponding TF and a given DNA sequence is computed by our previously published TRAP method (see Supplementary Material for details).

2.5 TF affinity predictions using TRAP

To predict the binding strength of a given TF to a promoter sequences we utilize the TRAP method (Roider *et al*., 2007). TRAP avoids the artificial

separation between binding sites and non-binding sites but instead computes the binding probability of a given TF to each site in the sequence. These binding probabilities are summed over all positions in a sequence to give an estimate on the total binding affinity of the TF for a putative promoter. The affinities are then used to rank all promoters for the given TF. For details see Supplementary Material.

2.6 Measuring TF–gene category associations

In order to detect an association between a TF and a given functional category, we test for the enrichment of genes from the category among the high-ranking genes of the TF (Fig. 1). Given binary assignments for all genes (see subsequently), the enrichment of target genes of a TF among the genes belonging to a category is evaluated by the following hypergeometric test:

$$P(x \geq X) = 1 - \sum_{k=0}^{X-1} \frac{\binom{T}{k}\binom{\text{N - T}}{\text{C - k}}}{\binom{\text{N}}{\text{C}}} \quad (1)$$

where N is the number of all genes in the input set, C is the number of genes assigned to a category, T is the number of targets for a TF and X is the number of observed targets in the category.

The significance of the TF–gene category associations obtained from the above hypergeometric test depends on the cut-off used to make a binary assignment of the genes to a category and on the cut-off on the predicted affinity, $\langle \text{N} \rangle$, used to specify the targets of a given TF. Since the optimal values for these two cut-offs are not known a priori, we loop over a set of cut-offs on both the values that determine association with a category (e.g. significance of the expression of a gene in a tissue) as well as on $\langle \text{N} \rangle$. For the gene list ranked by likelihood of the genes belonging to the category, the cut-off is chosen in such a way that sets containing $\{1, 2, \ldots, 99, 100, 110, \ldots, 290, 300, 400, \ldots, 1000\}$ genes (however maximal the number of genes in the input set) are generated. On $\langle \text{N} \rangle$ the cut-off is selected so that target gene sets of size $\{25, 50, \ldots, 150, 175, 200, 250, \ldots, 500, 600, \ldots, 1000\}$ are obtained. Together the two sets of cut-offs give a total of 2413 combinations for the number of genes in a given category and the number of target genes for a given TF. In general, each of these cut-off combinations will yield a different hypergeometric P-value. We assume that the smallest achieved hypergeometric P-value corresponds to the most meaningful detectable association between a given TF and a set of genes in a given category. The negative logarithms of the most significant P-values are used as scores to subsequently rank the associations for a given TF or a given tissue. The resulting ranking reflects the relative rather than the absolute association of the TFs with respect to a given category. To assign absolute P-values to the scores we compare the results to 10^6 resamplings, which have been pre-computed for any given input set size by randomly shuffling the rankings of the genes for both lists. This procedure allows for fast subsequent assessment of the significance of the enrichment scores and accounts for the dependency between consecutive test scores.

Besides expression data, groups of genes may also be derived from categorical data as presented by databases such as KEGG (Aoki and Kanehisa, 2005) or GO (Hill *et al*., 2002). In such a case, one might seek to find TFs that regulate the expression of genes unambiguously assigned to a particular metabolic pathway or cellular process. The genes belonging to such a category are not ranked and are thus all treated equal, that is, no additional cut-off is applied to the input list.

2.7 TF expression in predicted top ranking tissue

To test whether TFs are in general preferentially expressed in the tissues most significantly enriched with their target genes, we first select for each TF the PFM yielding the most significant hypergeometric P-value for a given tissue. This is done to avoid any bias potentially introduced when evaluating multiple PFMs for a given TF. In order to assign a TRANSFAC matrix to the related EST cluster, we mapped the protein sequence of the respective TF to the mouse or human EST cluster with highest sequence similarity according to BLASTX. TFs with EST cluster P-value $<10^{-6}$ in the corresponding tissue were selected as specifically expressed. Subsequently, all cases where a TF is specifically expressed in its top ranking tissue were put in a first bin, all cases where a TF is specifically expressed in its second to top tissue in a second bin and so forth. For each TF, this procedure was repeated over all its 72 tissue associations. The ultimate assessment of the size of the resulting bins is complicated by the fact that tissue categories with few ESTs are not only less likely to express the TF, but are also less likely to produce significant hypergeometric P-values. Therefore, there exists an intrinsic negative correlation between the ranks of the tissue and the number of TFs expressed per tissue. To assess whether the enrichment is higher than expected, we repeated the entire analysis 10 times, every time assigning a random 200-bp long DNA sequence to each of the genes. The difference between the actual results and the ones obtained from the random sequences in each of the 72 bins was finally evaluated by the following t-statistic:

$$t_i = \frac{bin_{i,g} - \overline{bin}_{i,r}}{\sigma_r} \quad (2)$$

where $bin_{i,g}$ is the number of TFs assigned to bin i using the real genomic sequences, $bin_{i,r}$ is the average number of TFs assigned to bin i over all 10 random sequence sets and σ_r is the SD of the number of TFs in bin i obtained over the 10 random sets.

