Zhong *et al. BMC Genomics* 2013, **14**:796
http://www.biomedcentral.com/1471-2164/14/796

BMC Genomics

METHODOLOGY ARTICLE Open Access

Predicting tissue specific transcription factor binding sites

Shan Zhong, Xin He and Ziv Bar-Joseph*

Abstract

Background: Studies of gene regulation often utilize genome-wide predictions of transcription factor (TF) binding sites. Most existing prediction methods are based on sequence information alone, ignoring biological contexts such as developmental stages and tissue types. Experimental methods to study *in vivo* binding, including ChIP-chip and ChIP-seq, can only study one transcription factor in a single cell type and under a specific condition in each experiment, and therefore cannot scale to determine the full set of regulatory interactions in mammalian transcriptional regulatory networks.

Results: We developed a new computational approach, PIPES, for predicting tissue-specific TF binding. PIPES integrates *in vitro* protein binding microarrays (PBMs), sequence conservation and tissue-specific epigenetic (DNase I hypersensitivity) information. We demonstrate that PIPES improves over existing methods on distinguishing between *in vivo* bound and unbound sequences using ChIP-seq data for 11 mouse TFs. In addition, our predictions are in good agreement with current knowledge of tissue-specific TF regulation.

Conclusions: We provide a systematic map of computationally predicted tissue-specific binding targets for 284 mouse TFs across 55 tissue/cell types. Such comprehensive resource is useful for researchers studying gene regulation.

Background

To reconstruct and model transcriptional regulatory networks (TRNs) we need to know the genome-wide binding sites of transcription factors (TFs) [1,2]. Chromatin immunoprecipitation(ChIP) followed by microarray (ChIP-chip) [3] or sequencing (ChIP-seq) [4] has been extensively used to study the *in vivo* binding locations of individual transcription factors and cofactors in a wide range of species and tissues [1,2,5-9]. Despite their popularity, such methods can only study a single TF in a single cell type, under a specific condition, in each experiment. Thus, it is difficult to use these methods to obtain a comprehensive understanding of the complicated mammalian TRNs. These networks can involve hundreds or thousands of TFs whose activities change across different tissues and conditions. Using computational methods to integrate other genomic resources in order to predict tissue-specific transcription factor binding is therefore an important research challenge.

*Correspondence: zivbj@cs.cmu.edu
Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Several methods have been developed to use *in vitro* data characterizing TF binding specificities to identify TF binding sites across the genome. Specifically, data from universal protein binding microarray (PBM) [10,11] is often used for such analysis. PBM is capable of analyzing the interaction of a sequence-specific TF with tens of thousands of short DNA sequences (probes) in a single experiment, and thus provides a highly detailed picture of TF-DNA interactions. It has been successfully applied to reveal the binding profiles of hundreds of TFs in yeast [12], worm [13], mouse [14] and arabidopsis [15]. Some of the proposed methods for using PBM data represent TF binding preference by position weight matrices (PWMs) [11,16,17]. However, PWMs, although popular due to their simplicity, assume independence between positions, an assumption which may not hold in many cases [14,18,19]. In contrast, more sophisticated models (e.g. using *k*-mers) may better represent the full binding profiles of TFs, without loss of information from using PWM. For instance, it has been suggested that many TFs have more than one binding preference [14] and these are easier to represent using k-mers.

While *in vitro* data provides important information regarding binding specificities, such data is context independent. Actual binding is highly dependent on tissue-specific conditions including chromatin accessibility, the presence of co-factors, etc [20]. Recently, a number of studies have reported that epigenetic information including certain histone modifications and hypersensitivity to DNase I cleavage correlate with TF binding *in vivo* [21,22]. Moreover, functional TFBSs tend to be under stronger negative selection, leaving a "phylogenetic footprint" in the genomic sequences. Several methods for predicting *in vivo* TF binding sites have attempted to combine such information with PWMs to predict global binding preferences [23-26]. However, as mentioned above, PWM may not be the best representation of TF binding. As we show, by using a model that retains the dependence between positions in the motif we can improve upon methods that integrate epigenetic and PWM data. In addition, none of these methods have so far been applied to elucidate the complete set of targets for TFs across a large number of tissues.

To predict accurate tissue-specific TFBS, we integrate multiple types of genomic data. The first part of our model is a biophysically-motivated *k*-mer based method for analyzing PBM data, which allows secondary binding profiles and nucleotide dependencies in different positions of the TF binding sites. Next, we develop a new method, PIPES (*p*robabilistic *i*ntegration of *P*BM, *e*pigenetics and *s*equence data), to combine the results from the PBM model with DNase I hypersensitivity (DHS) data and evolutionary conservation data to predict tissue-specific TFBS *in vivo*. We demonstrate that such an integrative model significantly boosts context specific prediction results compared with using PBM data alone. We also show that PIPES improves upon other methods developed for integrating data to predict TFBS [24,26], in some cases significantly so. Finally, we created a resource for tissue-specific TRNs using PBM data for 284 mouse TFs from UniPROBE [27] and DNase I hypersensitivity data for 55 mouse tissue/cell types from the mouse ENCODE project [28]. We predict the activities of TFs across different tissues, and, as we show, many of these predictions agree with current knowledge regarding tissue-specific roles of TFs. Our tissue specific activity predictions are also supported by global analysis of TF expression data. The comprehensive resource of TF binding sites we built thus provides a reference map for understanding complex gene expression patterns.

Results

An overview of our PIPES method is shown in Figure 1. Our model has two components: the left part of the figure shows our model for the PBM data, and the right part our model of epigenetics and conservation data. Starting with raw fluorescent intensities measured by PBM, we first infer binding probabilities to each individual *k*-mer with a biophysically-motivated model (Figure 1a). This information, based on PBM alone, can be used to score a sequence for potential TFBS (Figure 1b). Next, we use tissue specific DNase I hypersensitivity data to determine chromatin accessibility (Figure 1c, d), and combine such information with sequence conservation and the PBM derived scores to predict *in vivo* binding sites (Figure 1e).

K-mer based PBM analysis can accurately infer TF binding specificities

While a number of methods have been suggested to use PBM data for predicting TF binding sites (in most cases using PWMs), we decided to extend *k*-mer based methods using a biophysically-motivated model (Figure 1a). *K*-mer based methods were shown to achieve the best performance among several techniques for the analysis of PBM data [29]. Such methods allow an intuitive representation of potential alternative binding motifs and can account for dependency among positions in a motif. We use lasso regression with positive constraints to learn model parameters that represent binding probabilities to individual *k*-mers, where *k* is determined as part of the learning procedure. This results in a sparse model with relatively few *k*-mers having nonzero binding probabilities. On average, the number of *k*-mers with non-zero probabilities is 398.4 ± 253.2 across 284 mouse TFs with PBM data available. The model combines the benefits of recent PWM-based biophysical methods (for example, BEEML-PBM [16]) with the ability of PBMs to capture dependencies between positions in a given motif (see Methods and Supplementary Methods in Additional file 1 for details).

We illustrate the results of our PBM model using four TFs including Sox12, Esrra, Klf7 and Pou2f1. Figure 2 presents the PWMs derived from the PBM data for these TFs by the Seed-and-Wobble algorithm [11] (denoted as S&W PWMs), PWMs in TRANSFAC [30] for the corresponding TFs when available, and all the *k*-mers estimated by our model that have binding probabilities above 0.5. For S&W PWMs, when a secondary binding preference was derived [14], both the primary and secondary PWMs are shown. As can be seen, our *k*-mer model does well for this data. For Sox12, the learned *k*-mers match well with both the primary and secondary S&W PWMs (Figure 2a). For Esrra and Klf7, *k*-mers matching the consensus sequences of the primary S&W PWMs, respectively, are predicted to have high binding probabilities (Figure 2b and c), whereas *k*-mers matching the secondary S&W PWMs have lower predicted binding probabilities ranging from 0.2 to 0.44 (not shown in the figure). In the case of Pou2f1, none of the inferred top *k*-mers match the S&W PWM (Figure 2d). However, many of these

