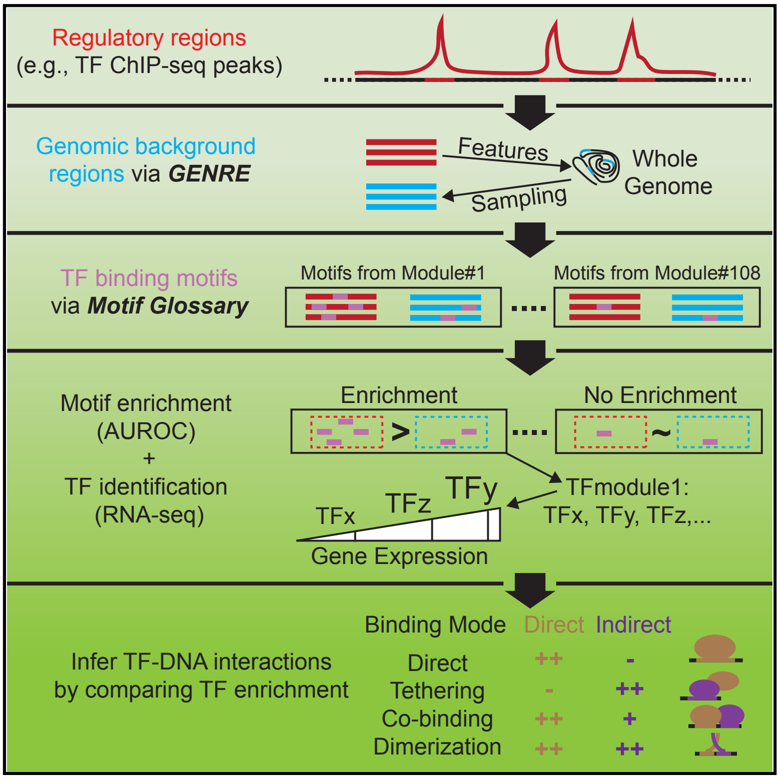
Cell Systems

Article

Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds

Graphical Abstract



Highlights

- Non-redundant TF-8mer glossary constructed from collection of universal PBM datasets
- GENRE software builds backgrounds matched to user-defined genomic and compositional features
- Glossary and GENRE outperform other tools in motif analysis of ENCODE ChIP-seq data
- Results suggest novel indirect TF-DNA tethering interactions

Authors

Luca Mariani, Kathryn Weinand, Anastasia Vedenko, Luis A. Barrera, Martha L. Bulyk

Correspondence

mlbulyk@genetics.med.harvard.edu

In Brief

Motif enrichment analysis of ChIP-seq data can elucidate the molecular mechanisms by which transcription factors regulate gene expression in a tissue-specific manner. In the current study, we created a glossary of the intrinsic DNA binding specificity of ∼40% of human TFs, developed a method to construct matched genomic background sequences, and showed that the combination of these two tools improves the identification of TF binding modes within regulatory regions.

Mariani et al., 2017, Cell Systems *5*, 187–201
September 27, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2017.06.015

Cell Systems
Focus on RECOMB
Article

Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds

Luca Mariani,[1] Kathryn Weinand,[1] Anastasia Vedenko,[1] Luis A. Barrera,[1,2,3] and Martha L. Bulyk[1,2,3,4,5,*]
[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[2]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA
[3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA
[4]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[5]Lead Contact
*Correspondence: mlbulyk@genetics.med.harvard.edu
http://dx.doi.org/10.1016/j.cels.2017.06.015

SUMMARY

Transcription factors (TFs) control cellular processes by binding specific DNA motifs to modulate gene expression. Motif enrichment analysis of regulatory regions can identify direct and indirect TF binding sites. Here, we created a glossary of 108 non-redundant TF-8mer "modules" of shared specificity for 671 metazoan TFs from publicly available and new universal protein binding microarray data. Analysis of 239 ENCODE TF chromatin immunoprecipitation sequencing datasets and associated RNA sequencing profiles suggest the 8mer modules are more precise than position weight matrices in identifying indirect binding motifs and their associated tethering TFs. We also developed GENRE (genomically equivalent negative regions), a tunable tool for construction of matched genomic background sequences for analysis of regulatory regions. GENRE outperformed four state-of-the-art approaches to background sequence construction. We used our TF-8mer glossary and GENRE in the analysis of the indirect binding motifs for the co-occurrence of tethering factors, suggesting novel TF-TF interactions. We anticipate that these tools will aid in elucidating tissue-specific gene-regulatory programs.

INTRODUCTION

Tissue-specific gene-expression patterns are encoded in metazoan genomes primarily via DNA sequence motifs that are recognized by sequence-specific transcription factors (TFs). Chromatin immunoprecipitation sequencing (ChIP-seq) data on *in vivo* TF genomic occupancy have been used to infer *cis* regulatory elements and TF DNA binding sites (Garber et al., 2012; Kundaje et al., 2015; Lara-Astiaso et al., 2014; Neph et al., 2012). However, accurate identification of the bound TFs can be complicated by cofactors modulating TF-DNA recognition *in vivo* (Biddie et al., 2011; Jolma et al., 2015; Shiina et al., 2015; Slattery et al., 2011). Depending on the TF, indirect DNA association through tethering by a sequence-specific TF with a different binding motif ("tethered binding") can explain a significant fraction of a TF's *in vivo* binding events (Gordan et al., 2009). To discriminate between direct and indirect bindings and infer recruiting factors (Gordan et al., 2009), ChIP-seq data are typically analyzed together with data on intrinsic TF DNA binding specificities, such as those obtained from protein binding microarrays (PBMs) (Berger et al., 2006) or HT-SELEX (Jolma et al., 2013). Universal PBMs, which assay the binding of a TF to all possible 8mers, have been used to screen ~30% of the mammalian TF repertoire (Badis et al., 2009; Weirauch et al., 2014), and PBM-derived *in vitro* binding specificities have been shown to correlate highly with *in vivo* binding data (Berger et al., 2008; Siggers et al., 2011a; Wei et al., 2010; Weirauch et al., 2013).

Although only a minority of the ~1,400 human TFs (Vaquerizas et al., 2009) have been assayed for DNA binding specificity, the binding preferences of many of the remaining TFs can be inferred from close homologs since members of TF families often exhibit highly similar DNA binding specificities (Badis et al., 2009) and share motif recognition (Berger et al., 2008; Nakagawa et al., 2013; Wei et al., 2010). Analysis of the motif repertoire in a large TF ChIP-seq collection (119 TFs in 72 cell lines) yielded just 79 unique motifs, 67 of which were already reported (Wang et al., 2012a). Although it remains unclear how many unique motifs will be sufficient to describe the human TF motif repertoire, a well-curated, non-redundant core set of motifs would expedite the identification of TF binding events in genomic sequences. While exhaustive collections of motifs typically include thousands of position weight matrices (PWMs) (Kheradpour and Kellis, 2014; Weirauch et al., 2014), their quantitative clustering suggested that core sets of approximately 100 motifs were sufficient to predict TF targets across the genome (Kheradpour et al., 2007) or to assess enrichment of motif combinations in developmental enhancers (Gisselbrecht et al., 2013). However, since those motif sets represented only available *Drosophila* TF binding, and were derived from both *in vitro* and *in vivo* data, they are not sufficiently comprehensive and accurate to identify the genomic occupancies of human TFs.

While a PWM can provide a good approximation of a TF's DNA binding specificity, PWM models can vary significantly between motif-derivation algorithms in terms of their ability to capture *in vivo* TF binding sites (Weirauch et al., 2013). Moreover, TF subfamilies can recognize additional, distinct sequences (Badis et al., 2009; Berger et al., 2008; Nakagawa et al., 2013; Siggers et al., 2011a) that might be not captured by a single PWM representing the shared binding preferences of the TF family (Gordân et al., 2013; Nakagawa et al., 2013). Furthermore, some TFs can recognize two or more distinct sets of sequences that are better represented as multiple motifs (Badis et al., 2009; Nakagawa et al., 2013). Going beyond PWMs, *k*-mer-based models describe a TF's preferences for binding sequences of length *k* and can capture higher-order complexities in DNA sequence recognition, such as variable spacers or position interdependence, than do typical PWMs (Arvey et al., 2012).

The DREAM5 TF-DNA Motif Recognition Challenge (Weirauch et al., 2013) compared a large selection of PBM-derived models for TF specificity and concluded that for most of the examined TFs, PWMs performed similarly to more complex models. Moreover, although *k*-mer models outperformed PWMs in predicting *in vitro* TF DNA binding, PWMs performed better in distinguishing *in vivo* binding (Weirauch et al., 2013). However, the DREAM5 competition evaluated performance on relatively few *in vivo* datasets (five mouse ChIP-seq and four yeast ChIP-exo experiments) and focused exclusively on the detection of direct DNA binding with no assessment of indirect binding events.

Motif enrichment in ChIP-seq "bound" regions can be quantified by the area under the receiver operating characteristic (AUROC) curve, the established metric to evaluate how well a motif distinguishes the bound (foreground) from unbound (background) sequences (Gordan et al., 2009; Weirauch et al., 2013). Use of an appropriate background set, therefore, is crucial (Worsley Hunt et al., 2014); because of the sequence biases in the genome (Badis et al., 2009; Plotkin and Kudla, 2011), accurate identification of motif enrichment depends on the selection of a background with contextual and compositional features similar to those of the foreground (Hung and Weng, 2017). For example, GC content varies across the genome (Nekrutenko and Li, 2000), and, if not controlled for, can significantly bias the identification of GC-rich sequence motifs in foreground sequences.

A popular method to minimize this GC content bias is sequence reshuffling with preservation of dinucleotide frequencies ("dinucleotide shuffle") (Bailey, 2011; Barrera et al., 2016; Jiang et al., 2008; Kundaje et al., 2015; McLeay and Bailey, 2010; Weirauch et al., 2013, 2014). However, the resulting randomized sequences are not native to the genome, and thus potentially neglect other local and/or higher-order biases inherent in the foreground, such as shape preferences (Abe et al., 2015) or the periodic alternation of GC and AT nucleotides within nucleosomal DNA (Struhl and Segal, 2013). Genomic regions flanking the foreground sequences ("flanking regions") have been used as background to overcome this concern (Bailey and Machanick, 2012; Orenstein and Shamir, 2014; Setty and Leslie, 2015; Siggers et al., 2011a; Wang et al., 2012a), although they do not exclude GC bias. Alternatively, background sequences obtained by supervised, semi-random selection from the genome facilitate the control for sequence biases. In an analysis of 43 ChIP-seq datasets, GC-controlled random selections of genomic sequences implemented by HOMER (Heinz et al., 2010) and BiasAway (Worsley Hunt et al., 2014) software ("GC-HOMER" and "GC-BiasAway") outperformed sequences obtained by dinucleotide shuffle in detection of enriched motifs (Worsley Hunt et al., 2014); however, that comparison did not consider flanking regions as an alternate background (Worsley Hunt et al., 2014) and did not evaluate motif enrichment by AUROC (Weirauch et al., 2013).

In this study, we first assembled published universal PBM data (Badis et al., 2009; Barrera et al., 2016; Berger et al., 2008; Nakagawa et al., 2013; Peterson et al., 2012; Siggers et al., 2011a; Wei et al., 2010; Weirauch et al., 2013, 2014; Zhu et al., 2012) for 650 metazoan TFs and expanded it with new data for 21 additional human TFs (Table S1). Clustering the PBM 8mer data yielded a "glossary" of 108 TF-8mer specificity "modules," where each module associates one or more TFs with their shared set of recognized 8mers. We evaluated different types of background sequences, including those generated by our newly developed GENRE tool, which selects background genomic sequence sets that are controlled for user-defined sequence properties. GENRE outperformed other state-of-the-art methods in identifying motifs enriched in a collection of 239 ENCODE ChIP-seq experiments (ENCODE Project Consortium, 2012). When we interfaced the glossary 8mer modules with GENRE, we found that 8mers are more specific but less sensitive than PWMs in suggesting motifs for tethering factors. Integration of gene-expression data from ENCODE RNA sequencing (RNA-seq) experiments (ENCODE Project Consortium, 2012) into our analysis enabled the precise identification of which TF, among all the TFs associated with an enriched motif, is likely to be responsible for tethering the ChIPed factor. This analysis also suggested that 8mer modules are preferable to reveal additional motifs recruiting TFs indirectly. Finally, we identified several TF tethering interactions, both known and novel, which we characterized by comparing the peak overlap of co-binding factors and the co-occurrence of the indirect motif.