2.8 Comparison to PAP, z-statistics and Clover

The PASTAA algorithm was compared to three widely used methods for predicting TF–tissue associations.

Promoter Analysis Pipeline (PAP) (Chang *et al*., 2007) was accessed via logging into bioinformatics.wustl.edu/webTools/portalModule/PromoterSearch.do. Clover (Frith *et al*., 2004a) was obtained from http://zlab.bu.edu/clover/. As input we provided the set of 589 TRANSFAC vertebrate PFMs, a sequence set corresponding to the 200 bp promoters of the mouse genes in a given tissue category, and as background DNA all mouse promoter sequences not contained in the tissue category. Input and background sequence sets hereby had very similar overall GC content. For both PAP and Clover we used default settings. For the z-score statistics we referred to a standard statistical method for the annotation of TF hits (Rahmann *et al*., 2003), which balances the expected number of true and false positive binding site predictions and subsequently applied a z-score statistics [as used for instance by oPOSSUM (Ho Sui *et al*., 2007)] to detect any binding site enrichment in a given tissue category. As input to the z-score statistics, we used the same 589 TRANSFAC PFMs and 200 bp proximal promoters as above. In order to make the results optimally comparable between PASTAA, Clover, z-score statistics and PAP, we restrict the tissue sets for the comparison to only those genes whose IDs could be unambiguously matched to entries in the PAP database.

3 RESULTS

3.1 PASTAA identifies meaningful cut-offs

To validate PASTAA's ability to select meaningful sets of input genes and predicted target genes from ranked lists, we utilized the large-scale ChIP–chip and PBM binding data from Harbison *et al.* (2004) and Mukherjee *et al*. (2004) where the binding between a given TF and all approximately 6000 intergenic regions from yeast has been measured.

As a first test case, we supply PASTAA with a list of all intergenic regions ranked either according to their measured *in vitro* binding strength with the factor Abf1 or ranked according to their predicted affinities based on the matrix ABF1_01 from TRANSFAC. PASTAA obtains the most significant overlap between experimentally bound

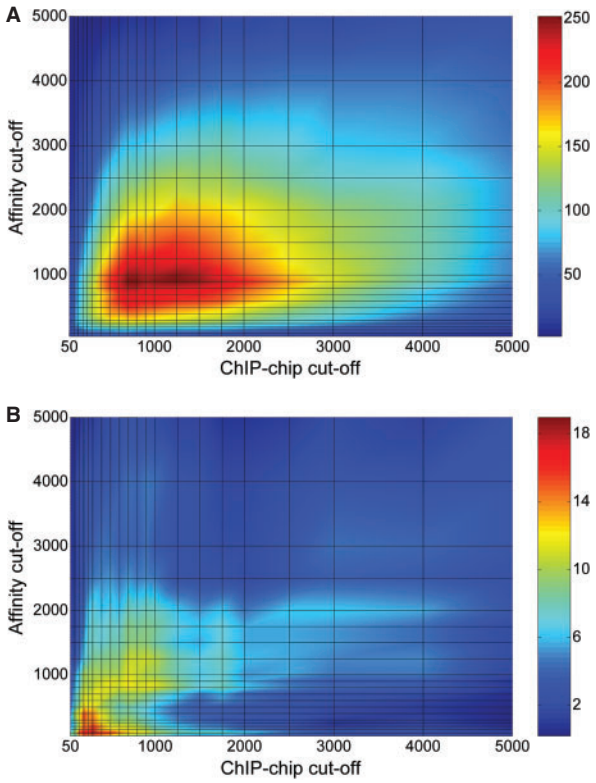

Fig. 2. Cut-off space for the hypergeometric test. (**A**) The −log hypergeometric *P*-values (indicated by colour) for ABF1_01 and the Abf1 *in vitro* dataset depending on the cut-off combination employed for the predicted affinity and PBM binding values. The most significant target enrichment (*P*-value 7.3×10^{-253}) is found when using the top 800 genes according to PBM and top 900 genes according to affinity. The steepest increase in −log *P*-values is found at the origin of the plot. (**B**) Same analysis as in (A) but for the factor PHO4_01 and the Pho4 ChIP–chip dataset (phosphate-deprived condition). According to the fact that Pho4 has far less targets than Abf1 an optimal hypergeometric *P*-value of 7.9×10^{-20} is found when using only the top 300 genes according to ChIP–chip data and top 100 genes according to affinity.