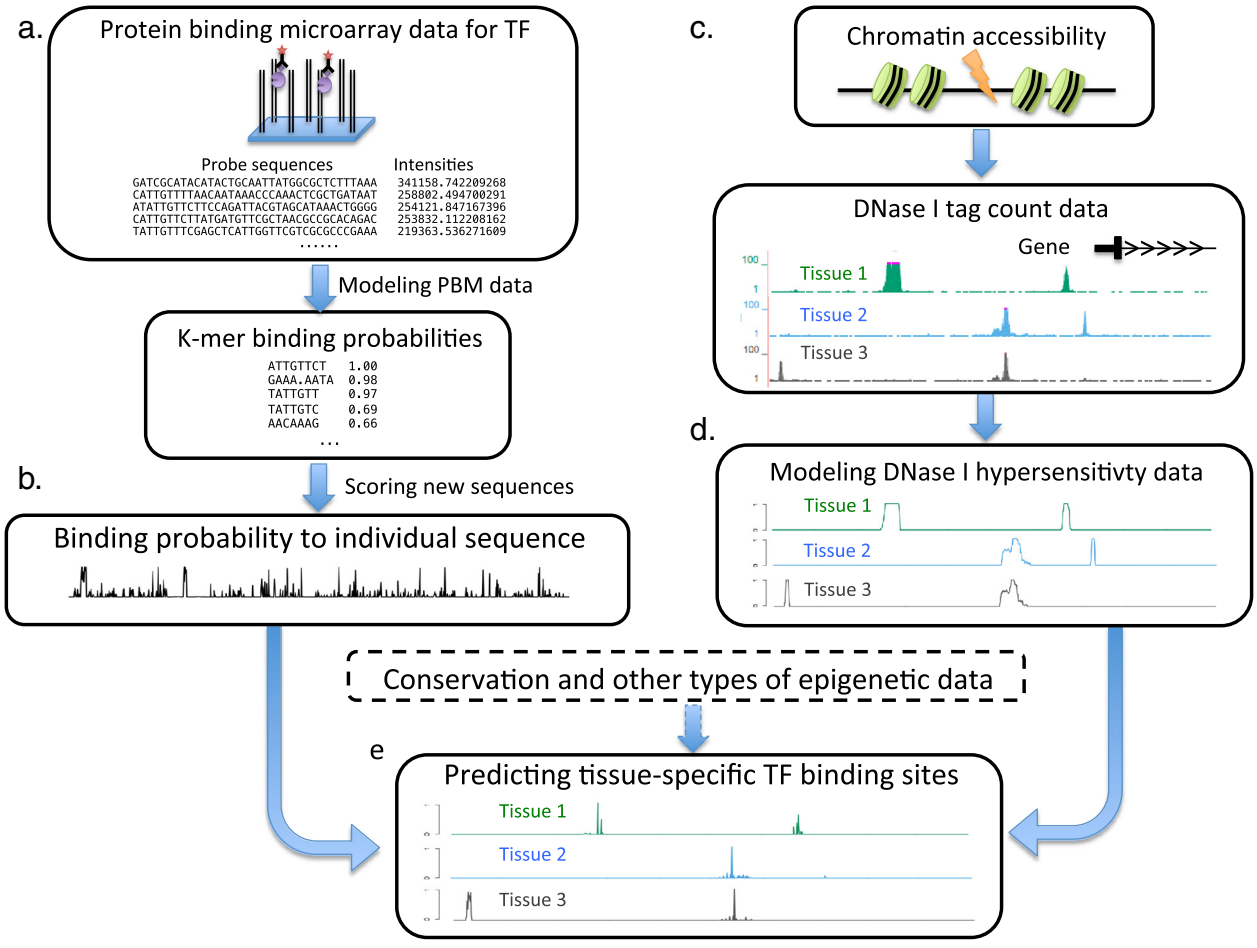

Figure 1 Overview of PIPES. (a, b) Starting with raw PBM data for a TF represented as fluorescence intensities to each individual probes, we first infer binding probabilities to individual short *k*-mers, and then a given sequence can be scored by such inferred binding probabilities. **(c, d)** To predict *in vivo* TF binding, we take as input the tissue-specific DNase I hypersensitivity data (tag counts) and convert them to probabilities that represent chromatin accessibility for each position in the genome at each tissue/cell types. **(e)** The PBM data, DNase data and other types of data including sequence conservation are combined using an integrative model. See Methods for details.

top *k*-mers closely match the consensus sequence of the TRANSFAC motif for Pou2f1 derived from literature evidence (Figure 2d). Overall, these results support the use of a biophysically-motivated model: the binding probabilities of *k*-mers are largely consistent with the results from independent methods and known motifs from TRANSFAC.

To test our PBM model, and as a baseline, we next used the inferred binding probabilities to predict *in vivo* TF binding. We collected 11 published mouse ChIP-seq datasets for which the PBM data for the same TF or for a TF with a similar DNA-binding domain is available (Additional file 2). From each ChIP-seq dataset, the top 3000 peaks with highest enrichment are extracted, and the 600bp genomic regions centered on the reported peaks are used as the positive sequences bound by the TF. Then, 600bp sequences that (1) are upstream of and (2) 300bp apart from each positive sequence, and (3) do not overlap with any other positive sequences, are used as negative sequences. We also explored two alternative options for constructing negative sequences, including using size-matched random promoter sequences and randomly generated sequences. The choice of negative sequence sets makes very little difference on the AUC values, so we will only report the results based on the first negative set here (see Supplementary Results in Additional file 1, and Additional file 3 for details). We compared our PBM model with seven other methods that predict affinities of TF binding to given sequences, and the different methods are evaluated using areas under the ROC curve (AUC) as a measure of their abilities to correctly classify the two sets of sequences (see Methods, Supplementary Results in Additional file 1 and Figure S1 therein for details).

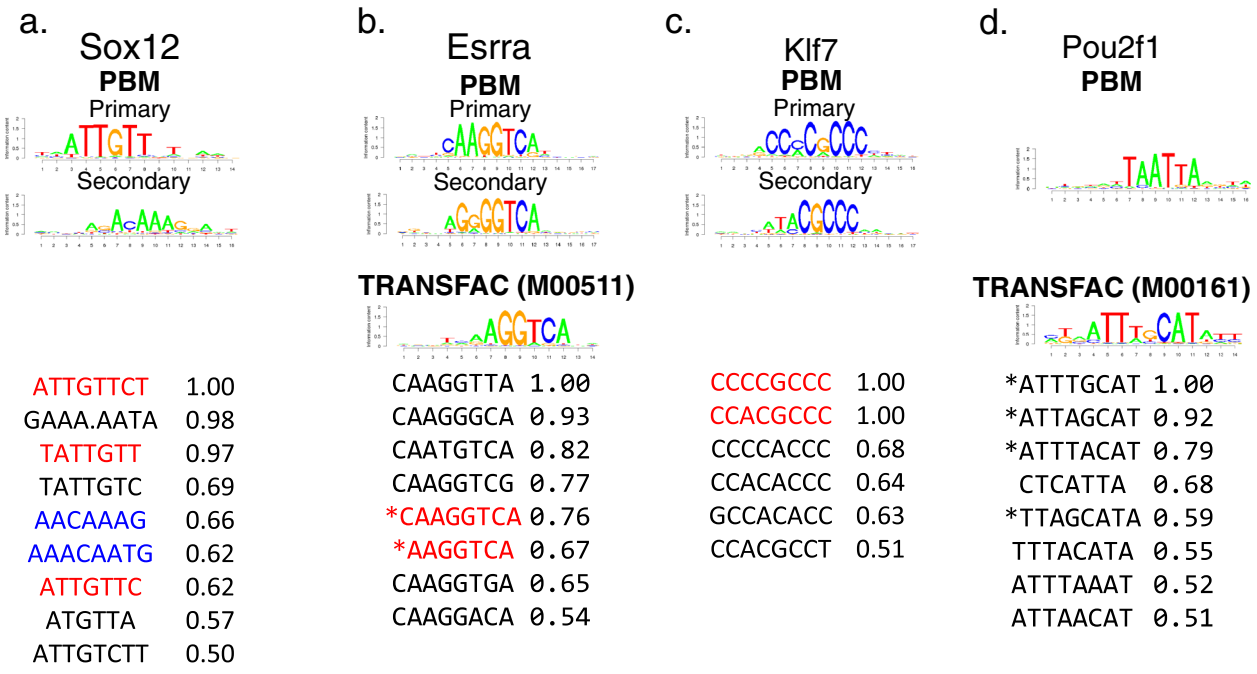

Figure 2 Top inferred *k*-mer binding probabilities for (a) Sox12, (b) Esrra, (c) Klf7 and (d) Pou2f1. The S&W PWMs and TRANSFAC motifs (when available) for these four factors are also shown. *k*-mers are colored according to whether they match the consensus sequences of the primary S&W PWM (red) or secondary S&W PWM (blue). *k*-mers matching the TRANSFAC motifs are indicated by a "*" in the front. Only k-mers with coefficients above 0.5 (normalized so that the maximum is equal to 1) are shown.

The results indicate that the performance of our PBM method is at least comparable, and in some cases better than, previous methods. For 4 of the 11 TFs we tested (Esrrb, Sox2, Oct4 and Crx), our method improved over all other methods. In all other cases, the AUC of our method ranks within the top 4 (Supplementary Results in Additional file 1 and Additional file 1: Figure S1 therein). Notably, in such cases, none of the other methods consistently achieves the best AUC. The PWM-based and E-score based methods tend to work well for some cases (for example, BEEML PWM for Klf7 and Max E-score for Srf), but for others their performance is not as good. Overall, our PBM model has the highest average AUC over the 11 TFs tested (Supplementary Results in Additional file 1, and Additional file 4).