RESULTS

Constructing a Glossary of 8 bp Sequences Specific for TF (Sub)Families

We collected 773 universal PBM 8mer datasets for 650 metazoan TFs from the UniPROBE (Hume et al., 2015) and CIS-BP (Weirauch et al., 2014) databases (Table S1). In addition, to expand the available data for TF DNA binding domain (DBD) classes with poorly (e.g., ARID) or incompletely (e.g., C2H2 zinc-finger and nuclear hormone receptor) characterized DNA binding specificities, we performed universal PBM experiments for 63 TFs, 21 of which previously lacked PBM data (Table S1). After filtering the PBM data according to quality control criteria (STAR Methods), our final dataset includes 671 TFs from 10 different species assayed in 715 PBM experiments, often merging two independent replicates with different PBM array designs (Berger et al., 2006). This dataset encompasses 10,428 bound 8mers with more than 20 DBD classes represented by at least 3 TFs (Figure 1A).

To group TFs with their recognized 8mers into TF-8mer "modules," we performed hierarchical clustering on all the TFs and "bound" 8mers according to their E-scores (Badis et al., 2009) (Figure S1A). Because TF families and subfamilies are highly variable in their motif information content, we reasoned that

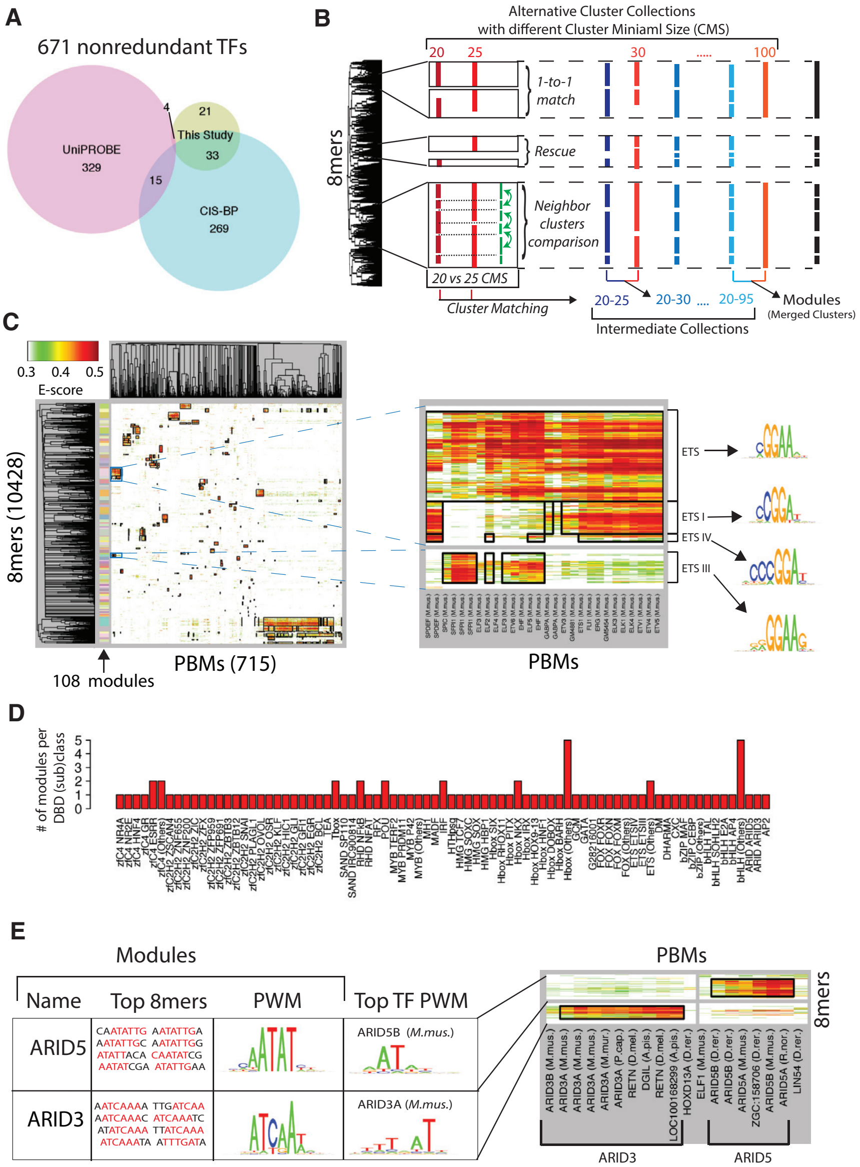


Figure 1. Glossary of 108 TF-8mer Modules Derived from PBM Data

(A) Venn diagram depicting sources of PBM data analyzed in this study (Table S1).

(B) Schema of the TF-8mer clustering strategy. The 8mer dendrogram (left) is initially cut into several alternative cluster collections (red to orange) defined solely by the minimal number of 8mers contained in each cluster, called the cluster minimal size (CMS), which we varied between 20 and 100 8mers. We created an optimal collection of "merged clusters" by progressively matching collection pairs (blue to cyan); i.e., the comparison of the collections with CMS 20 and 25 produces the intermediate collection 20–25, which is then compared and matched to CMS 30, and then iteratively up to the CMS 100, which gives the final merged clusters (black, Modules). Alternative clusters for the same 8mers are matched according to the depicted rules (STAR Methods): *1-to-1 match*, *Cluster rescue*, and *Neighbor clusters comparison*.

(C) Left: 2D hierarchical re-clustering of 671 TFs (715 PBM datasets) and 10,428 ungapped 8mers bound ($E \geq 0.3$) in at least one of the PBM datasets. Black frames within the heatmap outline the 108 modules of correlated TF-8mer binding profiles; colored bars to the left of the heatmap denote glossary modules. Right: zoom-ins on the modules specific to ETS TFs.

(D) Distribution of TF-8mer modules across TF DNA binding domain (DBD) (sub)families.

(E) ARID module comparison. Right: Zoom-in on the two ARID modules from the heatmap shown in (B). Left: for each ARID module: 8mers with the highest E-score (Top 8mers with core consensus sequence in red), PWMs of module 8mers (PWM), and the CIS-BP PWMs of the TF with highest mean E-score specificity within each module (Top TF PWM) (Table S2).

imposing a fixed PBM 8mer E-score threshold to cut the dendrogram into branches might hinder the recognition of local 8mer patterns corresponding to different TFs with similar binding preferences and likely within the same TF family. Therefore, we first created several independent collections of 8mer clusters by repeatedly cutting the 8mer dendrogram in a size-dynamic manner (Langfelder et al., 2008) (Figure 1B). To associate specific TFs to each 8mer cluster, we applied two-means clustering on the E-score profile to discriminate TFs exhibiting high versus low specificity for the 8mers (Figure S1B). We then created an optimal collection of "merged clusters" by progressively comparing and matching collection pairs (Figures 1B and S1C STAR Methods). After one round of re-clustering, this analysis of TF specificity yielded a "glossary" of 108 TF-8mer modules used in the rest of this study (Figure 1C; Table S2), which we also used to construct PWMs for each module ("8mer PWM") and to identify the most representative PWM ("TF PWMs").

TF-8mer Glossary Discriminates TFs and TF Subfamilies with Distinct Specificities

To investigate the degree to which the TF-8mer modules capture families of closely related TFs, we evaluated each module's homogeneity in terms of the DBD structural classes of its member TFs. The vast majority of the modules (90/108) predominantly comprised TFs from single DBD classes (Figure 1D) and recapitulate previously reported TF specificities (Table S2).

Although prior studies have analyzed PBM data to explore the heterogeneity of DNA sequence recognition by TFs from various DBD classes, those studies were either based on much smaller datasets (Badis et al., 2009; Gordan et al., 2011) or were focused on specific TF families and did not consider binding specificities shared by TFs from different DBD classes (Berger et al., 2008; Grove et al., 2009; Nakagawa et al., 2013; Wei et al., 2010; Weirauch et al., 2014). Here, by comparing the 8mer binding preferences of a large collection of TFs across a wide range of DBD classes, our glossary reveals distinct sequence preferences of TF subfamilies, such as ETS (Figure 1C), IRF, and FOX, and of individual TFs, such as TEA (Figure S1D). For

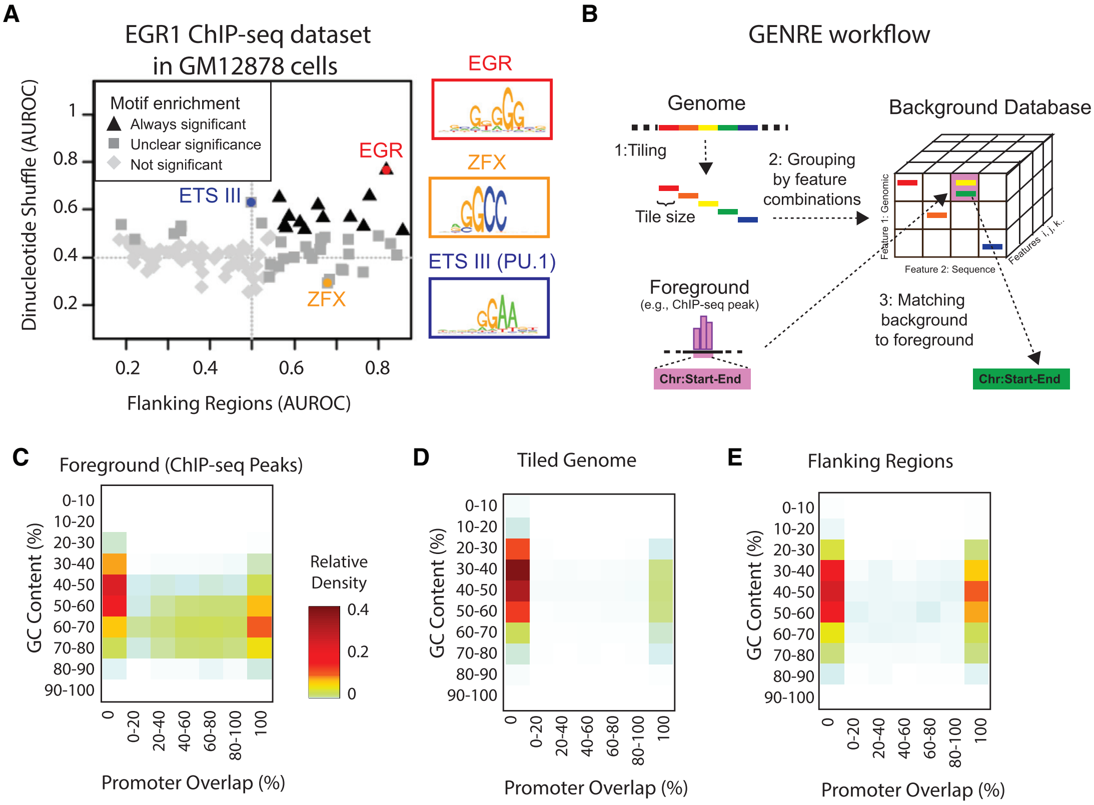


Figure 2. GENRE: A Tool for Constructing Matched Genomic Background Sequence Sets

(A) Motivating example showing lack of discriminatory power of background sequence sets based on dinucleotide shuffle or flanking regions. Motif enrichment for EGR1 ChIP-seq peaks in GM12878 cells (Table S3) calculated using 108 individual TF PWMs (Table S2; STAR Methods) representing each of the glossary modules. Scatterplot of motif enrichment AUROC values obtained using dinucleotide shuffle (y axis) versus flanking regions (x axis) as the background. Black triangles and dark gray squares indicate motifs found enriched using both or just one background type, respectively (AUROC > 0.5 and $p < 0.05$ Fisher's exact test with false discovery rate correction). Colored dots highlight the PWMs represented in the sequence logos.

(B) GENRE schema. The whole genome is tiled in non-overlapping regions of fixed size, as determined by the foreground set (Tiling). The regions are grouped according to similarities in sequence and genomic features chosen by the user (Grouping by feature combinations). For each foreground sequence, a background sequence is randomly sampled from the bin with the same feature grouping (Matching background to foreground). In this study, the sequence features are GC content and CpG frequency, and the genomic features are promoter overlap, i.e., percentage of the sequence located within 2 kb upstream of a transcription start site, and repeat overlap.