sequences and predicted targets by selecting the top 800 intergenic regions according to ChIP measurements and the top 900 genes according to predicted affinities. These sets share a total of 474 intergenic regions (*P*-value $\sim 10^{-253}$). The dependency of the *P*-values on the chosen cut-offs and their apparent convergence to an optimal value is illustrated in Figure 2A. Given that most factors have considerably fewer real targets than Abf1, which is a global transcriptional regulator involved in the regulation of a multitude of genes (Miyake *et al.*, 2004), the optimal cut-offs for more specifically acting factors are expected to yield fewer than 1000 genes for both the target and measurement set. In fact, all other 25 tested yeast factors had optimal cut-offs below 1000 genes, as illustrated in Figure 2B for the matrix PHO4_01 and its corresponding ChIP–chip dataset. For this factor, with only a few target genes (Springer *et al.*, 2003), the most significant association (*P*-value $\sim 10^{-19}$) is found when using the top 300 genes according to ChIP–chip data and top 100 genes according to predicted affinity. For efficiency we thus restrict the further analyses to the top 1000 genes in either list.

Table 1. Top associated PFMs for the HNF and MYC target gene sets

HNF1	HNF4	HNF6	MYC
HNF1_Q6	HNF4_Q6_01	CDPCR1_01	E2F1_Q3
HNF1_01	HNF4_01	OCT_01	E2F_Q2
HNF1_Q6_01	HNF4_01_B	CDP_02	MYC_Q2
HNF1_C	HNF4_Q6	CDPCR3HD_0	ETF_Q6
AR_02	STAF_02	HNF6_Q6	E2F1_Q4
HNF4_Q6_01	HNF4_DR1_Q3	PBX1_02	ZF5_01
AR_03	COUPTF_Q6	HNF1_C	MYCMAX_B
FHL_01	T3R_01	CDPCR3_01	CHCH_01
CEBP_Q2	COUP_01	CDP_01	AP2ALPHA_01
RORA1_01	STAT_01	E2F_Q3_01	ZF5_B

Top ranking PFMs for the HNF1, HNF4 and the HNF6 ChIP–chip datasets and the cMYC ChIP–PET dataset. Matching PFMs are indicated in red. Matrices for E2F, a co-regulator of MYC genes, are indicated in yellow.

3.2 Recovery of yeast ChIP–chip data

We next assess how well PASTAA detects the TFs corresponding to a given PBM or ChIP–chip dataset by evaluating the association between a given dataset and all 56 yeast matrices in TRANSFAC. To this end, we rank all PFMs according to their association scores (−log of the most significant hypergeometric *P*-value). For 21 out of 24 Chip–chip datasets, for which a corresponding matrix is available, PASTAA recovers the correct PFM among the five top ranking matrices. In several cases, non-matching TFs, which, however, share a similar binding motif to the correct TF, are among the top ranking factors. For instance, ADR1_01 (consensus GGGGT) and STRE (AGGGG) are among the top ranking PFMs for the Mig1 dataset (AAAATCTGGGGT). In addition to such matching motifs, PASTAA detects known co-factors for many of the datasets. For example, Lac9, a co-regulator for galactose response genes (Salmeron *et al.*, 1989), is the second highest ranked TF for the Gal4 dataset; while heat shock factor, a known co-regulator of Msn2 (Grably *et al.*, 2002), is identified as second highest ranking TF for the Msn2 dataset (see Supplementary Table S1 for details).

In many cases, the association scores drop several-fold between the top ranking PFMs and the subsequent ranking matrices. For instance, the three top ranking PFMs for the Abf1 dataset obtain scores of 300.0, 252.1 and 173.0; while the next motif, REPCAR1_01, attains a score of 7.6. To assess more quantitatively how far down the list the ranking remains meaningful, we assess the probability of generating a given score by chance. To this end we compare the PASTAA scores to that of 10^6 random resamplings. The large majority of matching PFMs obtain resampling *P*-values of $<10^{-4}$, while many unconfirmed associations are less significant (Supplementary Table S1).