To further assess the benefits of using PBM to derive TF binding specificities, we use two other collections of PWMs for AUC evaluation. The HOMER PWMs [31] were derived from ChIP-seq datasets, while the JASPAR PWMs were from multiple sources (including ChIP-seq, literature curation and PBM data). Overall, the results from the JASPAR PWMs are very similar to those obtained from the Seed and Wobble PWMs, and both are weaker than our method. The use of HOMER PWMs lead to better overall performance. However, given that HOMER trains the PWMs from ChIP-seq data and the same datasets may be used for evaluation, this is clearly not a fair comparison. We did notice that for some TFs, our method outperforms HOMER PWMs (e.g. Max: 0.809 vs 0.757). The full details are shown in Supplementary Results in Additional file 1 and Additional file 4.

Integrated model of PBM and DNase I hypersensitivity data significantly improves TFBS prediction accuracy

PBM data, although powerful, only measures *in vitro* binding. Therefore, even when using sophisticated methods, the ability to predict *in vivo* binding based on PBM data is limited. DNase I hypersensitive (HS) sites are regions of chromatin that are very sensitive to DNase I cleavage [32], and previous studies have shown that such hypersensitivity correlates with TF binding [21,22]. To better predict tissue-specific *in vivo* binding sites, we developed PIPES, a probabilistic graphical model for integrating DNase I HS data with PBM data. For windows containing a 36bp genomic region ("*site*"), we assume the chromatin of the site could exist in two states: *open* or *closed*, and that only in the open state the chromatin is accessible to binding by a TF. We infer the chromatin state by using a mixture model for the DNase HS data: the open state should be associated with higher tag densities from the DNase data, and the closed state with lower densities. The *in vivo* occupancy of a site is then estimated as the probability of binding *in vitro* estimated using result from the PBM model multiplied by the probability that the site

is in an open state inferred from the DNase HS data (see Methods for a detailed description of PIPES).

Figure 3 presents the AUCs from applying PIPES to predict *in vivo* TF binding in the corresponding tissues for the same 11 TFs studied in the previous section. Compared with using PBM data alone (black bars), the incorporation of DNase I HS data in the corresponding tissues (green bars) improves performance for 10 of the 11 TFs. Overall, the improvement of AUC from adding DNase HS data across all TFs is statistically significant ($p = 0.0049$, one sided Wilcoxon rank-sum test). The biggest improvements are seen for TFs for which the results when using only PBM data are relatively poor. For example, when predicting Srf binding sites in heart, even the best methods analyzed above achieve an AUC only slightly better than random (Figure 3). By integrating PBM and DNase data, the performance of our method is improved by 46% from 0.539 to 0.787 (Figure 3). Similar improvement is also observed for Oct4 (from 0.532 to 0.847). As another baseline, we evaluated the AUCs from using DNase HS data alone. Somewhat surprisingly, this feature alone seems quite discriminative (Figure 3, red bars; see also Additional file 4): the mean AUC is 0.822. Nevertheless, in 8 out of 11 cases, our model using both DNase and PBM data improves these baseline results and its mean AUC is also higher at 0.866. We also point out that in practice it is not appropriate to use the HS data alone to predict binding sites for a TF as the predictions would not be specific to the TF of interest.

In addition to DNase I hypersensitivity data, *bona fide* TF binding sites are usually under evolutionary pressure and therefore more conserved [33,34]. We thus further extended PIPES to incorporate phastCons scores [35] for each site (Methods). Performance of the full model that incorporates PhastCons information is shown in Figure 3 (purple bars). As can be seen, while in some cases adding the conservation information very slightly improves performance (for example for Srf and Oct4), overall using conservation data does not lead to a significant improvement in prediction accuracy. When the DNase HS data is not available, using PhastCons in addition to the PBM data provides a slight improvement of the AUCs from using PBM alone (improving the results for 8 out of 11 TFs, Additional file 4). The average AUC is increased by 1%, and for some TFs, the improvement can be quite significant (e.g. Srf, AUC changes from 0.539 to 0.592 by adding PhastCons).

Finally, we compare PIPES with recent methods proposed for integrating DNase and motif information to predict TFBS. The first method we compare against, termed 'Intersection", was used by Neph et al. [26]. This method intersects sites that have high-scoring PWM matches for a TF with DNase HS sites to predict in vivo binding (Supplementary Methods in Additional file 1). We also compare PIPES with CENTIPEDE [24]. CENTIPEDE uses a probabilistic model to integrate the prior information of putative sites, such as sequence conservation and matches to PWMs, with the epigenetic data to predict binding sites. While probabilistic, CENTIPEDE does, however, rely on a stringent cutoff for PWM match scores to achieve low false positive rates.

The results are presented in Figure 3. As can be seen, the intersection method leads to AUCs that are significantly lower than the ones obtained by our method for all 11 TFs (cyan bars, $p = 9.77 \cdot 10^{-4}$, one sided Wilcoxon rank-sum test), indicating that strict cutoffs (as opposed to probabilistic integration) may lead to a high rate of false negatives. Similarly, using the default settings for

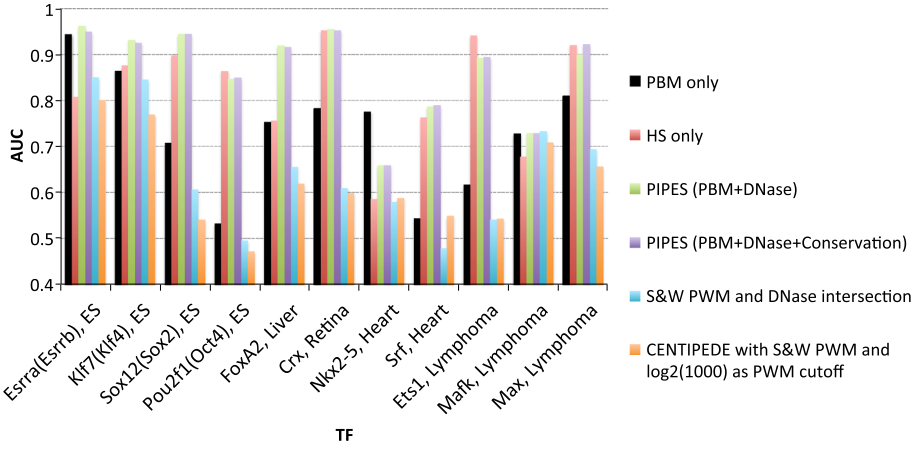

Figure 3 AUCs of different integrative methods for predicting *in vivo* TF binding. The first four bars are from our *k*-mer PBM data analysis method (black), simple baseline using only DNase data (red), or from PIPES (green and purple). Cyan bars: Simple overlapping approach that overlays PWM matches with DNase data. Orange bars: CENTIPEDE using S&W PWMs to scan sequences and an odd log score of $\log_2(1000)$ as cutoff. See main text and also Additional file 1 for more details.

CENTIPEDE led to AUC scores that are much lower than ours (Figure 3, orange bars). To further explore this, we varied the setting of CENTIPEDE, including using different PWMs and a range of cutoffs for defining putative binding sites, but the results remained the same (See Supplementary Methods and Results in Additional file 1, and Additional file 4). The difference in AUCs is highly significant: $p = 4.88 \cdot 10^{-4}$ (one sided Wilcoxon rank-sum test) using the best combination of PWM and cutoff for CENTIPEDE. These results indicate that our PIPES model, which relies on *k*-mer based representation and avoids strict cutoffs, can improve *in vivo* predictions of TFBS.

We performed additional analysis to the integration model. For these, we replace the binding probabilities predicted from the *k*-mer model, with those predicted when using PWMs. Three different versions of PWMs were used: the Seed-and-Wobble PWMs and RAP PWMs learned from the same PBM data, and the JASPAR PWMs. In all cases, the AUCs are substantially higher than the ones from the Intersection method and CENTIPEDE (below or close to the full PIPES model). These results indicate that the probabilistic integration step alone is enough to improve upon prior methods (Additional file 4). The usefulness of the *k*-mer based analysis provides additional advantage, as independently demonstrated in the earlier section.

Combining PBM and DNase data enables the prediction of tissue-specific TF activities

The recently released mouse ENCODE project data provides DNase I hypersensitivity data for more than 50 mouse tissue/cell types (Methods). We set out to combine the PBM data for 284 mouse TFs in UniPROBE with such DNase data to predict tissue-specific TF targets and determine tissue-specific TF activities (Methods).