(C–E) Distributions of GC content and promoter overlap differ across foreground, random genome, and flanking regions. 2D density plot for promoter overlap and GC content (C) in all the ChIP-seq foreground regions analyzed in this study (Table S3), (D) in a representative subset of the default GENRE tiles for the human genome (hg19 version, STAR Methods), and (E) in the 200 bp flanking regions located 1 kb upstream of the foreground ChIP-seq peaks.

example, the glossary divides ARID TFs, which were previously not clearly distinguishable according to their TF PWMs, into two modules with distinct specificities, with the ARID5 module TFs recognizing an ATATTG motif and the ARID3 module TFs recognizing an ATCAA motif (Figure 1E).

GENRE, a Method for Constructing Background Sequences Matched for Genomic Regulatory Regions

We utilized our TF-8mer glossary to evaluate how different background generation methods in motif analysis influence the motif enrichment statistics. We first compared four different commonly used methods for background construction, described above: dinucleotide shuffle, flanking regions, GC-HOMER, and GC-BiasAway. For our study, we compiled 239 ChIP-seq datasets from 80 sequence-specific TFs examined in 33 different cell types by the ENCODE Consortium (Table S3) (ENCODE Project Consortium, 2012). For each dataset, we took as foreground sets the top 500 ChIP-seq peaks, trimmed at ±100 bp around the peak summit. As in DREAM5 (Weirauch et al., 2013), we scored motif enrichment in each ChIP-seq foreground set by AUROC.

Overall, these different types of background sequences yielded similar, strong enrichment of the motifs corresponding to direct DNA binding by the ChIPed TFs (Table S4). For example, the EGR motif was highly enriched within EGR1 ChIP-seq peaks in GM12878 cells using either dinucleotide shuffle background or flanking regions background (Figure 2A). In contrast, the milder enrichment of additional motifs was inconsistent using different background methods (dark gray squares, Figure 2A). Accurate identification of such additional motifs, like those for the hematopoietic factors PU.1 (also known as SFPI1 and encoded by SPI1 gene) and ZFX (Zon, 2008), may suggest tethering factors mediating indirect DNA binding of the ChIPed TF.

To enhance the ability to identify co-regulatory motifs in foreground genomic regions, we developed GENRE (genomically equivalent negative regions), a genome randomization method for construction of genomically matched background sequences. To overcome limitations of previous methods for generating background sequences, we designed GENRE to allow a user to choose genomic features that represent potential biases for the sequence, e.g., GC content, CpG dinucleotide frequency (Bhasin and Ting, 2016; Spruijt and Vermeulen, 2014), and the overlaps with repeat sequences (Boeva, 2016) ("repeat overlap") and with the promoters ("promoter overlap"), defined as the 2 kb regions upstream of transcription start sites (STAR Methods).

In brief, GENRE first builds a database of putative background sequences by tiling a genome in non-overlapping regions and then grouping the sequence tiles according to the user-specified sequence features (Figure 2B; STAR Methods). Next, GENRE matches each foreground sequence to the database bin with the same grouping of features, and then randomly samples a background sequence from that bin. To investigate contextual and compositional biases that might impact the construction of background sequences, we again considered the top 500 ChIP-seq peaks from each of 239 ENCODE ChIP-seq datasets as foreground sets. We then used GENRE to construct background sets, controlling for all four of the potential sequence bias features discussed above.

We noticed that the feature distributions in the ENCODE ChIP-seq foreground sets were either all-or-none, such as for promoter (Figures 2C, 2D, S2A, and S2B) and repeat (Figures S2C and S2D) overlaps, or pronouncedly skewed, such as for GC content (Figures 2C and 2D) or CpG frequency (Figure S2). To optimize the binning, GENRE classifies features binarily for those with all-or-none distributions and utilizes quartiles for features with skewed distributions. The interdependency between features can also affect background construction, as noticeable in the high GC content exhibited by most promoters (Boeva, 2016) (Figure 2C), whereas promoters and GC-rich regions are rare across the whole genome. Random genomic promoters (rightmost column, Figure 2D) exhibited a GC content distribution in between that of the foreground peaks (Figure 2C) and the whole genome (Figure 2D). By selecting only from the random genomic promoters, GENRE can better match the foreground sequences with high GC content (STAR Methods). Notably, background sets constructed by flanking regions (Figure 2E) exhibited a GC content of overlapping promoters markedly reduced compared with foreground sequences (Figure 2C), indicating that within flanking regions not all of the foreground features' attributes are accurately represented.

Glossary-Based Evaluation of Construction Methods for Background Regulatory Sequences

We compared GENRE with dinucleotide shuffle, flanking regions, GC-HOMER, and GC-BiasAway for their impact on motif enrichment statistics, considering the 239 ENCODE ChIP-seq datasets and 108 glossary modules as a benchmark (Table S3) (ENCODE Project Consortium, 2012). To evaluate the performance of each background construction method, we considered two criteria. First, we noted that since the binding motif anchoring a ChIPed TF tends to occur near the center of a ChIP-seq peak (Bailey and Machanick, 2012; Wang et al., 2012a), its enrichment and median distance to the peak summit tend to be correlated. For example, TEAD4 ChIP-seq peaks are enriched for the TEA motif in hepatocellular carcinoma (HepG2) cell lines (Figure 3A) and human embryonic stem cells (H1-hESCs, data not shown), consistent with direct binding of TEAD4 to the regions bound *in vivo*. Accordingly, the TEA motif tends to occur near the peaks' centers. In contrast, the GATA, rather than the TEA, motif is both enriched and centered within TEAD4 ChIP-seq peaks in the human leukemia K562 cell line (Figure 3A), suggesting that TEAD4 associates indirectly with DNA via a directly bound GATA factor in K562 cells (Figure 3A). Therefore, we calculated the Spearman rank correlation coefficient rho between the motif enrichment AUROC value and the motif-summit distance ("AUROC distance correlation") as a summary statistic across all 239 ChIP-seq datasets and 108 glossary modules (Figure 3B).

Our second criterion was the median absolute deviation (MAD) of the AUROC distribution, which we used to estimate the specificity of a background in guarding against false-positive motif enrichment. As noted in previous analysis of motif over-representation (Worsley Hunt et al., 2014), the genomic occupancy of a factor typically involves just a few different TF binding site motifs. Most of the AUROC values obtained by the glossary across the ChIP-seq datasets are thus distributed around a central bulk of "no enrichment" (AUROC = 0.5) (Figures 3B and S3). Therefore, we reasoned that the ability of a background method to filter out false positives is associated with low AUROC dispersion, which we measured by the MAD. As opposed to the standard deviation, the MAD can more robustly exclude the effect of outliers, which we expected to occur from enrichment of a TF's direct motif (Table S4).

Considering these two criteria, the best method is the one with an AUROC distance correlation closest to −1 and AUROC MAD closest to 0 (gray lines in Figure 3C). Thus, the methods based on genome randomization (GENRE, GC-HOMER, and GC-BiasAway) outperformed dinucleotide shuffle and flanking regions, with GENRE performing the best overall (Figures 3C and S3). Each of the GENRE features was validated using an all-but-one analysis as well (Figure S3). To assess whether the performance of a background depends on the sequence specificity of the ChIPed TFs, we repeated the comparison on subsets of the ChIP-seq datasets, sorted by TF family (Figures 3D and 3E). Indeed, for certain families, such as TEA (Figure 3D), the results were highly concordant across different types of background while for others, such as KLF, which includes EGR and SP TFs (Figure 3E), the choice of the background had a strong influence on the motif enrichment. Overall, GENRE typically outperforms the other background construction methods (Figures 3C–3E and S4).

Glossary TF-8mer Modules Outperform PWMs in Identifying Enriched Motifs within ChIP-Seq "Bound" Regions

Since the DREAM5 evaluation on ChIP data was performed for just nine TFs and considered the enrichment of just the direct binding motif, in this study we undertook an extensive comparison of the ability of our glossary's TF-8mer modules versus TF PWMs to identify direct versus indirect binding in 239 human ENCODE ChIP-seq datasets in comparison with background sets constructed by GENRE. To evaluate the performance of

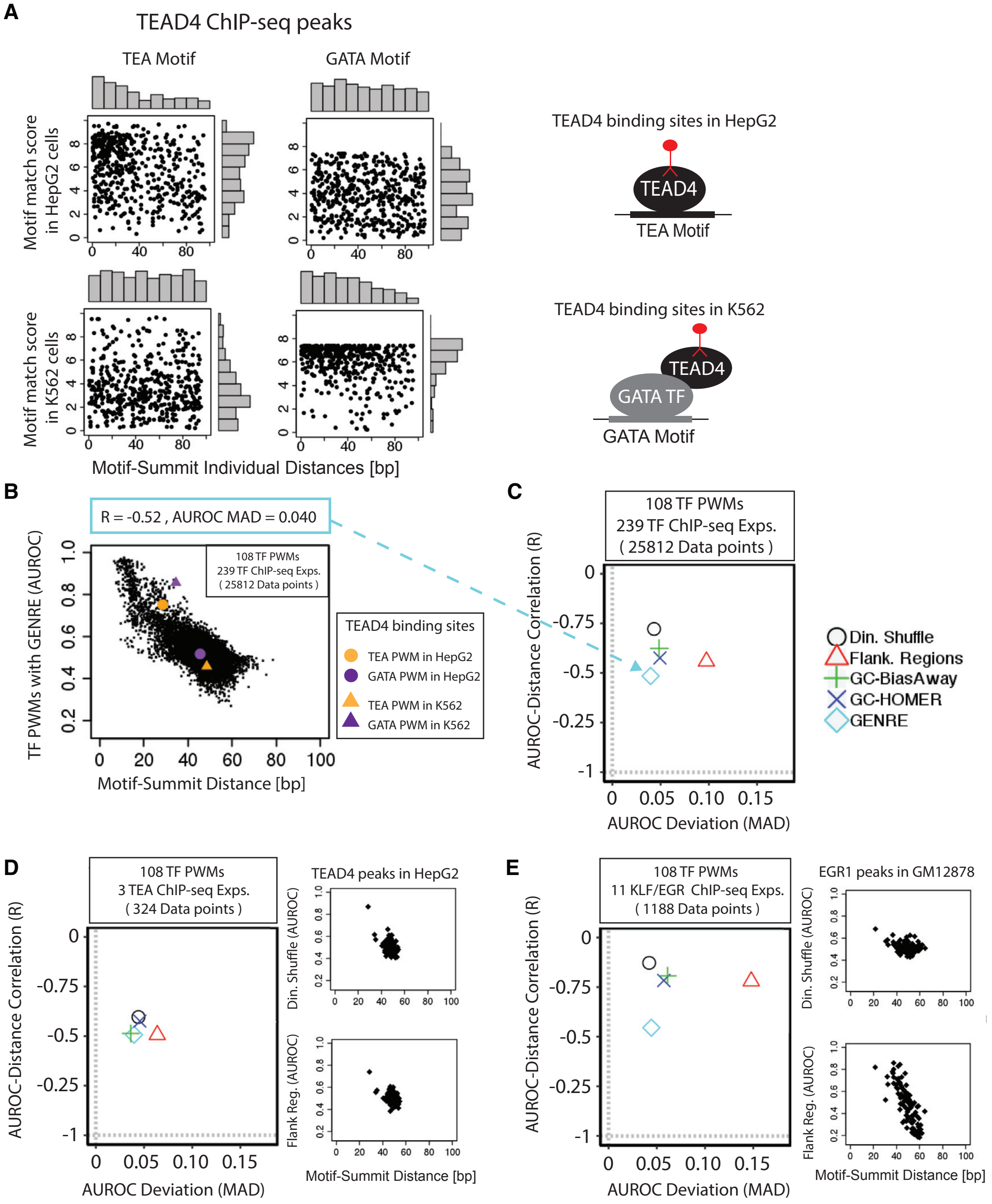


Figure 3. Evaluation of Different Background Models for Analysis of Motif Enrichment in ChIP-Seq Peaks

(A) Scatterplots depicting correlation between the position of the best scoring TF PWM match (Table S2) relative to each peak's summit (Motif-Summit Individual Distances) and the score of that best TF PWM match (Motif match score). The histograms adjacent to each plot represent the marginal distributions. PWMs associated with TEA (left panels) and GATA (right panels) modules were used to scan the top 500 TEAD4 ChIP-seq peaks in HepG2 (upper panels) or in K562 (lower panels) cells from ENCODE. Cartoons depict molecular interpretations of the motif enrichment results: TF with direct DNA binding (black ovals), TF with indirect binding (gray oval), and ChIPed antibody (red sticks).