3.3 PASTAA accounts for ChIP data from human

To assess PASTAA's ability to detect an enrichment of TF targets in a set of vertebrate sequences, we analysed ChIP–chip data available for the three hepatic TFs HNF1, HNF4 and HNF6 (Odom *et al.*, 2004). As input for PASTAA, we ranked all approximately 13 000 promoter sequences assessed by the experimenters according to how strongly they were bound by a given HNF factor. As shown in Table 1, for the HNF1 and HNF4 gene sets PASTAA correctly finds the highest association for the PFMs corresponding

to HNF1 and HNF4, respectively (out of 589 vertebrate PFMs present in TRANSFAC). For the HNF6 dataset the single PFM present in TRANSFAC is ranked at position five, while four PFMs corresponding to another homoeodomain factor CDP are ranked top. These results match the findings of a dedicated study that identified the same factors as top associated with the HNF6 ChIP–chip dataset (Smith *et al.*, 2005). For the cMYC dataset (Zeller *et al.*, 2006) shown in the rightmost column of Table 1, we rank two cMYC matrices among the top 10 PFMs and another MYC matrix at position 13. Interestingly, among the top matrices we also detect E2F, a key co-regulator of Myc genes (Leone *et al.*, 2001).

3.4 PASTAA predicts tissue-specific TFs

We now turn to searching for TFs involved in the regulation of sets of tissue-specific genes. To this end, we define tissue categories based on EST data and determine the significance of expression of the gene in every category. As above we produce two ranked lists, one according to the significance of expression, and one according to the predicted affinity. PASTAA then determines the most significant overlap between these lists. Affinities are hereby computed for the 200 bp upstream of the transcription start sites of all approximately 26 000 Ensembl mouse genes (Birney *et al.*, 2006). In the following, we analyse the tissues of muscle, heart, liver, leucocyte and retina. For each of these tissues a number of key regulators are known from experimental as well as computational studies, which we expect to recover with our method. As shown in Table 2, the most significantly associated matrices for muscle and heart are PFMs corresponding to muscle enhancer factor 2 (MEF2), serum response factor (SRF) and muscle specific TATA (MTATA). This is in accordance with previous findings by Wasserman and Fickett (1998). For the liver category HNF1, HNF4 are dominating the ranking (Odom *et al.*, 2004), while for leucocyte PFMs corresponding to immune related TFs such as NF-kappaB and c-Ets-1 are found (Pennacchio *et al.*, 2007). Lastly, for retina, PASTAA detects the eye-specific factors CRX [cone rod homoeobox protein (Furukawa *et al.*, 2002; Qian *et al.*, 2005)] and CHX10 (Dorval *et al.*, 2006).

Aside from these well-studied cases, we also find functional associations for several other tissues. For instance, the pancreatic TFs IPF1 [insulin promoter factor 1 (Ohlsson *et al.*, 1993)] and PTF1 [pancreas-specific transcription factor 1 (Roux *et al.*, 1989)] are listed among the top 10 factors for pancreas. The lung- and thyroid-specific factor TTF1 [thyroid transcription factor 1 (Kimura *et al.*, 1999)] is detected as the top ranking factor in the lung category and among the top 10 factors in the thyroid category. Another example is PIT1 [pituitary-specific positive transcription factor 1 (Li *et al.*, 1990)] which is detected at rank 4 in the pituitary gland category.

Importantly, very similar results are found for many tissues when analysing gene sets derived from the GNF microarray dataset instead of EST data (see Supplementary Material for results obtained from GNF data).

3.5 Comparison to alternative approaches

In order to evaluate the usefulness of PASTAA, we compared its performance to that of three alternative methods: (i) Clover (Frith *et al.*, 2004a); (ii) PAP (Chang *et al.*, 2007); and (iii) a z-score statistics [as was used for instance in oPOSSUM (Ho Sui *et al.*, 2007)] applied to a standard hit-based annotation that balances the number of false and true binding site predictions (Rahmann *et al.*, 2003).