Identifying TFs that are highly active in specific tissues is useful for determining the function of such TFs, and serves as an initial step for reconstructing the tissue-specific transcriptional regulatory networks. We predict how likely a TF is functional in any given tissue/cell type with an activity score for each TF-tissue pair. Our hypothesis is that if the TF is active in a tissue, it will bind a number of target sequences, thus the putative TF binding sites will be overrepresented in the DNase HS regions (see Methods for details). A higher activity score indicates that the TF is more active in the corresponding tissue (the expected value is 1 for non-active TFs). We use a binomial test to assess the statistical significance of the activity scores (see Methods). The complete results, including activity scores and *p*-values, are provided in Additional file 5. In Figure 4a we illustrate these results by focusing on the activity scores calculated for 4 TFs (Gata3, Pou6f1, Crx and Hnf4a) across 18 representative tissue/cell types. Gata3 is known to function in mouse fetal liver haematopoiesis [36], and its expression had also been observed in leukemia cells [37]. Our results are in good agreement with the prior knowledge regarding Gata3's activity: the top two tissues predicted for Gata3 are E14.5 liver cells and the adult leukemia cell line. Similarly the top tissue for Pou6f1 is E14.5 whole brain, in agreement with its known role in brain development [38]. Crx is an important TF for regulating photoreceptor genes in retina [39,40], and our method correctly determined that its activity score in that tissue is the highest. Finally, Hnf4a is a well known master regulator of liver- and kidney-specific genes [41,42], as correctly predicted by our method. While we only show 18 tissues, for all four TFs the correct tissues shown in Figure 4a have the highest scores among all 55 tissues we tested (Additional file 5).

To more globally validate these tissue-specific TF activities, we compared the correlation between our predicted TF activity scores and mRNA levels for the same TFs in the corresponding tissue (measured by qRT-PCR [43]) . Eight tissues and 222 TFs that are common to both datasets are used (Additional file 6). Even though the two types of data (PBM and DNase vs. expression) measure completely different aspects of cellular activity, we observe a Pearson correlation coefficient of 0.229, which is highly statistically significant ($p < 10^{-8}$, permutation test, Figure 4b). Since many TFs are only post-transcriptionally regulated, such a significant correlation provides strong support to the predictions computed by our method.

Existing literature strongly supports predicted TF activities in several tissues

To further validate our predictions and investigate their potentials to lead to new biological insights, we took a closer look at the TFs predicted to be active in the adult liver tissue. The top five such predictions are shown in Table 1A ($p < 10^{-100}$ for all, binomial test). Besides Hnf4a discussed above, Rara, Nr2f2, Rxra and Tcf7 are all known to either regulate liver-specific genes or are involved in maintaining liver metabolism and homeostasis (Table 1A). The 7*th* ranked factor Tcf7l2 (activity score of 1.38) was linked to type 2 diabetes risk in previous studies using SNP data [44], but the mechanism for its involvement was unclear. Our result indicates that it may have a regulatory role in liver metabolism. Indeed, a very recent study confirms its role in regulating key liver-specific metabolic genes [45]. Our result also assign a high liver activity score to Cutl1 (1.36, rank 8/284). Cutl1 was a known transcriptional repressor of terminal differentiation genes in several cell lineages including hepatocyte [46]. Recently, Cutl1 was identified as target of the liver-specific microRNA miR122 and a central mediator of the effects caused by the deregulation of miR122 in hepatocellular carcinoma [47]. Further down

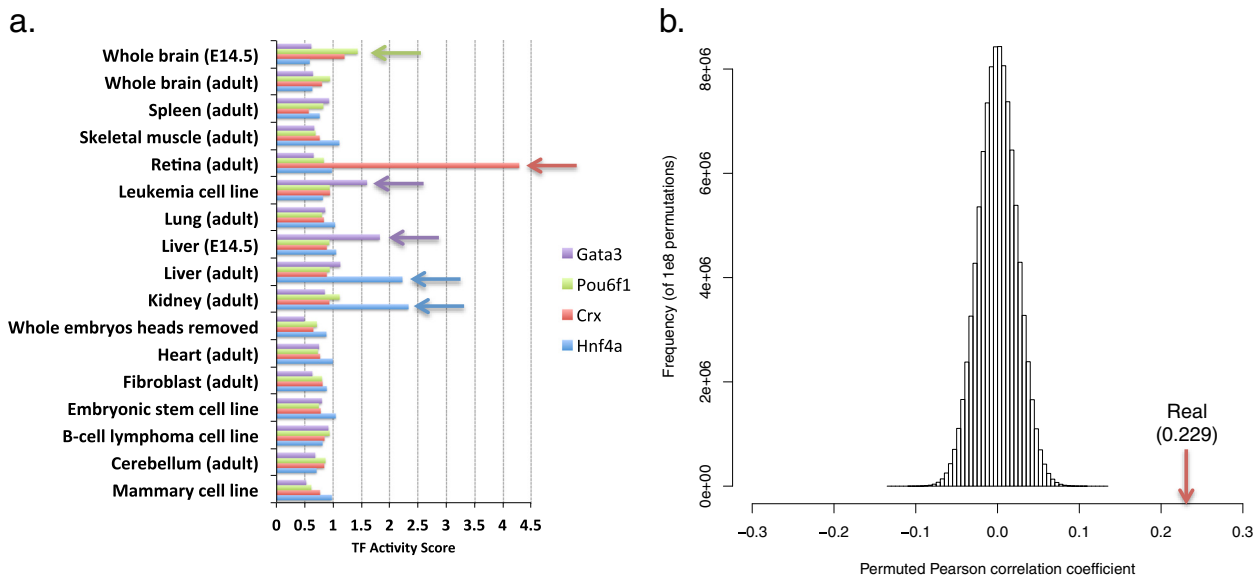

Figure 4 Results for tissue-specificity TF activity prediction. (a) Predicted tissue activity scores for 4 TFs across 18 representative tissue/cell types. Arrows indicate known functions of the TF in the corresponding tissue as supported by literature evidence. In all cases, the highest activity score matches a known tissue for the factor. **(b)** Pearson correlation coefficients between tissue specific expression experiments and the activity level predicted by our method. The distribution is based on 10^8 permutations of the activity scores. The value from the real predictions is indicated by the arrow on the right.

the list, Foxa2 (1.32, rank 10/284) is known to regulate lipid metabolism and ketogenesis related genes in liver [48], and Lef1 (1.31, rank 11/284) is a prognostic biomarker for liver metastasis in colorectal cancer. Other TFs ranked within the top 20 for liver include Tcf1 and Tcf2, members of the T-cell factor (Tcf) family that are critical for hepatocyte metabolism and function [49,50]; Bhlhb2, which is involved in the regulation of lipogenesis in liver [51]; and Hmbox1, whose expression levels was shown to be reduced in liver cancer compared with surrounding normal tissues [52]. Overall, our predicted set of liver regulators is comprehensive, spanning several different classes of liver related activities including glucose and lipid metabolism and cancer, and including both repressors and activators. In addition, Table 1B and C presents the top 5 predicted TFs for two more tissues (retina and B cell, $p < 10^{-100}$ for all, binomial test). As can be seen, for almost all of these TFs there is strong support for their tissue-specific activity in the predicted tissue.

Table 1 Top five predicted TFs for liver, retina and B cell

TF	Score	Known functions in the corresponding tissue
A. Liver		
Hnf4a	2.22	Essential for maintaining hepatic gene expression and lipid homeostasis [41]
Rara	1.90	Important in maintaining liver homeostasis, and its disruption is linked to hepatocarcinogenesis [57]
Nr2f2	1.56	Expressed in liver, and known to regulate liver-specific genes [58]
Rxra	1.45	Important role in liver metabolism [59]
Tcf7	1.44	Downstream regulator in Wnt signaling which is critical in liver physiology and pathology [60]
B. Retina		
Crx	4.28	Regulates photoreceptor gene expression [40]
Pitx3	4.26	Required for normal retina formation in Xenopus and zebrafish [61,62]
E2F3	4.06	Involved in retina progenitor cell development [63]
Pitx2	3.92	Pitx2-deficient mouse exhibits ocular abnormalities [64]
Gsc	3.89	Unknown function in retina.
C. CD19+ B cell		
Sfpi1	2.09	Essential regulator of B-cell differentiation [65]
Pou2f2	2.08	Required for T-cell independent B cell activation [66]
Spic	1.98	Promotes B cell differentiation [67]
Pou2f3	1.94	Unknown function in B cell, but has almost the same binding preference as Pou2f2
Elf4	1.70	Regulates proliferation of B cells [68]

Discussion

A number of recent projects including ENCODE [53], modENCODE [54,55] and the Roadmap Epigenomics Project [56] have generated large amounts of genomic data. An important research goal is to translate these resources into accurate, tissue and condition-sensitive, molecular-level networks. Constructing tissue-specific

maps of TF binding sites is a central part of this overall research effort. Using several different datasets and a novel computational strategy, PIPES, we demonstrated that such high-quality computational predictions can be obtained. We used PIPES to compile a resource that includes comprehensive predictions for more than 200 TFs across 50 tissues.