(B) Scatterplot of the AUROC values for PWM match score calculated using GENRE background (TF PWMs with GENRE) versus the median distance of the best PWM match from the peak's summit (Motif-Summit Distance). Each point represents one of the 108 TF PWMs applied to one of the 239 ENCODE TF ChIP-seq datasets (as in Figure 2A). Colored points indicate the 2 PWMs and 2 ChIP-seq datasets analyzed in (A). Numbers in blue box above the plot report the two metrics used for background evaluation: R, the Spearman rank correlation coefficient rho between all the values of Motif-Summit Distance and PWM AUROC in the scatterplot ($R = -0.53$, $p << 0.001$); AUROC MAD, the median absolute deviation across all the AUROC values in the scatterplot.

(legend continued on next page)

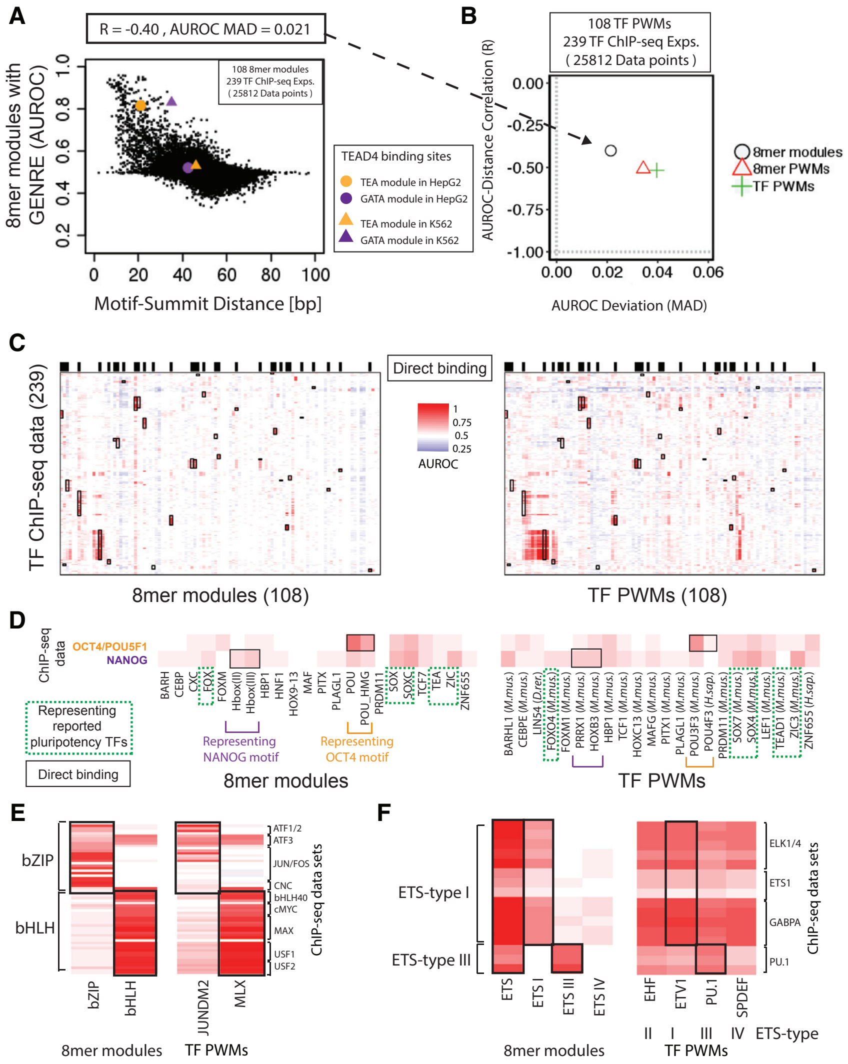


Figure 4. PBM-Derived TF-8mer Glossary Precisely Identifies *In Vivo* TF Binding Events

(A) As in Figure 3B, but with 8mer modules instead of PWMs used to calculate AUROC values and motif-summit distances.

(B) As in Figure 3C, but with calculation of AUROC values and identification of motif occurrences within ChIP-seq peaks using either 8mer modules, 8mer-derived PWMs, or TF PWMs representing each of the modules (Table S2). All the reported values of Spearman rho (R) were highly significant (p << 0.001).

(C) Motif enrichment in ChIP-seq peaks for 239 sequence-specific TFs (ENCODE Project Consortium, 2012), calculated using all 108 glossary 8mer modules versus individual TF PWMs representing each of the glossary modules (Table S2; STAR Methods). Black boxes highlight AUROC values corresponding to the expected direct binding of the ChIPed TF to its cognate motif.

(D) Significantly enriched 8mer modules or corresponding TF PWMs in H1-hESC ChIP-seq peaks of NANOG and OCT4 (encoded by the POU5F1 gene). Black boxes highlight AUROC values corresponding to the expected direct binding of the ChIPed TFs to their cognate motifs. Green dotted boxes highlight other reported pluripotency TFs. AUROC color bar as in (C).

(E) Enrichment of 8mer modules versus TF PWMs in ChIP-seq peaks of the corresponding bZIP and bHLH TFs. AUROC color bar as in (C).

(F) Same as in (E) for ETS TFs.

each motif model, we utilized the AUROC distance correlation and AUROC MAD criteria described above. For the TF-8mer modules, we computed the AUROC by comparing the presence of these 8mers in the foreground versus background sequences (Barrera et al., 2016) (STAR Methods).

The 8mer modules exhibited AUROC values that were more narrowly distributed (MAD = 0.021) and less correlated with the motif-summit distance (Spearman rank correlation coefficient rho [R] = −0.4) than did TF PWMs (MAD = 0.04 and R = −0.52, respectively, Figures 3B and 4A). Thus, according to these criteria, 8mer modules are more specific, but less sensitive, than PWMs in detecting motif enrichment, suggesting that the module-based enrichment analysis yields fewer potential false positives. To rule out the possibility that these differences might have arisen from artifacts due to glossary specificity, we evaluated the performance of PWMs constructed from the 8mers of each module. These "8mer-derived PWMs" recapitulated the performance of the individual TF PWMs (Figure 4B), supporting our conclusion that the difference in motif enrichment predicted by PWMs and 8mers was a property of the type of motif model used in the analysis. Across the entire matrix of 108 glossary motifs and 239 ChIP-seq datasets, the 8mer modules resulted in an overall sparser set of enriched motifs, providing a "sharper," potentially more precise view of motif enrichment (Figure 4C). This ability, which was consistent in ChIP-seq data across TF families (data not shown), can furnish a reasonable trade-off since an experimentalist would prefer greater confidence in a smaller number of predicted interactions to test experimentally.

(C) Choice of background type has an impact on the performance of motif enrichment analysis. For each background type, AUROC MAD, and AUROC distance correlation R were evaluated aggregately across the 239 ChIP-seq datasets as in (B). All the reported values of Spearman rho (R) were highly significant (p << 0.001). Gray dashed lines indicate the limits of specific (x = 0, minimal deviation) and sensitive (y = −1, maximal inverse correlation) outcomes.

(D) As in (C), but for ChIP-seq data for TEA TFs. Right insets depict AUROC values versus Motif-Summit Distances obtained using the 108 TF PWMs applied to TEAD4 ChIP-seq peaks from HepG2 cells and either flanking regions (lower plot) or dinucleotide shuffle (upper plot) as the background model. All the reported values of Spearman rho (R) were highly significant (p << 0.001).

(E) As in (D), but for ChIP-seq data for KLF/EGR TFs in GM12878 cells.

DeepBind, a machine-learning approach to describe TF sequence specificity by a set of motif detectors that function similarly to PWMs, had been reported recently to outperform the models previously tested in the DREAM5 Challenge (Alipanahi et al., 2015). Therefore, for the 13 glossary modules also present as DeepBind models, we compared the motif enrichment across the 239 ChIP-seq datasets and found that DeepBind models behaved similarly to single PWMs (Figure S5).

To corroborate our findings, we looked at relevant biological cases provided by our datasets. Within the ChIP-seq peaks of the two pluripotency TFs OCT4/POU5F1 and NANOG in human embryonic stem cells (Figure 4D), both models resulted in enrichment of the POU and homeodomain motifs, consistent with direct binding by OCT4 and NANOG, and also of the motifs for other pluripotency TFs (e.g., SOX, ZIC, TEA, and FOX) (Luo et al., 2015; Yagi et al., 2007; Young, 2011). However, in contrast to the TF-8mer modules, 46 TF PWMs of TF families (e.g., IRF, E2A, and CEBP) that, to our knowledge, have not been implicated in pluripotency and that are not expressed at this stage, were enriched, suggesting that the TF-8mer modules result in more specific, biologically relevant motif enrichment. Similarly, the TF-8mer modules of bZIP and bHLH TFs overall outperformed PWMs in terms of higher AUROC values, consistent with direct DNA binding by the ChIPed TF, and the motif enrichment being more specific to the ChIPed TF's family (Figure 4E). The motif enrichment results for ETS TFs also highlighted the higher specificity of TF-8mer modules in distinguishing direct DNA binding of ETS subfamilies, compared with PWMs derived for individual members of the ETS subfamilies (Figure 4F).

Gene-Expression Profiles Corroborate *k*-mer-Based Predictions of Lineage-Specifying TFs

TF expression levels are highly regulated during cellular differentiation as they play a key role in controlling the underlying network of regulatory interactions. Cell types can be distinguished by the upregulation of particular TF families, such as IRF TFs in GM12878 cells (Figure 5A), or of a specific TF from a TF family, such as POU5F1 in H1-hESCs (Figure S6A). We considered the expression patterns of 7 cell lines highly represented in the 239 ENCODE ChIP-seq datasets analyzed in this study (Table S5). We noticed that, within a given cell line, the motifs enriched within ChIP-seq peaks often suggested indirect binding and, moreover, matched the specificities of highly expressed TFs (Figures 5B and S6B). For example, in GM12878 cells, peaks for non-IRF TFs frequently exhibited an enriched IRF motif (right inset, Figure 5B) centered within the ChIP-seq peaks (Figure 5C). This suggests that an IRF TF, such as IRF1, IRF4, or IRF8, which are expressed highly in GM12878 cells (Figure 5A), mediates indirect DNA binding for the ChIPed non-IRF TFs (e.g., PU.1 and BATF). These results both recapitulate well-characterized interactions, such as between IRF4 and PU.1 or BATF (Glasmacher et al., 2012; Murphy et al., 2013), and suggest additional binding partners for IRF TFs (e.g., CEBPB, ATF2, RUNX3, and MEF2A). We noticed similar behavior for other TF families in different cell lines, such as FOX in HepG2 cells, GATA in K562 cells, and TEA and ZIC in H1-hESCs (Figure S6B).

While motifs of highly expressed TFs were found to be similarly enriched regardless of the use of PWMs or 8mer modules, or of different types of background (with the notable exception of flanking regions), the enrichment of motifs of TFs that were not highly expressed was much more subject to the choice of motif model and background (Figures 5B and S6B). To evaluate the impact of different parameter settings in the motif enrichment analysis, we inspected the gene-expression levels of the TFs whose motifs are enriched. Since motif enrichment results suggestive of indirect binding were much more variable on the choice of motif model and background type than direct binding (Table S4), we focused our analysis on indirect binding. Here, we considered all members of 12 TF families with well-characterized, distinct motifs, and which are known to drive particular cellular differentiation processes (STAR Methods).

To quantify the upregulation of the putative tethering TFs in each cell line, we calculated the *Z* score of their enrichment among genes highly expressed in the respective cell line (STAR Methods). To identify a reliable expression range for factors that are upregulated in a highly cell-type-specific manner and are known to mediate regulatory interactions in those cells, we first assembled a reference set of cell-type-specific regulators (e.g., Sox2 and Oct4 in H1-hESCs) (STAR Methods) and noticed that their expression ranked within the top 3,000 most highly expressed genes (Figure S6C). Among these highly expressed genes, we found that the tethering TFs predicted using the glossary 8mer modules with GENRE background were significantly enriched (*Z* score = 4) (Figure 5D).

Having established this threshold and the *Z* score metric for the most highly expressed genes, we then compared the enrichment of tethering TFs predicted using the different motif models and background types (Figure 5E). While flanking regions strikingly resulted in complete lack of enrichment, most of the background types yielded similar, modest enrichment (Figure 5E); therefore, in subsequent analyses, we used only GENRE background sets and found that 8mer modules resulted in strong, highly significant enrichment, in contrast to TF PWMs and 8mer-derived PWMs (Figure 5E). These gene-expression *Z* score values correlated with the AUROC MAD values (Figure 5F), supporting the conclusions that (1) the use of 8mer modules rather than PWMs results in more reliable predictions of tethering factors responsible for indirect binding, and (2) regions flanking peaks do not represent a reliable background.