Table 2. Result for tissues with Known TF associations

CLOVER	z-score	PAP	PASTAA		
Muscle					
SP1_Q2_01	SRF_01	TATA_01	SRF_Q5_01	7.3	1E-06
MAZ_Q6	SRF_C	T3R_Q6	SRF_01	6.2	2E-05
MEF2_Q6_01	SRF_Q5_02	MTATA_B	SRF_Q5_02	6.0	4E-05
TATA_01	SRF_Q6	SF1_Q6	SRF_C	5.9	5E-05
TBP_01	SRF_Q4	SPZ1_01	MTATA_B	5.8	5E-05
Heart					
SP1_Q4_01	SRF_01	SF1_Q6	MEF2_Q6_01	8.0	0.0
SP1_Q2_01	MEF2_02	ERR1_Q2	SRF_C	6.4	3E-05
SP1_Q6	SP1_Q4_01	ER_Q6_02	RSRFC4_01	6.1	5E-05
GC_01	UF1H3B._Q6	T3R_Q6	MTATA_B	6.0	9E-05
SP1_Q6_01	SRF_Q5_02	TATA_01	MEF2_02	5.9	9E-05
Liver					
SP1_Q4_01	HNF4_Q6_01	CEBP_Q2_01	HNF4_Q6_01	21.3	0.0
SP1_Q2_01	HNF1_01	PBX1_03	HNF1_01	20.7	0.0
GC_01	HNF4_01	CEBP*	HNF4_01	20.5	0.0
SP1_Q6_01	HNF1_Q6	GR_Q6_01	HNF1_Q6	19.3	0.0
SP1_Q6	HNF1_Q6_01	HNF1_Q6	HNF1_C	17.4	0.0
Retina					
SP1_Q2_01	UF1H3B._Q6	SREBP1_Q6	GATA1_03	12.4	0.0
CACB._Q6	SP1_Q4_01	LFA1_Q6	CRX_Q4	7.9	0.0
SP1_Q6_01	SP1_Q2_01	ZIC2_01	VMAF_01	5.1	6E-04
WT1_Q6	KROX_Q6	TFIII_Q6	SREBP1_02	4.9	9E-04
SP1_01	SP1_Q6	PAX4_03	CHX10_01	4.5	2E-03
Leukocyte					
SP1_Q4_01	SP1_Q4_01	ETS_Q6	NFK.B65_01	13.0	0.0
SP1_Q6_01	SP1_Q6	PEA3_Q6	NFK.B_01	12.2	0.0
SP1_Q2_01	GC_01	PU1_Q6	NFKB_Q6_01	11.7	0.0
GC_01	SP1_Q6_01	ETS_Q4	CREL_01	11.2	0.0
SP1_Q6	SP1_Q2_01	cREL*	ETS_Q6	10.0	0.0

Top ranking PFMs according to PASTAA and three alternative approaches. Predictions corresponding to experimentally characterized TF–tissue associations are shown in red. Associations in blue correspond to matrices for the general factor SP1 and the basal TATA box. The last two columns indicate PASTAA's association scores as well as the corresponding resampling P-values.
*JASPAR matrices (Sandelin *et al.*, 2004) used only by PAP.

Clover and the z-score statistics were used with 200 bp proximal promoters as input, while PAP uses larger promoter regions refined by phylogenetic footprinting. As shown in Table 2, PAP detects well-characterized associations especially for the liver and leucocyte tissue categories. Clover finds GC-rich motifs such as the general TF SP1 (Kaczynski *et al.*, 2003) as highly enriched in all tested categories. Accordingly, more specific associations appear at higher ranks (these results are obtained regardless of whether or not a background gene set is provided to Clover). The z-score statistics recovers many of the known muscle-, heart- and liver-specific associations but also detects GC-rich motifs as top ranking in several categories. Interestingly, neither method ranked CRX or CHX10 among the top PFMs for retina.

When analysing the HNF ChIP–chip datasets Clover and z-score statistics yield similar results to PASTAA. In contrast, for the MYC dataset, Clover ranked the first MYC matrix at position 48 while all but one other MYC matrix were considered anti-correlated with the input set. Similarly, for this dataset the z-score statistics ranked the first MYC matrix at position 28 while the top matrices correspond to immune-related and heat-shock factors.

3.6 TFs are over-expressed in their top ranked tissues

Above we showed that PASTAA successfully detects important TFs for groups of co-expressed genes. Here we address the reverse question: given a TF can we detect in which tissue the factor plays a