A recent benchmark study that compared many methods for analyzing PBM data concluded that PWM-based methods work as well as other models for predicting TFBS [69]. Our results differ from these previous studies. This could either be the result of the Lasso based method we have used or the specific dataset we used for the comparison. Additional work is required to reach a definitive conclusion regarding the importance of independence assumption used by PWMs when modeling TFBS. Here we focused on integrating a number of datasets for predicting TFBSs. For our biophysical approach, using a *k*-mer based method allowed us to capture both the dependency among positions within the binding site as well as multiple different motifs for a single TF. Such method worked well in classifying bound and unbound sequences from in vivo ChIP-seq data for many TFs. Moreover, integrating PBM data with chromatin accessibility from DNase I HS data greatly improves the accuracy of TF binding predictions. Several recent papers explored related ideas. Chromia [70] used a hidden Markov model to combine sequence-specific TF binding with histone modification data, but their predictions were based on PWM scoring and only focused on a dozen of TFs in mouse embryonic stem cells. Ernst et al. [23] combined experimental data from a number of tissues to generate a single (global) TF-target prediction map. However, that method has also relied on PWMs and no tissue specific predictions were made. CENTIPEDE [24] used unsupervised methods to integrate TF-DNA interaction, epigenetic and evolutionary data, and is most similar to our efforts. However, CENTIPEDE relies on a footprint in the DNase data that TFs leave. Such DNase footprints are the actual locations where the TF binds and are therefore protected from DNase cleavage within the DNase HS site. Unfortunately, DNase footprint data is expensive to obtain (indeed, it was not available for most of the tissues we analyzed) since it requires very high coverage when sequencing. In addition, CENTIPEDE uses a stringent cutoff (based on PWM matching) to define putative binding sites, and thus may lose significant information in relatively weak binding sites, which have been shown to be collectively important for TF binding [71]. Neph et al. [26,72] also combined DNase footprints with PWMs to predict TF-TF interactions (though not TF-gene interactions) across a large number of human tissues, using a simple method to intersect motif matches with DHS sites. Such hard cutoffs may miss sites that score high (but just below the cutoff) for both types of data which are found by our method.

In general, we find that binding sites for the TFs we looked at are only modestly conserved when compared with controls (using PhastCons scores alone classifies ChIP-seq sequences quite poorly, see Additional file 4). This is largely consistent with the recent findings that functional non-coding sequences evolve rather rapidly [9]. As a result, adding PhastCons in the integrated model does not lead to improvements in AUC values. Nevertheless, there are a number of advantages for models that can incorporate sequence conservation. First, when DNase data is not available, adding conservation in the model leads to slight improvement over models that only use PBM data (Additional file 4). Second, the conservation of binding events can vary greatly among tissues. For example, enhancers in brain are far more constrained than those in the heart [73]. Thus it is quite possible that sequence conservation would be more informative for other ChIP-seq studies.

The application of PIPES to predict tissue specific TFBS led to results that agree well with existing knowledge regarding TF roles in specific tissues. The overall results are significantly correlated with independent gene expression data measured for these TFs across tissues even though such expression data was not used at all in our analysis.

Several extensions of our current work are possible. Recently Jiang et al. [74] reported the interesting phenomenon of sticky *k*-mers: these are *k*-mers that appear to bind to TFs with relatively high affinities in a large number of PBM experiments. The sticky *k*-mers likely represent background noises in the PBM experiments and an interesting research direction is to expand our regression method to remove such noises. In another recent study, Ballare et al. [75] reported that functional TFBSs are not always associated with high chromatin accessibility, an assumption implicitly made by us and other related methods. Rather, nucleosomes may occupy TFBSs at basal conditions, and are only remodeled or displaced upon change of cellular conditions (e.g. by hormone stimulation). Despite this unexpected relationship between TFBS and chromatin states, the paper does report that such sites, while associated with high nucleosome occupancy before stimulation, often overlap with DNase HS sites. It remains to be seen how common such cases are and what is the impact on methods such as ours that rely on DNase data to predict condition specific TF binding. Moreover, our integrative framework for utilizing additional information sources when predicting binding events on a genome wide scale could also be used for large scale comparison of different PWM methods and methods that use more complicated models to represent

TF binding preferences [76-78]. This requires a detailed study and is left for future work.

Conclusions

Combining PBM and DNase data, we presented the first major effort to provide a systematic map of computationally predicted tissue-specific targets for hundreds of TFs across a large number of tissues in mouse. We complied a resource that provides TF-target predictions for all 284 TFs studied across the 55 tissue/cell types (Supplementary Methods and Supplementary Website in Additional file 4). We believe that such comprehensive resource would be useful for both biology and computation-oriented researchers studying gene regulation [79-82].

Methods

K-mer based method that uses PBM data to predict TF binding

The binding specificities of TFs are often represented by position weight matrices, which assume that each position of a site contributes independently to the overall binding affinity of the site (independence assumption). The PBM data simultaneously measures binding of a TF to tens of thousands of probes, and can be used to construct a much more detailed and accurate model of TF binding specificities.

A biophysically-motivated model for PBM data

Our k-mer based PBM model (Figure 1a) is motivated by the biophysics of TF binding to the probes in PBM experiments. Following Zhao et al. [16], we denote by Y_i the experimentally measured intensity of the i-th probe on the PBM array. We denote by $F(i)$ the (unobserved) binding probability of the TF to this probe. While these two quantities are related, due to experimental errors and scaling they are not identical. We thus assume a simple linear model for the mapping between the two:

$$Y_i = a + cF(i) + \epsilon_i \tag{1}$$

where a and c are constants, and ϵ_i is the error term. Since each probe is much longer than the motif itself (probe length is 36bp while motifs are generally less than 20bp with a typical length at 12bp [83]) we follow BEEML-PBM [16], and express the binding probability $F(i)$ as the sum of the binding probabilities over all k-mers in the probe. Let k be the length of a TF binding site and L be the length of the variable region on the probe, we have:

$$F(i) = \sum_{j=1}^{L-k+1} \lambda_j \cdot \beta_{S_i(j)} \tag{2}$$

where λ_j is the position effect at position j (see Supplementary Methods in Additional file 1) and $\beta_{S_i(j)}$ is the binding probability to $S_i(j)$, the k-mer at the j-th position of the i-th probe. The term β_s is symmetric for any k-mer s, i.e. $\beta_s = \beta_{\bar{s}}$, where $\bar{s}$ is the reverse complement of s. We note that our model allows features/k-mers to overlap, so the k-mers can be of different lengths and can contain gaps. The model relies on Lasso regression to estimate the contribution of each k-mer (see below).

By plugging in the equation of $F(i)$ into the linear model, we can couple the different values we obtain for probe intensities as a function of the individual k-mer contributions, see Supplementary Methods in Additional file 1 for full details.

Learning the parameters of the linear model

The above linear model has approximately $4^k/2$ parameters (one for each k-mer and its reverse complement). To avoid overfitting, we use the lasso regression [84] to estimate the coefficients. Lasso is a widely used approach to linear regression that encourages a sparse model where most of the coefficients are zero. In our problem, we have the additional requirement that the coefficients must be non-negative, and this is known as positive lasso [84] (Supplementary Methods in Additional file 1). After learning the model parameters using positive lasso, we set $\beta'_s = \beta_s / \max_s \beta_s$ to be the binding probability of the TF to the k-mer s up to a scaling constant.

Since we do not know the width of the motif bound by each TF, our method searches for k-mers of different lengths. In order to speed up the calculation, we first run positive lasso using all short 4–6 mers. To allow longer k-mers to be considered, after the first run, all pairs from the top 100 such k-mers (based on regression coefficients) are tested to see if the prefix of one matches the suffix of the other, yielding longer $(k+1)$-mers. This process is repeated until up to 8-mers have been added to the feature set. In addition, we also allow for gapped k-mers to be considered (Supplementary methods in Additional file 1).