Novel TF Tethering Interactions

We further investigated the significantly enriched TF-8mer modules obtained using GENRE background (Figure 6A) to further characterize the potential interactions between a ChIPed TF (Tethered TF A) and motifs bound indirectly (Indirect Motif from Module B) through other TFs (Tethering TF B) (Figure 6B). By fixing thresholds for AUROC (>0.6) and motif-summit distance (<40 bp) that reliably distinguished the enriched, peak-centered modules from the background noise (Figure 6A), we identified 48 indirect *in vivo* interactions sometimes occurring in multiple cell types (Table S6), of which 16 were previously validated within the literature, 19 had been proposed in prior studies but had not been confirmed experimentally, and 13 that to our knowledge were novel to this study.

We noticed that the enrichment of the cognate motif of the Tethered TF A (i.e., Module A 8mers) and that of the indirect motif recognized by the Tethering TF B (i.e., Module B 8mers) were not highly correlated (Figure 6C). However, we found that a markedly

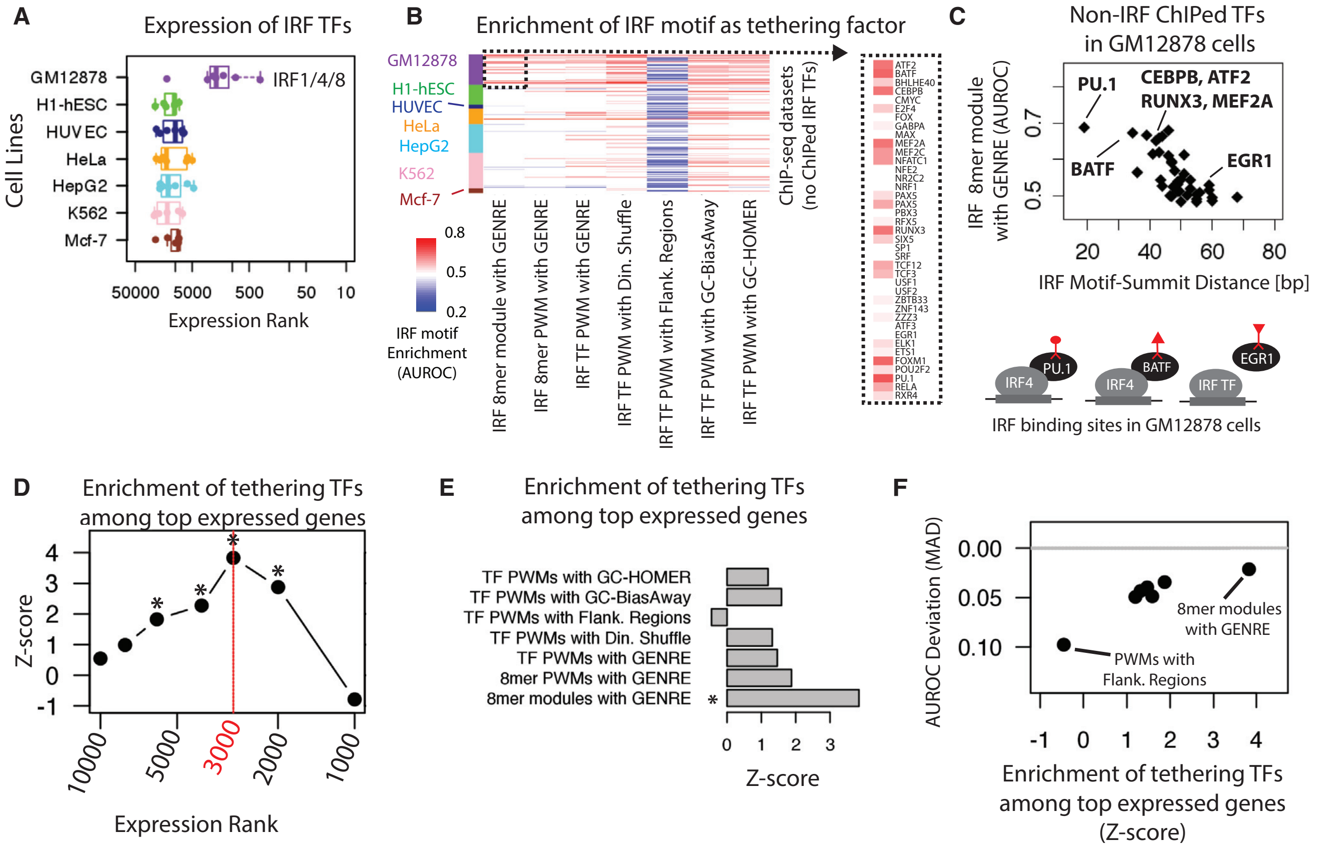


Figure 5. Gene-Expression Data Corroborate the Glossary-Derived Motif Enrichment and Identify Lineage-Specific Regulators

(A) Expression ranks of IRF TFs in the indicated cell types, as determined by independently ranking the whole genome by RNA-seq expression levels (fragments per kilobase of transcript per million fragments mapped) (Trapnell et al., 2010) (Table S5).

(B) Enrichment of IRF motif in ChIP-seq peaks of non-IRF TFs in various cell lines (rows) calculated using the indicated background sets (as in Figure 3C) and specificity models (as in Figure 4B). Enrichment suggests DNA tethering of the ChIPed TF through an IRF TF. Bars along the y axis (left) indicate cell lines used in ChIP-seq experiments, color-coded as in (A). Inset: zoom-in on GM12878 peaks for the indicated ChIPed TFs.

(C) Scatterplot of IRF 8mer module AUROC values calculated using GENRE background versus the median distance between the best IRF motif match from the summit of a non-IRF TF's ChIP-seq peak (IRF Motif-Summit Distance) (as in Figure 3B). Only non-IRF TF peaks were used, as in (B). Known immune factors are highlighted.

(D) Enrichment of putative tethering TFs among the top expressed genes. Motif enrichment was calculated using the 8mer modules with GENRE background (as in Figure 4A). x axis: threshold gene-expression rank used to call highly expressed genes; y axis: *Z* scores calculated using 200 randomly reshuffled genomes. *p < 0.01, calculated as the fraction of the randomly reshuffled genomes that resulted in more putative tethering TFs being included among the top expressed genes than that obtained with the real genome.

(E) Enrichment of putative tethering TFs among the top 3,000 most highly expressed genes (Table S5), calculated using the different motif models and background types. *p < 0.01, calculated as in (D).

(F) Scatterplot of AUROC deviation (MAD) (see also Figures 3C and 4B) calculated using the different motif models and background types versus *Z* scores presented in (E).

higher enrichment of indirect motifs discriminated cases where the tethered TF A is recruited by a tethering TF B, such as TAL1 by GATA1/2 in K562 cells (Wadman et al., 1997). In contrast, for the bZIP TF BACH1, which binds DNA as a heterodimer with MAFK, we instead found that the direct bZIP motif (Module A) was enriched to a similar degree as the indirect MAF motif in H1-hESC (Module B), in agreement with the fact that the BACH1:MAFK dimer jointly recognizes both motifs (Newman and Keating, 2003). For known stabilizing, co-binding interactions, such as IRF4:PU.1 or IRF4:BATF, the enrichment of direct motifs (e.g., ETSIII and bZIP as Modules A, respectively) was slightly higher than that of the indirect motifs (e.g., IRF as Module B) (Escalante et al., 2002; Murphy et al., 2013). Therefore, we conclude that the relative enrichment of Modules A and B in these analyses can distinguish which *in vivo* interactions are truly indirect TF-DNA interactions, versus stabilized, DNA co-binding events.

To further investigate the identified indirect interactions, we analyzed the degree of overlap between the sets of ChIP-seq peaks bound by Tethered TF A and Tethering TF B (Table S7). We examined all the potential Tethering TFs that were also assayed by ChIP-seq in the same cell line as was the Tethered TF A (Table S3) and selected the Tethering TF B as the one that shared the largest percentage of peaks with Tethered TF A (STAR Methods). Overall, strong enrichment of the indirect motif was consistent with a high degree of peak overlap,

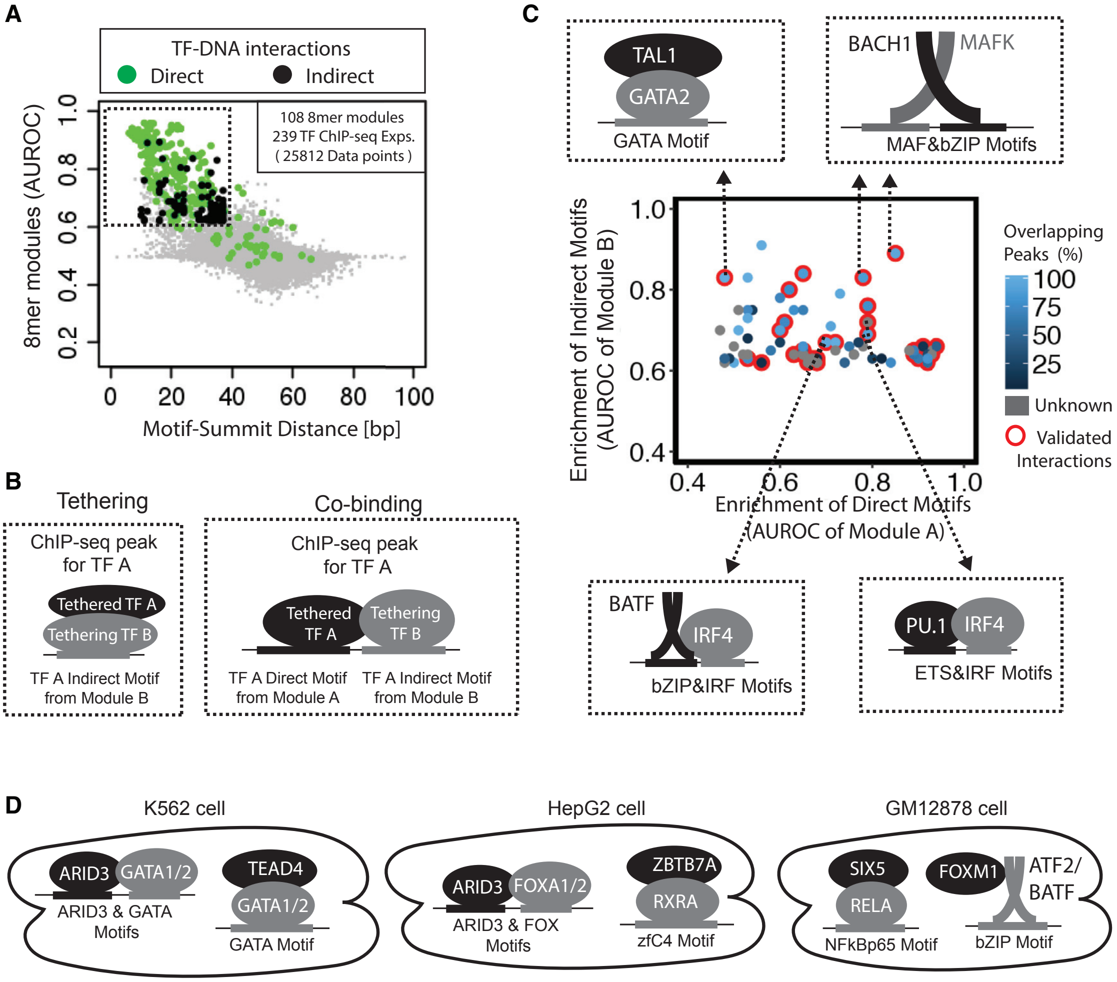


Figure 6. Enrichment of Modules for Indirect TF Binding Predicts TF Co-occupancy

(A) Scatterplot as in Figure 4A, highlighting indirect or direct TF binding (Table S6) as determined by the TF's known cognate motif (direct) or a different motif (indirect).

(B) Cartoons depicting two possible modes of indirect binding: fully tethering or cooperative co-binding.

(C) Indirect interactions from (A). Enrichment of the direct binding motif of the ChIPed Tethered TF A (AUROC of module A) (x axis) and of its indirect binding motif (AUROC of module B) (y axis). Blue scale indicates the percentage of overlapping peaks shared by the Tethered TF A and a candidate Tethering TF B associated with Indirect Module B (Table S7). Gray points signify that no putative tethering TF was assayed by ChIP-seq (Table S3). Cartoons depict literature validated (denoted by red circle) molecular models following the coloring scheme in (B).