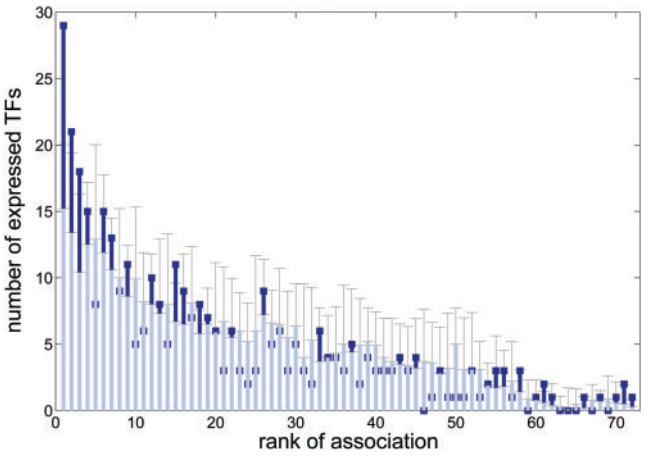

Fig. 3. TFs are over-expressed in their top ranking tissues. Height of bins indicates the number of TFs expressed in the associated tissue of given rank based on the real sequence data (dark blue) or on the results obtained from 10 random sequence sets (light blue). Error bars show the 95% confidence interval for the results obtained from the 10 random sequence sets. Tissues top ranking for a given TF express the factor more often than expected, while bottom ranking tissues express the TF equally or less often than expected. The enrichment is particularly significant for the first three bins corresponding to all three top ranking TF–tissue associations (*P*-value of enrichment for bins 1–3 combined: 2.2×10^{-12}). The general trend in the light blue bins indicates the technical bias caused by the different number of ESTs in each tissue category.

role? To assess, in an unbiased fashion, for all TFs how meaningful the top ranking tissue associations are, we analysed the expression patterns of the TFs themselves. The underlying assumption is that a TF specifically expressed in a certain tissue is likely to assert a regulatory function there. Consequently, a TF should be over-expressed more frequently among its top-ranking tissues rather than among randomly assigned tissues. In the entire dataset of 72 tissues, there are 352 TF–tissue associations where the TF is specifically expressed in the corresponding tissue. In 29 of these cases the tissue is indeed top ranking for the TF. This constitutes a 2-fold increase (*P*-value: 1.3×10^{-6}) over what would be expected by chance (see Section 2). In 21 cases the tissue is ranked second (1.6-fold increase, *P*-value: 0.019) and in 17 cases third to top (1.7-fold increase, *P*-value: 0.017). Over all the 72 possible tissue ranks a clear trend exists for the higher ranking tissues to express the corresponding TFs more often than expected, while lower ranking tissues tend to express the TFs at lower levels (Fig. 3).

It has to be noted that this verification method fails for factors such as SRF and HNF1, which are broadly expressed despite their known tissue-specific activities, or for factors such as PTF1, which do not have enough support by EST data to assess their expression patterns. To validate such TF–tissue associations, we performed an extensive manual PubMed search seeking for strong evidence for the involvement of a TF in the regulation of a tissue. This procedure confirmed an additional 149 top associations including HNF3 (Kaestner *et al.*, 1999) and PTF1 (Roux *et al.*, 1989) with pancreas, MEF2-muscle (Wasserman and Fickett, 1998), RFX-testis (Reith *et al.*, 1994) and NRSF-brain (Chen *et al.*, 1998) (Table 3).

Table 3. Top ranking tissues for a selected group of PFMs

AMEF2	Striated mus.	HEN1	Brain	IRF	Spleen	POU1F	Brain
	Heart		Cerebrum		Macrophage		Pituitary g.
	Muscle		Nerv ous		Leukocyte		Placenta
CHX10	Retina	HNF3	Pancreas	NERF	Leukocyte	PTF1β	Pancreas
	Eye		Stomach		Spleen		Dendritic cell
	Skin		Liver		Lymphocyte		Islet
CRX	Retina	HNF4	Liver	NGFIC	Cerebrum	SRF	Heart
	Eye		Kidney		Brain		Striated mus.
	Pineal		Intestine		Pituitary g.		Muscle
ETS	Leukocyte	HNF6	Liver	NRSF	Brain	STAT6	Leukocyte
	Thymus		Diaphragm		Cerebrum		Dendritic cell
	Lymphocyte		Pancreas		Retina		Lymphocyte
RFX	Testis	IPF1	Islet	PITX2	Retina	TITF1	Thyroid
	Bladder		Pancreas		Eye		Salivary g.
	Vesic. g.		Kidney		Testis		Lung

Associations supported extensively by literature or by specific expression of the TF in the respective tissue are indicated in yellow and red, respectively.

4 DISCUSSION

TFs play an important role in the regulation of genes specifically expressed in different cell stages and conditions. In order to detect functional associations between TFs and groups of co-regulated genes, we utilize the full qualitative information from functional genomics data and TF binding predictions. For the latter we have applied a biophysical model to predict binding affinities to regulatory regions. Combining the resulting rankings with an iterative search for the most significant overlap between genes in a category and target gene sets of a TF allows to robustly detect functional TF–tissue associations without the need for *ad hoc* cut-off selections. It has to be stressed that cut-offs applied to the affinity measure occur at the level of promoters and not at individual binding sites. While this still constitutes a rather artificial separation between TF target promoters and non-targets the subsequent hypergeometric test statistics is more powerful than a z-score test, which would avoid the target separation but tends to run into problems when trying to optimize the input lists (data not shown).