Predicting TF binding to any sequences

Our model is trained on sequences of 36bp in length (the length of the variable region of probes in PBM experiments), however, in practice, we often need to predict TF binding to longer sequences, e.g. promoter regions up to thousands of base pairs long. Searching for binding sites in long sequences often involves sliding windows with relative small size so that signals are not diluted over long regions. For simplicity, we use 36bp as our window size; otherwise, additional normalization would be needed for the PBM scores trained from 36bp probes. To predict the binding of a TF to a longer sequence (Figure 1b), we first define the binding probabilities of the TF to each overlapping 36bp region ("*site*") in that sequence:

$$B = \frac{1}{B_{\max}} \sum_{s \in \Sigma^k} \beta_s C_s \tag{3}$$

in which β_s is the binding probability to the k-mer s learned by the regression model, C_s is the number of times that s occurs in this site, and $B_{\max}$ is the highest possible unscaled binding probability of any 36-mer that can be achieved for the TF and is used as a scaling constant. The interpretation of this equation is that the binding probability to a 36bp sequence is the sum of binding probabilities to each of the k-mer of the 36bp sequence [85,86]. In practice, $B_{\max}$ is estimated, for each TF, from the highest unscaled binding probabilities to 100,000 randomly sampled 36bp sites. Then, the binding probability to the entire sequence is defined as the highest binding probability to any 36bp site in that sequence.

Integrated model of TF binding *in vivo*

PBM experiments measure TF binding *in vitro*. *In vivo* binding depends on several factors including the cellular environment and the chromatin state of the bound region. In addition, it has been shown that functional TFBSs tend to be evolutionary constrained [33,34]. In this section, we describe PIPES, a method that integrates our PBM motif learning and scanning method with these additional data sources in order to determine tissue specific binding.

Incorporating DNase I hypersensitivity data

Let B_i be the probability of binding of the TF of interest to a 36bp genomic region (a *site*) indexed by i based on the PBM model (Equation 3), reflecting the potential of TF binding *in vitro*. We are interested in the *in vivo* occupancy of the site, denoted as X_i. We assume that X_i is influenced by the chromatin state, which can be represented as a simple binary indicator variable, A_i (it is 1 if the chromatin is open/accessible and 0 otherwise). When the chromatin is open ($A_i = 1$), the occupancy X_i equals B_i; whereas a closed chromatin at that location means that $X_i = 0$. Thus, X_i is simply the product of B_i and $P(A_i = 1)$. The chromatin state variable can be partially determined using experimental data. Here we use DNase I hypersensitivity (HS) data (Figure 1c) which is available for several mouse and human tissues (Results). See Supplementary Methods in Additional file 1 for details.

The full integrated model

To further incorporate the conservation data into PIPES (Figure 1e), we consider the following graphical model:

$$X_i \leftarrow Z_i \rightarrow C_i \rightarrow S_i \tag{4}$$

Here X_i is the occupancy of site i as described above, Z_i is a binary variable indicating whether site i is a true binding site *in vivo* or not, C_i is a binary variable indicating whether site i is conserved or not, and S_i is a measure of the evolutionary conservation of the site. The model assumes that true TFBSs have a higher occupancy. Similarly, when $Z_i = 1$, C_i is more likely to be 1 as well (a true binding site is more likely to be conserved), and this is reflected by a higher conservation score S_i. The goal is to infer Z_i from the observed data X_i and S_i. The evolutionary conservation measure we used is the phastCons score [35] (phastCons 46way vertebrates) downloaded from the the UCSC Genome Browser (http://genome.ucsc.edu).

In Supplementary Methods in Additional file 1 we discuss the specific distributions we assume for each of the conditional probabilities in our model and how we learn the parameters for these distributions. After these parameters are estimated, we can compute the probability that the i-th site is bound by a specific TF. See Supplement for details.

Identifying tissue-specific TF activities

We used PIPES to identify TFs likely to be active in each tissue. Intuitively, if a TF f is active in a tissue T, then the binding sites of f should be overrepresented in the open chromatin regions of T. To quantify this overrepresentation, we define $R(f, T)$ as the fraction of DNase hypersensitive sites in tissue T that contain high-scoring binding sites of f. The high scoring sites are defined as those that have binding probabilities (according to the PBM model, as defined in Equation 3) higher than the top 0.1% of the binding probabilities for all possible sites for that TF (the exact percentage cutoff has little impact, data not shown). In practice the binding probability distribution of a TF is estimated from the 100,000 sampled sites. For each tissue, the open sites in the promoter regions are defined as those sites whose DNase tag densities are higher than 15. This threshold is chosen so that the inferred probability of chromatin being open is close to 1 according to our model. In order to identify TFs having active functions in specific cell types, we exclude the binding sites that are not tissue specific (defined as open in more than 1/3 of all tissues). Such broadly-active sites are not interesting for the purpose of finding tissue-specific TFs.

The activity score of a TF f in tissue T is defined as:

$$Activity\,(f, T) = \frac{R(f, T)}{R(f, \bar{T})} \tag{5}$$

where $\bar{T}$ denotes all tissues other than T. This is used as a measure of the likely activity of the TF in that tissue. We use a simple binomial test to evaluate the significance of the activity score defined here. Suppose we observe n high-scoring binding sites of the TF f in the tissue, and among these n sites, x sites fall into DNase hypersensitive regions. By chance, the expected fraction of binding sites in the DNase HS regions is $p = R(f, \bar{T})$, thus we perform the one-sided binomial test of x successes in n trials under the null model that the probability of success is equal to p.

A comprehensive collection of predicted TFBSs across 55 mouse tissues

PBM data for 284 mouse TFs were downloaded from UniPROBE [27]. DNase data for 55 mouse tissue/cell types were downloaded from the mouse ENCODE website at http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDnase/, and for each tissue/cell type, parameters were learned as described in Methods. A list of all the TFs and tissue/cell types is provided in Additional files 7 and 8. To predict targets of TFs, the promoter regions (+/- 10kb around transcription start sites) for all mouse genes were scanned. This choice of promoter regions is consistent with several recent publications [23,70]. We assessed the quality of our predictions in two ways. (1) The false discovery rates of top TF/tissue combinations were assessed, and in all but one of 25 combinations we examined, the FDR is below 15%. (2) For each of the 11 TFs we evaluated in the Results, we compared the set of predicted targeted genes from our method with the genome-wide targeted genes from the mouse ENCODE ChIP-seq datasets. The overlap of the two sets are highly significant for almost all TFs. See Supplementary Methods and Results in Additional file 1, Additional files 9, 10 and Supplementary Website for details.

Comparison with other methods

To evaluate our methods we obtained ChIP-seq data for 11 TFs for which PBM data and tissue specific DNase I hypersensitivity data were available (Additional file 2). ChIP-seq data was downloaded from NCBI GEO or ENCODE using the GEO IDs or UCSC Accession IDs listed in Additional file 2.

We performed a comprehensive comparison of our PBM model with several other methods that could be or have been used in predicting TF binding on real sequences. Since most prior methods relied on PWMs, we used the PWMs reported in UniPROBE [27] for these TFs, which were obtained by applying the Seed-and-Wobble algorithm on the PBM data (S&W PWM) [11]. We also compared with BEEML [16] using both the energy matrices (BEEML Energy) and converted PWMs (BEEML PWMs), PWMs identified by RAP (RAP PWMs) [87], the max E-score of k-mers (Max E-score) [11], the use of occupancy score proposed by [12], a support vector regression-based method (SVR) [88], and FeatureREDUCE (unpublished). In addition, we also compared using PWMs from external sources including the JASPAR database [89] and PWMs derived from HOMER on ChIP-seq data [31]. Moreover, we also compared our integrative model with an intersection strategy that combines PWM scanning with DNase data [72] across multiple tissues, a simple method that combines PWM scanning with our DNase model, and an integrative method CENTIPEDE [24] that uses PWMs, DNase HS and conservation data. A description of the details for the settings of all methods is provided in Supplementary Methods in Additional file 1.

Availability of supporting data

The genome-wide tissue-specific TFBS predictions for 284 mouse TFs and 55 tissue/cell types and codes for the k-mer based PBM modeling method are available from the supporting website at http://www.sb.cs.cmu.edu/PIPES.

Additional files

Additional file 1: Supplementary methods, results and figures. This file contains Supplementary Methods, Results and Figures.

Additional file 2: List of TF and ChIP-seq experiments used in evaluation. This file contains a list of information about the 11 TFs and corresponding ChIP-seq experiments used in the evaluation.

Additional file 3: Detailed AUCs of different methods using alternative negative sequence sets and background models (PWM-based methods). This file lists the detailed AUCs for methods using PBM data alone, obtained when alternative negative sequence sets or background models are used. See Supplementary Methods in Additional file 1 for details.

Additional file 4: Details of the AUCs for comparing different methods that predict in vivo TF binding. This file contains the AUCs of different methods that classify the positive and negative sequences from ChIP-seq experiments. Methods based on PBM data alone include S&W PWMs, BEEML PWMs, BEEML Energy, RAP PWMs, Max E-score, Occupancy score, SVR, FeatureREDUCE and our method. Methods that use external PWMs include JASPAR PWMs and HOMER PWMs. Methods based on integrative modeling that also use DNase and phastCons data include our integrative models (without and with phastCons data), baselines for our integrative models (DNase alone, phastCons alone and PBM+phastCons), simple overlapping primary S&W PWM matches with DNase, combining S&W, RAP or JASPAR PWMs with our DNase models, and CENTIPEDE under different settings. See text and Additional file 1 for details.