(D) Examples of indirect interactions, which to our knowledge have not been previously reported, identified by integrating motif enrichment results with the extent of ChIP-seq peak overlap (Tables 1 and S8).

corroborating the identified indirect interactions (Figure 6C). For many indirect interactions, we observed that cutoffs larger than 5,000 peaks gave a nearly complete overlap in the sets of peaks bound by TFs A and B (Table S7), such as for TAL1 and GATA2 in K562 cells (top left, Figure 6C). In cases where the degree of peak overlap was intermediate, we investigated whether the presence of the indirect motif distinguished overlapping peaks (Table S8) and was thus more likely to indicate a tethering interaction. For most of these interactions (37 of 44), we observed significant co-occurrence ($p < 0.05$) of the indirect motif and the ChIP-seq peaks of the tethering TF, such as for BACH1 being tethered by MAFK in H1-hESC (top right, Figure 6C).

Overall, our analysis suggests previously unknown TF tethering interactions (Table 1; Figure 6D). For example, ARID3 may co-bind DNA through the lineage-specifying factors GATA1/2 in K562 cells and FOXA1/2 in HepG2 cells. TEAD4 seems fully tethered by GATA1/2 in K562 cells, but in HepG2 and H1-hESC cells it binds DNA directly (Table S4). In HepG2 cells, ZBTB7A does not recognize DNA directly, but through the tethering factor RXRA, and not the lineage-specifiers FOXA1/2. Similarly, in GM12878 cells, SIX5 and FOXM1 bind DNA indirectly through NFkBp65 and bZIP motifs respectively, rather than by the lineage-specifier IRF TFs.

DISCUSSION

Tethering interactions expand the combinatorial complexity of the underlying regulatory networks, allowing genes to be

Table 1. Previously Proposed or Novel TF Tethering Interactions Predicted Using the Glossary 8mer Modules with GENRE Background

Tethered TF	Direct Module	Cell Line	Direct AUROC	Indirect Module	Indirect AUROC	Tethered TF Peaks with Indirect Motif	Tethering TFs (Peak Overlap Percentages)
Motif Enrichment Consistent with Previously Proposed TFs							
ATF3	bZIP	GM12878	0.60	bHLH	0.75	64.0%	USF1 (71%), MAX (70%), USF2 (64%), CMYC (25%)
		H1-hESC	0.65		0.75	63.6%	USF1 (67%), USF2 (65%), MAX (58%)
		HepG2	0.63		0.73	63.2%	MAX (79%), USF1 (66%), USF2 (56%)
		K562	0.71		0.71	59.8%	MAX (93%), USF1 (63%), USF2 (46%)
ESRRA	ESRR	HepG2	0.82	HNF4	0.62	40.0%	HNF4A (75%), HNF4G (74%)
FOSL2	bZIP	HepG2	0.91	FOX	0.63	66.2%	FOXA1 (81%), FOXA2 (71%)
NFE2	bZIP	GM12878	0.56	bHLH	0.89	87.0%	USF2 (100%), USF1 (98%)
RFX5	RFX	HeLa	0.76	bZIP	0.64	43.0%	FOS (26%)
		HepG2	0.79		0.62	35.0%	CJUN (10%)
		K562	0.74		0.62	37.0%	FOS (37%), CJUN (20%)
RXRA	zfC4	HepG2	0.78	FOX	0.68	72.6%	FOXA1 (90%), FOXA2 (83%)
SP1	KLF	HepG2	0.53	FOX	0.68	70.8%	FOXA1 (81%), FOXA2 (78%)
TCF12	E2A	HepG2	0.53	FOX	0.73	85.4%	FOXA1 (98%), FOXA2 (94%)
Previously Proposed Motif, Newly Predicted Corresponding TFs							
ATF3	bZIP	GM12878	0.60	bHLH	0.75	64.0%	BHLHE40 (58%)
		HepG2	0.63		0.73	63.2%	BHLHE40 (69%)
		K562	0.71		0.71	59.8%	BHLHE40 (73%)
CEBPB	CEBPB	GM12878	0.57	ETS	0.64	75.8%	ETS1 (17%), ELK1 (14%)
CJUN	bZIP	H1-hESC	0.75	TEA	0.66	47.4%	TEAD4 (61%)
NFE2	bZIP	GM12878	0.56	bHLH	0.89	87.0%	MAX (78%), BHLHE40 (69%), CMYC (17%)
PRDM1	NA	HeLa	NA	bZIP	0.66	43.6%	FOS (27%)
RFX5	RFX	HepG2	0.79	bZIP	0.62	35.0%	FOSL2 (23%)
		K562	0.74		0.62	37.0%	ATF1 (46%), ATF3 (27%), JUNB (16%), FOSL1 (10%)
		GM12878	0.77	Hbox	0.67	58.0%	PBX3 (22%)
TEAD4	TEA	K562	0.53	GATA	0.83	96.6%	GATA2 (100%), GATA1 (87%)
Novel Motif and Interacting TFs							
ARID3	ARID3	HepG2	0.60	FOX	0.69	88.8%	FOXA1 (92%), FOXA2 (90%)
		K562	0.58	GATA	0.64	70.2%	GATA2 (62%), GATA1 (33%)
FOXM1	FOXM	GM12878	0.50	bZIP	0.62	37.8%	ATF2 (97%), BATF (94%)
MEF2A	NA	K562	NA	bZIP	0.64	39.6%	ATF1 (52%), ATF3 (50%), JUNB (41%), CJUN (40%), FOSL1 (35%), FOS (20%)
NR2F2	zfC4	K562	0.72	GATA	0.79	87.8%	GATA2 (97%), GATA1 (67%)
NRF1	NA	K562	NA	bHLH	0.64	47.6%	MAX (78%), BHLHE40 (47%), CMYC (42%), USF1 (29%)
		H1-hESC	NA		0.64	47.0%	MAX (28%), USF1 (25%), USF2 (18%)
		HeLa	NA		0.63	46.6%	MAX (81%), USF2 (27%)
SIX5	SIX	GM12878	0.49	NFkBp65	0.63	43.4%	RELA (20%)
TCF12	E2A	A549	0.51	bZIP	0.70	49.2%	FOSL2 (98%), ATF3 (96%)
ZBTB7A	zfC2H2_EGR_RHD	HepG2	0.69	zfC4	0.63	78.8%	RXRA (47%)

Tethered TF (or TF A) represents a ChIPed TF with an Indirect Module (or Module B) from the 8mer Glossary both enriched (Indirect AUROC) and centered in the best 500 peaks of TF A; its Direct Module (or Module A) is used to calculate the Direct AUROC. Tethered TF peaks with Indirect Motif returns the percentage of the best 500 peaks of TF A that contain 8mer(s) from Module B. Furthermore, Tethering TF (or TF B) represents a candidate TF (Table S3) from Indirect Module B, with the associated Peak Overlap Percentage between the best 500 peaks of Tethered TF A and all peaks of Tethering TF B (adjusted Fisher's exact test $P < 0.05$). Interactions without a TF B ChIP-seq dataset currently available (or none with significance) can be found in Table S7. NA, not applicable.

(legend continued on next page)

co-regulated by multiple TFs (Jolma et al., 2015; Spitz and Furlong, 2012; Wang et al., 2012a). For reasons of computational efficiency in the detection of these combinatorial *cis* regulatory codes, a reduced core set of non-redundant motifs is often more desirable than an exhaustive motif dictionary. Compared with prior approaches (Gisselbrecht et al., 2013; Kheradpour et al., 2007; Wang et al., 2012a), here we assembled a compact glossary of 108 TF binding motifs by clustering the intrinsic 8mer specificities obtained for individual TFs by PBMs. The primary limitation of our glossary is that it comprises only TF-8mer binding data obtained from universal PBMs, which are limited to short (<8 bp) motifs. Although a minority among human TFs, there are TFs with longer sequence recognition, such as P53 or CTCF, which are missing from the current glossary. For identification of longer binding sites, the glossary could be expanded in the future to incorporate data from other types of custom arrays (Siggers et al., 2011a, 2011b) and from SELEX experiments (Jolma et al., 2013, 2015).

In this study, we also developed GENRE, a construction method for matched genomic background sequences with the highest flexibility. In our comparative analysis, GENRE overall outperforms four popular background approaches. Strikingly, flanking genomic regions underperformed by far when measuring indirect binding enrichment (Figure 5B), potentially because they did not recapitulate the interdependencies between bias-prone features of ChIP-seq “bound” regions (Figure 2E). In contrast, GENRE controls for both the individual feature distributions and their interdependencies, thus giving backgrounds that precisely differentiate between noise and biologically relevant weak signals (Figures 3C and 5B). To our knowledge, GENRE allows matching on more features than prior methods for background construction. HOMER can account for at most two features (GC content or CpG frequency) within sequences that are either promoters or random genomic sequences, while BiasAway matches only GC content. While by default GENRE controls for four potential biases (GC content, CpG frequency, promoter overlap, and repeat overlap), our understanding of what genomic features are important for regulatory function is incomplete, so we made GENRE readily tunable to control for additional sequence features given by the user. In the future, GENRE could be further developed to build background sequences from: (1) data from other sequence-based techniques (such as DNase-seq or ATAC-seq), which may differ from ChIP-seq peaks in their sequence biases, and (2) different species, which may exhibit different biases in sequence composition.

Agreement on evaluation criteria for assessing the performance of background sets or models of binding specificity in motif enrichment analysis remain open questions. Compared with other criteria utilized previously in assessing motif enrichment (Worsley Hunt et al., 2014), our evaluation criteria offer a biologically intuitive interpretation in terms of sensitivity (AUROC distance correlation) and specificity (AUROC MAD) of the motif binding response. While AUROC distance correlation quantifies the enrichment of motifs located toward the center of the peak, i.e., those likely anchoring the ChIPed TF to the DNA, a low AUROC MAD signifies that few motifs are enriched within regulatory regions, reflecting that only a minority of TFs controls the transcriptional programs underlying cellular phenotypes. Our criteria have the further advantage of utilizing AUROC values, an accepted statistical metric in motif enrichment analysis (Gordan et al., 2009; Weirauch et al., 2013). In addition to these criteria, we used the TF expression profiles in various cell lines as an independent biological feature to evaluate which TF suggested by the enriched indirect binding motifs is likely the tethering TF. This approach could be improved in future studies by integrating additional information, such as protein-protein interaction data and chromatin-accessibility profiles, to provide further insights into the underlying regulatory networks of tissue-specific gene-expression programs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - ○ Protein Binding Microarray (PBM) Experiments
  - ○ PBM Data Processing
  - ○ PBM Evaluation for DNA-Binding Specificity
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Construction of the Glossary
  - ○ Motif Enrichment Analysis in ChIP-seq Peaks
  - ○ Building Background Sets for ChIP-seq Peaks
  - ○ GENRE Implementation and Benchmarking
  - ○ Motif Enrichment of Highly Expressed TFs
- DATA AND SOFTWARE AVAILABILITY
  - ○ PBM Data
  - ○ Software

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2017.06.015.

AUTHOR CONTRIBUTIONS

Conceptualization, L.M., K.W., and M.L.B.; Methodology, L.M., K.W., and L.A.B.; Software, L.M. and K.W.; Formal Analysis, L.M. and K.W.; Investigation, L.M., K.W., and A.V.; Resources, M.L.B.; Data Curation, L.M., K.W., and L.A.B.; Writing – Original Draft, L.M., K.W., and M.L.B.; Writing – Review & Editing, L.M., K.W., and M.L.B.; Visualization, L.M., K.W., and M.L.B.; Supervision, L.M. and M.L.B.; Funding Acquisition, M.L.B.

We described these interactions in three categories based on previous literature evidence. The Motif Enrichment Consistent with Previously Proposed TF subset summarizes Tethered TF A and Tethering TF B interactions proposed in (Wang et al., 2012b) and (Neph et al., 2012) and recapitulated by our analysis. The Previously Proposed Motif, Newly Predicted Corresponding TF subset shows novel Tethered-Tethering TF pairs from our analysis that agreed with previously proposed presence of an Indirect Module B motif in tethering TF A (references in Table S6). In the Novel Motif and Interacting TF subset, we present novel interactions between Tethered TF A and Indirect Module B/Tethering TF B from this analysis.

ACKNOWLEDGMENTS

This project was supported by NIH grants R01 HG003985 and R01 HG005287 to M.L.B. and P50 HG006193 to Aviv Regev and M.L.B., and a National Science Foundation Graduate Research Fellowship to L.A.B. We thank Max Hume for technical assistance, Tim Hughes for sharing expression clones and/or proteins, Steve Gisselbrecht for helpful discussion and reading the manuscript, and Julia Maria Rogers and Kian Hong Kock for helpful discussion. An early version of this paper was submitted to and peer reviewed at the 2017 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at *Cell Systems*. M.L.B. is a co-inventor on U.S. patent #8,530,638 on universal PBM technology.