Using PASTAA we are able to detect associations between TFs and gene groups stemming from various sources such as ChIP–chip data as well as EST or microarray-based expression data. For the HNF and cMYC datasets we find the corresponding PFMs with high specificity, while neither Clover nor the z-score statistics ranked a MYC matrix among the top PFMs for the MYC dataset. Together these findings suggest that important biological information about regulating TFs can straightforwardly be obtained from the ranking of the PFMs for a given dataset provided by PASTAA without the need of introducing cut-offs a priori.

When applied to the analysis of tissue-specific gene sets PASTAA detects on one hand well-known TF–tissue associations, like SRF-heart, MEF2-muscle and HNF1-liver, which are usually predicted by most alternative computational approaches. In these cases, the TF–tissue association signals are so strong that the successful recovery of functional associations seems to be insensitive to the choice of the method. On the other hand, for a number of tissues the top ranking TFs diverge considerably between different methods. Many of the

association found by PASTAA are hereby strongly supported by literature as in the case CRX-retina (Furukawa *et al*., 2002), PTF1-pancreas (Roux *et al*., 1989) and TTF1-lung (Kimura *et al*., 1999). Besides extensive validation through literature, our predicted associations are additionally supported by the observation that the corresponding TFs are significantly more often over-expressed in their top ranking tissues than expected based on random sequence sets.

Despite the progress reported here, there are still a number of tissues and TFs for which no experimentally validated association could be recovered. One reason for this might be the lack of EST expression data for several tissues. Therefore, while we observed that variations in the list of genes assigned to a certain tissue category do not strongly affect the ranking of TF–tissue associations, it may still be sensible to integrate different expression datasets [as suggested by Pennacchio *et al*. (2007)].

Another reason for missing associations may be caused by TFs mainly acting on enhancer elements that are located far upstream or downstream of the transcription start site (TSS). We attempted to incorporate such elements by using evolutionary conserved sequences within 10 kb upstream of the TSS to compute the TF binding affinities but found nearly identical TF rankings for the analysed tissues (Supplementary Table S2). This indicates that the majority of detectable tissue-specific sequence signals reside within proximal promoters while signals outside of this well-defined region get overshadowed by sequence noise. Recently, databases assigning enhancer elements to genes based on synteny became available (Engström *et al*., 2008), which in future will allow to incorporate more accurately the distal regulatory modules for the affinity predictions and potentially improve tissue-specific TF binding predictions.

In addition, recent data indicate that genes can be categorized as having either a sharp TSS usually associated with a TATA box or a broad TSS often residing in CpG islands (Carninci *et al*., 2006). In this context it is interesting to note that we find a strong TATA box enrichment in many of the tissue categories for which we also find functional TF–tissue associations (Supplementary Table S4). In this context, our definition of a sharp TSS may hamper the accurate selection of the putative proximal promoter region when dealing with broad TSSs.

In general, the successful recovery of functional TF associations is strongly dependent on the definition of an appropriate set of genes acting in the same biological context as the TF. Given the substantial number of TF–tissue associations recovered by our method, we anticipate that PASTAA could also be applied directly to a non-ranked group of genes acting in a different functional context such as a metabolic pathway. Recently, two papers by Sinha *et al*. (2008) and Warner *et al*. (2008) predicted a considerable number of TFs and motif combinations associated with distinct gene sets. Sinha *et al*. (2008) also make use of the advantage of integrating weak and strong TF binding signals, but in contrast to PASTAA, both methods rely on predefined gene sets.

It is important to realize that our method as well as others merely suggest likely regulators based on statistical arguments of over-representation and enrichment. While statistical significance does not ensure biological relevance, it is reassuring to observe that our method recovers many known associations among the top ranking predictions. Nevertheless, all statistical efforts are hampered by the complex interplay of important alternative regulatory mechanisms such as post-transcriptional modifications, DNA methylation or epigenetic modifications that may force a further subdivision of functionally related genes according to the underlying regulatory mechanisms.

Funding: Biosapiens project (contract LHSG-CT-2003-503265); German National Genome Research Network (NGFN); SFB project 618.

Conflict of Interest: none declared.

REFERENCES

Aoki,K.F. and Kanehisa,M. (2005) Unit 1. 12: Using the KEGG database resource, Chapter 1. *Curr. Protoc. Bioinformatics.*

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.

Birney,E. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.

Carninci,P. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.

Chang,L.W. *et al*. (2007) PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis. *Nucleic Acids Res.*, **35**, W238–W244.