Additional file 5: Full list of the predicted tissue-specific activity score for all TFs and tissues. This file lists the predicted activity scores and binomial test p-values for all the 284 TFs across the 55 cell/tissue types.

Additional file 6: List of TFs and tissues with mRNA measurement data available from [43]. This file lists the TFs and tissues that have mRNA expression data for the corresponding TF available.

Additional file 7: List of all 55 tissues studied. This file lists the 55 mouse tissue/cell types studied. DNase I hypersensitivity data for these tissues were downloaded from the UCSC Genome Browser.

Additional file 8: List of all 284 TFs in mouse with PBM data studied. This file lists the 284 TFs in mouse studied with PBM data available.

Additional file 9: Overlap of genome-wide predictions with ChIP-seq data at gene level. This file provides the number of overlapping genes in the genome-wide predictions with ChIP-seq data for the 11 TF/tissues used in the evaluations. See Supplementary Methods and Results in Additional file 1 for details.

Additional file 10: FDR estimates. This file provides the estimated false discovery rates for the genome-wide predictions of the 11 TF/tissues used in the evaluation and the 15 TF/tissues with highest activity scores. See Supplementary Methods and Results in Additional file 1 for details.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SZ collected and processed the data. SZ, XH and ZBJ designed the experiments, analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information
Shan Zhong and Xin He: Co-first author.

Acknowledgements
Work supported in part by NIH grant 1RO1 GM085022 and NSF DBI-0965316 award to Z.B.J.

Received: 30 October 2013 Accepted: 6 November 2013
Published: 15 November 2013

References

1. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, MacIsaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**(7004):99–104.
2. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133**(6):1106–1117.
3. Buck MJ, Lieb JD: **ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.** *Genomics* 2004, **83**(3):349–360.
4. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** *Nat Rev Genet* 2009, **10**(10):669–680.
5. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**(5500):2306–2309.
6. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316**(5830):1497–1502.
7. Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M: **Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo.** *Genes Dev* 2007, **21**(4):385–390.
8. Kaufmann K, Muiño JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC: **Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower.** *PLoS Biol* 2009, **7**(4):e1000090.
9. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.** *Science* 2010, **328**(5981):1036–1040.
10. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML: **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.** *Nat Genet* 2004, **36**(12):1331–1339.
11. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities.** *Nat Biotechnol* 2006, **24**(11):1429–1435.
12. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML: **High-resolution DNA-binding specificity analysis of yeast transcription factors.** *Genome Res* 2009, **19**(4):556–566.
13. Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJM: **A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors.** *Cell* 2009, **138**(2):314–327.
14. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML: **Diversity and complexity in DNA recognition by transcription factors.** *Science* 2009, **324**(5935):1720–1723.
15. Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, Huang SsC, Schmitz RJ, Urich MA, Kuo D, Nery JR, Qiao H, Yang A, Jamali A, Chen H, Ideker T, Ren B, Bar-Joseph Z, Hughes TR, Ecker JR: **Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis.** *eLife* 2013, **2:**e00675.
16. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity.** *Nat Biotechnol* 2011, **29**(6):480–483.
17. Orenstein Y, Linhart C, Shamir R: **Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data.** *PLoS ONE* 2012, **7**(9):e46145.
18. Maerkl SJ, Quake SR: **A systems approach to measuring the binding energy landscapes of transcription factors.** *Science* 2007, **315**(5809):233–237.
19. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordan R: **Stability selection for regression-based models of transcription factor-DNA binding specificity.** *Bioinformatics* 2013, **29**(13):i117–i125.
20. Spitz F, Furlong EEM: **Transcription factors: from enhancer binding to developmental control.** *Nat Rev Genet* 2012, **13**(9):613–626.
21. Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD: **The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding.** *Genome Biol* 2011, **12**(4):R34.
22. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: **Chromatin accessibility pre-determines glucocorticoid receptor binding patterns.** *Nat Genet* 2011, **43**(3):264–268.
23. Ernst J, Plasterer HL, Simon I, Bar-Joseph Z: **Integrating multiple evidence sources to predict transcription factor binding in the human genome.** *Genome Res* 2010, **20**(4):526–536.
24. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.** *Genome Res* 2011, **21**(3):447–455.
25. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL: **Epigenetic priors for identifying active transcription factor binding sites.** *Bioinformatics* 2012, **28:**56–62.
26. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA: **Circuitry and dynamics of human transcription factor regulatory networks.** *Cell* 2012, **150**(6):1274–1286.
27. Newburger DE, Bulyk ML: **UniPROBE: an online database of protein binding microarray data on protein-DNA interactions.** *Nucleic Acids Res* 2009, **37**(Database issue):D77–D82.
28. Mouse ENCODE Consortium: **An encyclopedia of mouse DNA elements (Mouse ENCODE).** *Genome Biol* 2012, **13**(8):418.
29. Annala M, Laurila K, Lähdesmäki H, Nykter M: **A linear model for transcription factor binding affinity prediction in protein binding microarrays.** *PLoS ONE* 2011, **6**(5):e20059.
30. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108–D110.
31. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol cell* 2010, **38**(4):576–589.
32. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin.** *Annu Rev Biochem* 1988, **57:**159–197.
33. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals.** *Nature* 2005, **434**(7031):338–345.
34. So AYL, Cooper SB, Feldman BJ, Manuchehri M, Yamamoto KR: **Conservation analysis predicts in vivo occupancy of glucocorticoid receptor-binding sequences at glucocorticoid-induced genes.** *Proc Natl Acad Sci USA* 2008, **105**(15):5745–5749.
35. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15**(8):1034–1050.
36. Pandolfi PP, Roth ME, Karis A, Leonard MW, Dzierzak E, Grosveld FG, Engel JD, Lindenbaum MH: **Targeted disruption of the GATA3 gene**