Received: February 9, 2017
Revised: June 3, 2017
Accepted: June 29, 2017
Published: September 27, 2017

REFERENCES

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. (2015). Deconvolving the recognition of DNA shape from sequence. Cell *161*, 307–318.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Arvey, A., Agius, P., Noble, W.S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res. *22*, 1723–1734.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.

Bailey, T.L., and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res. *40*, e128.

Barrera, L.A., Vedenko, A., Kurland, J.V., Rogers, J.M., Gisselbrecht, S.S., Rossin, E.J., Woodard, J., Mariani, L., Kock, K.H., Inukai, S., et al. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science *351*, 1450–1454.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Peña-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

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

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Bhasin, J.M., and Ting, A.H. (2016). Goldmine integrates information placing genomic ranges into meaningful biological contexts. Nucleic Acids Res. *44*, 5550–5556.

Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A., Schiltz, R.L., Miranda, T.B., Sung, M.H., Trump, S., Lightman, S.L., et al. (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Mol. Cell *43*, 145–155.

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front. Genet. *7*, 24.

Bresnick, E.H., Katsumura, K.R., Lee, H.-Y., Johnson, K.D., and Perkins, A.S. (2012). Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. Nucleic Acids Res. *40*, 5819–5831.

Davies, D.L., and Bouldin, D.W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. *1*, 224–227.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc. Natl. Acad. Sci. USA *99*, 7554–7559.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Escalante, C.R., Brass, A.L., Pongubala, J.M., Shatova, E., Shen, L., Singh, H., and Aggarwal, A.K. (2002). Crystal structure of PU.1/IRF-4/DNA ternary complex. Mol. Cell *10*, 1097–1105.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to identify peaks from ChIP-Seq data. Curr. Protoc. Bioinformatics *Chapter 2*. Unit 2 14.

Fong, A.P., Yao, Z., Zhong, J.W., Johnson, N.M., Farr, G.H., 3rd, Maves, L., and Tapscott, S.J. (2015). Conversion of MyoD to a neurogenic factor: binding site specificity determines lineage. Cell Rep. *10*, 1937–1946.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Gisselbrecht, S.S., Barrera, L.A., Porsch, M., Aboukhalil, A., Estep, P.W., 3rd, Vedenko, A., Palagi, A., Kim, Y., Zhu, X., Busser, B.W., et al. (2013). Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. Nat. Methods *10*, 774–780.

Glasmacher, E., Agrawal, S., Chang, A.B., Murphy, T.L., Zeng, W., Vander Lugt, B., Khan, A.A., Ciofani, M., Spooner, C.J., Rutz, S., et al. (2012). A genomic regulatory element that directs assembly and function of immune-specific AP-1-IRF complexes. Science *338*, 975–980.

Gordan, R., Hartemink, A.J., and Bulyk, M.L. (2009). Distinguishing direct versus indirect transcription factor-DNA interactions. Genome Res. *19*, 2090–2100.

Gordan, R., Murphy, K.F., McCord, R.P., Zhu, C., Vedenko, A., and Bulyk, M.L. (2011). Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol. *12*, R125.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. *3*, 1093–1104.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J.M. (2009). A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. Cell *138*, 314–327.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. *43*, D117–D122.

Hung, J.H., and Weng, Z. (2017). Motif finding. Cold Spring Harb. Protoc. *2017*, pdb.top093195.

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics *9*, 192.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. Genome Res. *17*, 1919–1931.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics *24*, 719–720.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.

Lee, C.S., Friedman, J.R., Fulmer, J.T., and Kaestner, K.H. (2005). The initiation of liver development is dependent on Foxa transcription factors. Nature *435*, 944–947.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Loytynoja, A., Vilella, A.J., and Goldman, N. (2012). Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics *28*, 1684–1691.

Luo, Z., Gao, X., Lin, C., Smith, E.R., Marshall, S.A., Swanson, S.K., Florens, L., Washburn, M.P., and Shilatifard, A. (2015). Zic2 is an enhancer-binding factor required for embryonic stem cell specification. Mol. Cell *57*, 685–694.

McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165.

Mendez, A., and Mendoza, L. (2016). A network model to describe the terminal differentiation of B cells. PLoS Comput. Biol. *12*, e1004696.

Murphy, T.L., Tussiwand, R., and Murphy, K.M. (2013). Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. Nat. Rev. Immunol. *13*, 499–509.

Nagy, P., Bisgaard, H.C., and Thorgeirsson, S.S. (1994). Expression of hepatic transcription factors during liver development and oval cell differentiation. J. Cell Biol. *126*, 223–233.

Nakagawa, S., Gisselbrecht, S.S., Rogers, J.M., Hartl, D.L., and Bulyk, M.L. (2013). DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc. Natl. Acad. Sci. USA *110*, 12349–12354.

Nekrutenko, A., and Li, W.H. (2000). Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. *10*, 1986–1995.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Newman, J.R., and Keating, A.E. (2003). Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science *300*, 2097–2101.

Orenstein, Y., and Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. Nucleic Acids Res. *42*, e63.

Peterson, K.A., Nishi, Y., Ma, W., Vedenko, A., Shokri, L., Zhang, X., McFarlane, M., Baizabal, J.-M., Junker, J.P., van Oudenaarden, A., et al. (2012). Neural-specific Sox2 input and differential Gli-binding affinity provide context and positional information in Shh-directed neural patterning. Genes Dev. *26*, 2802–2816.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. *12*, 32–42.

Quinlan, A.R. (2014). BEDTools: the Swiss-Army Tool for genome feature analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1–11.12.34.

Setty, M., and Leslie, C.S. (2015). SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. PLoS Comput. Biol. *11*, e1004271.

Shiina, M., Hamada, K., Inoue-Bungo, T., Shimamura, M., Uchiyama, A., Baba, S., Sato, K., Yamamoto, M., and Ogata, K. (2015). A novel allosteric mechanism on protein-DNA interactions underlying the phosphorylation-dependent regulation of Ets1 target gene expressions. J. Mol. Biol. *427*, 1655–1669.

Siggers, T., Chang, A.B., Teixeira, A., Wong, D., Williams, K.J., Ahmed, B., Ragoussis, J., Udalova, I.A., Smale, S.T., and Bulyk, M.L. (2011a). Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-κB family DNA binding. Nat. Immunol. *13*, 95–102.

Siggers, T., Duyzend, M.H., Reddy, J., Khan, S., and Bulyk, M.L. (2011b). Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Mol. Syst. Biol. *7*, 555.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., et al. (2016). The UCSC genome browser database: 2016 update. Nucleic Acids Res. *44*, D717–D725.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Spruijt, C.G., and Vermeulen, M. (2014). DNA methylation: old dog, new tricks? Nat. Struct. Mol. Biol. *21*, 949–954.

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. *20*, 267–273.

Takaku, M., Grimm, S.A., and Wade, P.A. (2015). GATA3 in breast cancer: tumor suppressor or oncogene? Gene Expr. *16*, 163–168.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. (1997). The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. EMBO J. *16*, 3145–3157.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012a). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812.

Wang, Z., Oron, E., Nelson, B., Razis, S., and Ivanova, N. (2012b). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. Cell Stem Cell *10*, 440–454.

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Worsley Hunt, R., Mathelier, A., Del Peso, L., and Wasserman, W.W. (2014). Improving analysis of transcription factor binding sites within ChIP-seq data based on topological motif enrichment. BMC Genomics *15*, 472.

Yagi, R., Kohn, M.J., Karavanova, I., Kaneko, K.J., Vullhorst, D., DePamphilis, M.L., and Buonanno, A. (2007). Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. Development *134*, 3827–3836.

Young, R.A. (2011). Control of the embryonic stem cell state. Cell *144*, 940–954.

Zhou, C., Yang, X., Sun, Y., Yu, H., Zhang, Y., and Jin, Y. (2016). Comprehensive profiling reveals mechanisms of SOX2-mediated cell fate specification in human ESCs and NPCs. Cell Res. *26*, 171–189.

Zhu, X., Ahmad, S.M., Aboukhalil, A., Busser, B.W., Kim, Y., Tansey, T.R., Haimovich, A., Jeffries, N., Bulyk, M.L., and Michelson, A.M. (2012). Differential regulation of mesodermal gene expression by *Drosophila* cell type-specific Forkhead transcription factors. Development *139*, 1457–1466.

Zon, L.I. (2008). Intrinsic and extrinsic control of haematopoietic stem-cell self-renewal. Nature *453*, 306–313.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-glutathione S-transferase, rabbit IgG fraction, Alexa Fluor 488 conjugate	Invitrogen	A11131; RRID: AB_2534137
Bacterial and Virus Strains		
E. coli C41 DE3 cells	Lucigen	60444
Chemicals, Peptides, and Recombinant Proteins		
Cy3-conjugated dUTP	GE Healthcare	PA53022
Protease, from Streptomyces griseus	Sigma	P6911
Thermo sequenase cycle sequencing kit	USB	78500
Deposited Data		
PBM data	UniPROBE	http://the_brain.bwh.harvard.edu/uniprobe/
PBM data	CIS-BP	http://cisbp.ccbr.utoronto.ca/
Human reference genome hg19	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/
ChIP-seq datasets	ENCODE	http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform
hg19 blacklisted regions	ENCODE	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz
hg19 uniquely mappable regions	ENCODE	https://personal.broadinstitute.org/anshul/projects/umap/encodeHg19Female/globalmap_k20tok54.tgz
Oligonucleotides		
HPLC-purified primer (unmodified) for double-stranding of DNA oligonucleotide array 5’-CAGCACGGACAACGGAACACAGAC-3’	Integrated DNA Technologies	https://www.idtdna.com/site
Software and Algorithms		
Masliner (v1.02)	(Dudley et al., 2002)	http://arep.med.harvard.edu/masliner/pgmlicense.html
Universal PBM Analysis Suite	(Berger and Bulyk, 2009)	http://the_brain.bwh.harvard.edu/PBMAnalysisSuite/index.html
R		https://www.r-project.org/
Python 2		https://www.python.org/
BEDTools	(Quinlan, 2014)	https://github.com/arq5x/bedtools2/
samtools	(Li et al., 2009)	https://github.com/samtools/samtools
SQLite	Hwaci	http://sqlite.org/
SQLite extension	Liam Healy	https://github.com/rstats-db/RSQLite/blob/master/src/vendor/sqlite3/extension-functions.c
uShuffle	(Jiang et al., 2008)	http://digital.cs.usu.edu/~mjiang/ushuffle/
HOMER	(Heinz et al., 2010)	http://homer.ucsd.edu/homer/download.html
BiasAway	(Worsley Hunt et al., 2014)	https://github.com/wassermanlab/BiasAway/
DeepBind	(Alipanahi et al., 2015)	http://tools.genes.toronto.edu/deepbind/
PAGAN	(Loytynoja et al., 2012)	http://wasabiapp.org/download/pagan/
MAFFT	(Katoh and Standley, 2013)	http://mafft.cbrc.jp/alignment/software/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GENRE	This paper	http://thebrain.bwh.harvard.edu/glossary-GENRE/download.html
TF-8mer glossary	This paper	http://thebrain.bwh.harvard.edu/glossary-GENRE/download.html
Other		
Custom-designed “universal all 10-mer” oligonucleotide arrays	Agilent Technologies	AMADID #015681 AMADID #016060 AMADID #030236
ActivePro in vitro transcription and translation kit	Ambion	AM1295 (Out of production)
PURExpress in vitro transcription and translation kit	NEB	E6800

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Martha L. Bulyk (mlbulyk@genetics.med.harvard.edu).

METHOD DETAILS

Protein Binding Microarray (PBM) Experiments

We expressed proteins by either (i) purification from *E. coli* C41 DE3 cells (Lucigen), or (ii) *in vitro* translation reactions (Ambion ActivePro Kit and NEB PURExpress) without purification, both as previously described (Badis et al., 2009; Berger et al., 2008). Briefly, GST-TF fusion proteins were purified from *E. coli* overexpression cultures by GST affinity chromatography. The quality and concentration of each such GST-tagged protein was estimated relative to a dilution series of GST standards on Coomassie-stained SDS-PAGE gels. Alternatively, GST-TF fusion proteins were expressed by *in vitro* transcription and translation following the manufacturer’s protocol without subsequent purification. Western blots were performed to assess the quality and to approximate the concentration of each resulting GST-TF fusion protein relative to a dilution series of recombinant GST standards. Custom-designed oligonucleotide arrays (Agilent) were double-stranded and PBM experiments were performed following previously described experimental protocols (Berger et al., 2006). The array designs employed were “all 10mer” universal arrays (Agilent Technologies). Experimental conditions used for all PBM experiments, including gene names, protein expression method, Agilent AMADID numbers, TF concentrations, cloning sequences and TF classes are described in Table S1.