Chen,Z.F. *et al*. (1998) NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat. Genet.*, **20**, 136–142.

Dorval,K.M. *et al*. (2006) CHX10 targets a subset of photoreceptor genes. *J. Biol. Chem.*, **281**, 744–751.

Eden,E. *et al*. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.

Engström,P.G. *et al*. (2008) Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.*, **9**, R34.1–R34.12.

Frith,M.C. *et al*. (2004a) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.

Frith,M.C. *et al*. (2004b) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.

Furukawa,A. *et al*. (2002) The mouse Crx 5′-upstream transgene sequence directs cell-specific and developmentally regulated expression in retinal photoreceptor cells. *J. Neurosci.*, **22**, 1640–1647.

Grably,M.R. *et al*. (2002) HSF and Msn2/4p can exclusively or cooperatively activate the yeast HSP104 gene. *Mol. Microbiol.*, **44**, 21–35.

Gupta,S. *et al*. (2005) T-STAG: resource and web-interface for tissue-specific transcripts and genes. *Nucleic Acids Res.*, **33**, W654–W658.

Haas,S.A. *et al*. (2000) GeneNest: automated generation and visualization of gene indices. *Trends Genet.*, **16**, 521–523.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hill,D.P. *et al*. (2002) Extension and integration of the Gene Ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res.*, **12**, 1982–1991.

Ho Sui,S.J. *et al*. (2007) oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res.*, **35**, W245–W252.

Huber,B.R. and Bulyk,M.L. (2006) Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics*, **7**, 229.

Kaczynski,J. *et al*. (2003) Sp1- and Kruppel-like transcription factors. *Genome Biol.*, **4**, 206.

Kaestner,K.H. *et al*. (1999) Inactivation of the winged helix transcription factor HNF3alpha affects glucose homeostasis and islet glucagon gene expression in vivo. *Genes Dev.*, **13**, 495–504.

Kimura,S. *et al*. (1999) Thyroid-specific enhancer-binding protein/thyroid transcription factor 1 is not required for the initial specification of the thyroid and lung primordia. *Biochimie*, **81**, 321–327.

Leone,G. *et al.* (2001) Myc requires distinct E2F activities to induce S phase and apoptosis. *Mol. Cell*, **8**, 105–113.

Li,S. *et al*. (1990) Dwarf locus mutants lacking three pituitary cell types result from mutations in the POU-domain gene pit-1. *Nature*, **347**, 528–533.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Miyake,T. *et al*. (2004) Genome-wide analysis of ARS (autonomously replicating sequence) binding factor 1 (Abf1p)-mediated transcriptional regulation in Saccharomyces cerevisiae. *J. Biol. Chem.*, **279**, 34865–34872.

Mukherjee,S. *et al*. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.

Odom,D.T. *et al.* (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.

Ohlsson,H. *et al*. (1993) IPF1, a homeodomain-containing transactivator of the insulin gene. *Embo. J.*, **12**, 4251–4259.

Pennacchio,L.A. *et al*. (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.

Qian,J. *et al*. (2005) Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic Acids Res.*, **33**, 3479–3491 .

Rahmann,S. *et al*. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.*, **2**, Article 7.

Reith,W. *et al*. (1994) RFX1, a transactivator of hepatitis B virus enhancer I, belongs to a novel family of homodimeric and heterodimeric DNA-binding proteins. *Mol. Cell Biol.*, **14**, 1230–1244.

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

Roux,E. *et al*. (1989) The cell-specific transcription factor PTF1 contains two different subunits that interact with the DNA. *Genes Dev.*, **3**, 1613–1624.

Salmeron,J.M.Jr, *et al*. (1989) Interaction between transcriptional activator protein LAC9 and negative regulatory protein GAL80. *Mol. Cell Biol.*, **9**, 2950–2956.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Sinha,S. *et al*. (2008) Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res.*, **18**, 477–488.

Smith,A.D. *et al*. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, **21 (Suppl. 1)**, i403–i412.

Smith,A.D. *et al*. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl Acad. Sci. USA*, **103**, 6275–6280.

Springer,M. *et al*. (2003) Partially phosphorylated Pho4 activates transcription of a subset of phosphate responsive genes. *PLoS Biol.*, **1**, E28.

van Helden,J. *et al*. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.

Warner,J.B. *et al*. (2008) Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods*, **5**, 347–353.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Yu,X. *et al*. (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.

Zeller,K.I. *et al.* (2006) Global mapping of c-myc binding sites and target gene networks in human B cells. *Proc. Natl Acad. Sci. USA*, **103**, 17834–17839.