causes severe abnormalities in the nervous system and in fetal liver haematopoiesis. *Nat Genet* 1995, **11:**40–44.
37. Minegishi N, Morita S, Minegishi M, Tsuchiya S, Konno T, Hayashi N, Yamamoto M: **Expression of GATA transcription factors in myelogenous and lymphoblastic leukemia cells.** *Int J Hematol* 1997, **65**(3):239–249.
38. Zhang L, Ju X, Cheng Y, Guo X, Wen T: **Identifying Tmem59 related gene regulatory network of mouse neural stem cell from a compendium of expression profiles.** *BMC Syst Biol* 2011, **5:**152.
39. Hennig AK, Peng GH, Chen S: **Regulation of photoreceptor gene expression by Crx-associated transcription factor network.** *Brain Res* 2008, **1192:**114–133.
40. Corbo JC, Lawrence KA, Karlstetter M, Myers CA, Abdelaziz M, Dirkes W, Weigelt K, Seifert M, Benes V, Fritsche LG, Weber BHF, Langmann T: **CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors.** *Genome Res* 2010, **20**(11):1512–1525.
41. Hayhurst GP, Lee YH, Lambert G, Ward JM, Gonzalez FJ: **Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis.** *Mol Cell Biol* 2001, **21**(4):1393–1403.
42. Lucas B, Grigo K, Erdmann S, Lausen J, Klein-Hitpass L, Ryffel GU: **HNF4alpha reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma.** *Oncogene* 2005, **24**(42):6418–6431.
43. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, Carninci P, Daub CO, Forrest ARR, Gough J, Grimmond S, Han JH, Hashimoto T, Hide W, Hofmann O, Kamburov A, Kaur M, Kawaji H, Kubosaki A, Lassmann T, van Nimwegen E, MacPherson CR, Ogawa C, Radovanovic A, Schwartz A, Teasdale RD, Tegnér J, Lenhard B, Teichmann SA, Arakawa T, Ninomiya N, Murakami K, Tagami M, Fukuda S, Imamura K, Kai C, Ishihara R, Kitazume Y, Kawai J, Hume DA, Ideker T, Hayashizaki Y: **An atlas of combinatorial transcriptional regulation in mouse and man.** *Cell* 2010, **140**(5):744–752.
44. Grant SFA, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K: **Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes.** *Nat Genet* 2006, **38**(3):320–323.
45. Boj SF, van Es, J H, Huch M, Li VSW, José A, Hatzis P, Mokry M, Haegebarth A, van den Born M, Chambon P, Voshol P, Dor Y, Cuppen E, Fillat C, Clevers H: **Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand.** *Cell* 2012, **151**(7):1595–1607.
46. Sansregret L, Nepveu A: **The multiple roles of CUX1: insights from mouse models and cell-based assays.** *Gene* 2008, **412**(1–2):84–94.
47. Kojima K, Takata A, Vadnais C, Otsuka M, Yoshikawa T, Akanuma M, Kondo Y, Kang YJ, Kishikawa T, Kato N, Xie Z, Zhang WJ, Yoshida H, Omata M, Nepveu A, Koike K: **MicroRNA122 is a key regulator of alpha-fetoprotein expression and influences the aggressiveness of hepatocellular carcinoma.** *Nat Commun* 2011, **2:**338.
48. Wolfrum C, Asilmaz E, Luca E, Friedman JM, Stoffel M: **Foxa2 regulates lipid metabolism and ketogenesis in the liver during fasting and in diabetes.** *Nature* 2004, **432**(7020):1027–1032.
49. Shih DQ, Bussen M, Sehayek E, Ananthanarayanan M, Shneider BL, Suchy FJ, Shefer S, Bollileni JS, Gonzalez FJ, Breslow JL, Stoffel M: **Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism.** *Nat Genet* 2001, **27**(4):375–382.
50. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA: **Control of pancreas and liver gene expression by HNF transcription factors.** *Science* 2004, **303**(5662):1378–1381.
51. Iizuka K, Horikawa Y: **Regulation of lipogenesis via BHLHB2/DEC1 and ChREBP feedback looping.** *Biochem Biophys Res Commun* 2008, **374:**95–100.
52. Dai J, Zhang C, Tian Z, Zhang J: **Expression profile of HMBOX1, a novel transcription factor, in human cancers using highly specific monoclonal antibodies.** *Exp Ther Med* 2011, **2**(3):487–490.
53. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57–74.
54. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, et al.: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.** *Science* 2010, **330**(6012):1775–1787.
55. The modENCODE Consortium: **Identification of functional elements and regulatory circuits by Drosophila modENCODE.** *Science* 2010, **330**(6012):1787–1797.
56. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA: **The NIH Roadmap Epigenomics Mapping Consortium.** *Nat Biotechnol* 2010, **28**(10):1045–1048.
57. Khetchoumian K, Teletin M, Tisserand J, Mark M, Herquel B, Ignat M, Zucman-Rossi J, Cammas F, Lerouge T, Thibault C, Metzger D, Chambon P, Losson R: **Loss of Trim24 (Tif1alpha) gene function confers oncogenic activity to retinoic acid receptor alpha.** *Nat Genet* 2007, **39**(12):1500–1506.
58. Zhang P, Bennoun M, Gogard C, Bossard P, Leclerc I, Kahn A, Vasseur-Cognet M: **Expression of COUP-TFII in metabolic tissues during development.** *Mech Dev* 2002, **119:**109–114.
59. Wan YJ, An D, Cai Y, Repa JJ, Hung-Po Chen T, Flores M, Postic C, Magnuson MA, Chen J, Chien KR, French S, Mangelsdorf DJ, Sucov HM: **Hepatocyte-specific mutation establishes retinoid X receptor alpha as a heterodimeric integrator of multiple physiological processes in the liver.** *Mol Cell Biol* 2000, **20**(12):4436–4444.
60. Thompson MD, Monga SPS: **WNT/beta-catenin signaling in liver health and disease.** *Hepatology* 2007, **45**(5):1298–1305.
61. Khosrowshahian F, Wolanski M, Chang WY, Fujiki K, Jacobs L, Crawford MJ: **Lens and retina formation require expression of Pitx3 in Xenopus pre-lens ectoderm.** *Dev Dyn* 2005, **234**(3):577–589.
62. Shi X, Bosenko DV, Zinkevich NS, Foley S, Hyde DR, Semina EV, Vihtelic TS: **Zebrafish pitx3 is necessary for normal lens and retinal development.** *Mech Dev* 2005, **122**(4):513–527.
63. Chen D, Pacal M, Wenzel P, Knoepfler PS, Leone G, Bremner R: **Division and apoptosis of E2f-deficient retinal progenitors.** *Nature* 2009, **462**(7275):925–929.
64. Gage PJ, Suh H, Camper SA: **Dosage requirement of Pitx2 for development of multiple organs.** *Development* 1999, **126**(20):4643–4651.
65. Sokalski KM, Li SKH, Welch I, Cadieux-Pitre HAT, Gruca MR, DeKoter RP: **Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia.** *Blood* 2011, **118**(10):2801–2808.
66. Corcoran LM, Karvelas M: **Oct-2 is required early in T cell-independent B cell activation for G1 progression and for proliferation.** *Immunity* 1994, **1**(8):635–645.
67. Schweitzer BL, Huang KJ, Kamath MB, Emelyanov AV, Birshtein BK, DeKoter RP: **Spi-C has opposing effects to PU.1 on gene expression in progenitor B cells.** *J Immunol* 2006, **177**(4):2195–2207.
68. Lacorazza HD, Miyazaki Y, Di Cristofano A, Deblasio A, Hedvat C, Zhang J, Cordon-Cardo C, Mao S, Pandolfi PP, Nimer SD: **The ETS protein MEF plays a critical role in perforin gene expression and the development of natural killer and NK-T cells.** *Immunity* 2002, **17**(4):437–449.
69. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S, DREAM5 Consortium, Agius P, Arvey A, Bucher P, Callan CG, Chang CW, Chen CY, Chen YS, Chu YW, Grau J, Grosse I, Jagannathan V, Keilwagen J, Kiełbasa SM, Kinney JB, Klein H, Kursa MB, Lähdesmäki H, Laurila K, Lei C, et al.: **Evaluation of methods for modeling transcription factor sequence specificity.** *Nat Biotechnol* 2013, **31**(2):126–134.
70. Won KJ, Ren B, Wang W: **Genome-wide prediction of transcription factor binding sites using an integrated model.** *Genome Biol* 2010, **11:**R7.
71. Tanay A: **Extensive low-affinity transcriptional interactions in the yeast genome.** *Genome Res* 2006, **16**(8):962–972.
72. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P,

Zhang M, Balasundaram G, et al.: **An expansive human regulatory lexicon encoded in transcription factor footprints.** *Nature* 2012, **489**(7414):83–90.
73. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al.: **ChIP-Seq identification of weakly conserved heart enhancers.** *Nature Genet* 2010, **42**(9):806–810.
74. Jiang B, Liu JS, Bulyk ML: **Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers.** *Bioinformatics* 2013, **29**(11):1390–1398.
75. Ballaré C, Castellano G, Gaveglia L, Althammer S, González-Vallinas J, Eyras E, Le Dily F, Zaurin R, Soronellas D, Vicent GP, Beato M: **Nucleosome-driven transcription factor binding and gene regulation.** *Mol Cell* 2012, **49**(1):67–79.
76. Mathelier A, Wasserman WW: **The next generation of transcription factor binding site prediction.** *PLoS Comput Biol* 2013, **9**(9):e1003214.
77. Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V: **From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites.** *J Bioinform Comput Biol* 2013, **11:**1340004.
78. Grau J, Posch S, Grosse I, Keilwagen J: **A general approach for discriminative de novo motif discovery from high-throughput data.** *Nucleic Acids Res* 2013. doi:10.1093/nar/gkt831.
79. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z: **Reconstructing dynamic regulatory maps.** *Mol Syst Biol* 2007, **3:**74.
80. Li H, Zhan M: **Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data.** *Bioinformatics* 2008, **24**(17):1874–1880.
81. Schulz MH, Devanny WE, Gitter A, Zhong S, Ernst J, Bar-Joseph Z: **DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data.** *BMC Syst Biol* 2012, **6:**104.
82. Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M: **Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks.** *Genome Res* 2012, **22**(7):1334–1349.
83. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152**(1–2):327–339.
84. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least angle regression.** *Ann Stat* 2004, **32**(2):407–499.
85. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.** *Bioinformatics* 2006, **22**(14):e141–e149.
86. He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH, Zhong S: **A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data.** *PLoS ONE* 2009, **4**(12):e8155.
87. Orenstein Y, Mick E, Shamir R: **Rap: Accurate and fast motif finding based on protein-binding microarray data.** *J Comput Biol* 2013, **20**(5):375–382.
88. Agius P, Arvey A, Chang W, Noble WS, Leslie C: **High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions.** *PLoS Comp Biol* 2010, **6**(9).
89. Bryne JC, Valen E, Tang MHE, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A: **JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.** *Nucleic Acids Res* 2008, **36**(Database issue):D102–D106.

doi:10.1186/1471-2164-14-796
Cite this article as: Zhong *et al.*: Predicting tissue specific transcription factor binding sites. *BMC Genomics* 2013 **14**:796.