PBM Data Processing

PBM scan images were obtained using a GenePix 4000A Microarray Scanner (Molecular Devices). The resulting image data were processed using GenePix Pro v7.2 to obtain signal intensity data for each spot. The data were then further processed using Masliner software (v1.02) (Dudley et al., 2002) to combine scans from different intensity settings, increasing the effective dynamic range of the signal intensity values. If a dataset had any negative background-subtracted intensity (BSI) values (which can occur if the region surrounding a spot is brighter than the spot itself), all BSI values were scaled such that they all became non-negative. All BSI values were normalized using the software for spatial de-trending provided in the Universal PBM Analysis Suite (Berger and Bulyk, 2009), as previously described (Berger et al., 2006). Briefly, Alexa488 signal was normalized by Cy3 signal for each spot to account for differences in the amounts of double-stranded DNA. To correct for any possible non-uniformities in protein binding, we further adjusted the Cy3-normalized Alexa488 signals according to the position of each spot within a 15 x 15 block centered on each spot on the microarray.

PBM Evaluation for DNA-Binding Specificity

We used the Universal PBM Analysis Suite (Berger and Bulyk, 2009) to calculate the PBM enrichment score (E-score) for each of the 32,768 non-redundant, ungapped 8mers for each protein. The E-score is a rank-based statistic that is closely related to the area under the receiver operating characteristic (ROC) curve and robust to technical variation across arrays (Berger and Bulyk, 2009). Larger E-score values reflect higher specificity for binding a particular 8mer. The presence of E-scores $\geq$ 0.45 has been reported as a viable quality control metric to identify successful PBM experiments (Berger et al., 2008; Weirauch et al., 2014). Here, we deemed a PBM experiment to be of acceptable quality if it contained at least five 8mers with an E-score $\geq$ 0.45.

QUANTIFICATION AND STATISTICAL ANALYSIS

Construction of the Glossary

Data Mining and Filtering

We collected 773 good quality PBM arrays (*i.e.*, those that contained at least five 8mers with an E-score $\geq$ 0.45) representing the sequence specificity of more than 671 metazoan TFs from our unpublished experiments, CIS-BP (Weirauch et al., 2014) database, and UniPROBE (Badis et al., 2009) database, noting gene names, organism of origin, original publication and TF classes (Table S1). We combined the E-scores of these PBM arrays into a single data table with each PBM experiment retained in a separate column and each 8mer in a separate row. We used the rank-based PBM enrichment score (E-score) to quantify the preference of a TF for each 8mer (Berger et al., 2006) and used an E > 0.3 threshold to capture preferences for binding to a wide range of affinity sequences (Barrera et al., 2016; Berger et al., 2008). We considered all 8mers with E-score > 0.3 as ‘bound’ by an assayed protein (Barrera et al., 2016), and discarded 8mers that did not exhibit an E-score > 0.45 in at least one experiment. This filter reduced the initial number of 8mers to 10,926.

First Clustering

We independently clustered the rows and the columns of these data using Pearson correlations as distance and hierarchical agglomerative complete-linkage (‘‘hclust’’ function in R). From them, we obtained one dendrogram for the 8mers and one for the PBMs (Figure S1A). We created several independent collections of 8mer clusters by repeatedly cutting the 8mer dendrogram with the function ‘‘cutreeDynamicTree’’ (from R package ‘‘dynamicTreeCut’’, setting: maxTreeHeight=0.9, deepSplit=FALSE). Each dendrogram treecut differed by the minimal cluster size (CMS), which we varied between 20 and 100 8mers (red to orange bars, Figure 1B). To associate specific PBM experiments to each 8mer cluster, we sorted the subtable defined by its 8mers and applied a 2-mean-clustering on the E-score profile to discriminate PBM exhibiting high versus low specificity. The highly specific PBMs then were coupled to the 8mer cluster, to form an 8mer-PBM cluster (black boxes in Figure S1B). We grouped the 8mer-PBM clusters obtained by each individual CMS to form a single CMS collection (i.e., ‘‘20 CMS collection’’). We evaluated the different cluster collections by the Davies-Bouldin index (DBI) (Davies and Bouldin, 1979), which compares the inter-cluster E-score difference against the intra-cluster E-score similarity (small DBI for good tree cuts). Since cluster collections with smaller DBI tended to filter out more 8mers, we corrected the DBI by the 8mer coverage (Figure S1C).

Merged Cluster Collection

To create a collection of 8mer-PBM merged clusters from the previous single CMS collections, we matched them through a cumulative pairwise comparison (blue to cyan bars, Figure 1B). For the comparison scheme, we started comparing the 20 and 25 CMS collections, creating a ‘‘20-25’’ intermediate collection. The intermediate ‘‘20-25’’ was then compared with the 30 CMS collection to create the ‘‘20-30’’ intermediate collection. This comparison scheme was iterated up to the 100 CMS collection, creating the ‘‘20-100’’ intermediate collection which, after a further comparison with an asymptotic collection formed by a single cluster for the whole 8mer dendrogram, furnished the final Merged Clusters. For the rules of cluster merging between two collections (A and B), we used the following assumption:

1-to-1 match: if the 8mers of a cluster from A overlapped uniquely with the 8mers of one and only one cluster from B, we merged their 8mers into a single cluster in the next intermediate collection.

Cluster rescue: if all the 8mers of a cluster from A did not fall in any cluster of B, we inserted the cluster in the next intermediate collection.

Neighbor clusters comparison: if the 8mers of two (or more) clusters from collection A jointly overlapped the 8mers of one (or more) clusters from collection B, we cut the 8mer interval in all the possible tree cut intervals suggested by the two tree cuts (dotted lines in Figure 1B), which creates a new set of minimal clusters (green bars in Figure 1B). Within this set, we compared each pair of contiguous 8mer-PBM clusters. To merge two clusters (green arrows in Figure 1B) we first required that their difference in the average E-scores was less than 0.12, as a criterion to ensure that the 8mers of the two clusters exhibit a sufficiently similar overall specificity to the associated TF PBMs. Moreover, within each of the two clusters, we quantified the average E-score of the PBMs over the 8mers. If the overall variability of the average E-score of the PBMs in the two clusters (i.e., the sum of the Standard Deviation of the average E-score for PBMs in the two clusters) was also less than 0.12, we merged the two clusters. If this variability was higher than 0.12, we further required that the average E-score values of the PBMs in the two clusters were positively correlated (Spearman rank correlation coefficient rho (R) > 0.1, $p < 0.05$) to merge 8mer clusters with similar E-score profiles for the associated PBMs. We empirically found that these conditions reliably merged two clusters exhibiting the same E-score pattern across PBM experiments. When compared to the collections obtained by single CMS, the merged cluster collection showed the smallest corrected DBI index (best clustering) (Figure S1C), the highest coverage, and its clusters encompassed the widest range of 8mer sizes (data not shown).

Re-clustering

To ensure convergence and robustness of the clustering, we re-clustered the merged cluster collection. To filter out noise, we first assigned the E-score basal value (0.3) to any 8mer-PWM pair not contained in the merged clusters (*background filtering*). For these filtered data, we obtained 8mer and PBM dendrograms, as described for the first clustering (*re-clustering*). The background E-score values were then re-assigned to each 8mer-PBM pair (*background rescue*), and the dendrogram were repeatedly cut and compared to create the merged clusters collection. We renamed this collection as ‘‘glossary’’, and its 8mer-PBM merged cluster as ‘‘TF-8mer module’’, naming each module according to the class of the TFs assayed by the most specific PBMs. Overall, the glossary contained 671 non-redundant TFs from 715 PBM datasets and 10,428 8mers. Importantly, each 8mer is assigned to just one module, but the same PBM (and therefore its assayed TF) can be associated with several modules (*right panel* Figure 1C, *middle-right panels* in Figure S1D).

Evaluation of Module Purity for TF Class

To evaluate how strongly each module was associated to a single TF class, we defined a "Purity Index" for the modules as follows. If all the PBMs associated with a certain module assayed TFs from the same class, then that module's Purity Index was assigned as 100% (*i.e.*, TEA, IRF and FOX modules in Figure S1D). If the module's PBMs were assaying TFs from several TF classes, each PBM in the module was assigned its average 8mer E-score (*i.e.*, E-score$_{PBM}$ = <E-scores(8mers$_{cluster}$, PBM)>). Each TF class was then assigned the highest value of its associated PBMs (*i.e.,* E-score$_{TFclassA}$ = max(E-score$_{TFclassA\_PBM1}$, E-score$_{TFclassA\_PBM2}$, ...). We than ranked the E-scores assigned to all the TF classes presented in the module in a decreasing order (*i.e.,* E-score$_{First\_TFclass}$ > E-score$_{Second\_TFclass}$ > ...> E-score$_{Last\_TFclass}$) and defined the Purity Index as following:

$$\text{Module Purity Index} = (\text{E-score}_{First\_TFclass} - \text{E-score}_{Second\_TFclass})/(\text{E-score}_{First\_TFclass} - \text{E-score}_{Last\_TFclass})$$

Derived PWMs from Glossary's 8mer Modules

To build a PWM for each module, we ranked 8mers by their average E-score. We oriented the ranked 8mers using the program PAGAN(Loytynoja et al., 2012), aligned them to create a PWM with the program MAFFT(Katoh and Standley, 2013), and visualize the resulting PWM via the R-package seqLogo (Table S2).

Motif Enrichment Analysis in ChIP-seq Peaks

To examine ChIP-seq peaks for the enrichment of a TF binding specificity motif, represented either by a PWM, an 8mer module, or a DeepBind motif model, we calculated the area under receiver operating characteristic curve (AUROC) statistics to evaluate the enrichment (Weirauch et al., 2013) in the 500 ChIP-seq peak sequences with the highest fold-change above the local distribution of reads modeled by a Poisson distribution (Feng et al., 2011). We trimmed the peaks to encompass a maximum of 200 bp by using the position of the peak center (also referred in the text as peak summit) provided by the original datasets. As the background set, we generated an equally sized set of 500 sequences with one of the methods for background construction discussed in the text and below.

To compute a PWM-based AUROC, each sequence in the foreground and background sets was scanned by the PWM (function "PWMmatch" in R package Biostrings) and then assigned a score corresponding to the best PWM match (function "PWMscoreStartingAt" in R package Biostrings). The position of the best PWM match within the sequence was also recorded and used to evaluate the motif distance from the peak summit ("Motif-Summit Distance"). The AUROC statistic was obtained by calculating sensitivity and specificity values as the score threshold for predicting a region to be bound was varied between 0 and the max score.

To compute 8mer-based AUROC for a glossary's module, we first looked through each foreground and background sequence for matches to the module's 8mers, scoring each sequence by the highest E-scores of its matching 8mers. This E-score value represented the binding probability of the sequence to be used in the AUROC analysis. As previously described (Barrera et al., 2016), the AUROC statistic was obtained by calculating sensitivity and specificity values as the E-score threshold for predicting a region to be bound was varied between 0.3 and 0.5 (the range of E-scores used in this study). The P-values associated with each AUROC value were calculated by using a Wilcoxon signed-rank test comparing the scores for foreground and background sequences, which we adjusted with a False Discovery Rate test for multiple hypotheses along the 108 motifs of the glossary. The position of the best-matched 8mer was also used to evaluate the Motif-Summit distance.

For each background method and TF specificity model, we evaluated across all the combinations of one ChIP-seq dataset and one glossary module: (1) the median absolute deviation (MAD) of the AUROC values and (2) the Spearman rank correlation coefficient rho (R) between AUROC values and Motif-Summit distance. We computed the AUROC MAD through the function "mad" in R-package "stats," which by definition returns the median of the absolute deviations from the median and adjusts it by a constant for consistency with the standard deviation in case of normal distributions:

$$\text{mad}(Xi) = 1.483*\text{median}(|Xi - \text{median}(Xi)|)$$

For the values in Figures 3C and 4B, the total number of ChIP-seq-module combinations, which are represented as data points in the plots, are:

$$239 \text{ ChIP-seq datasets x } 108 \text{ TF-8mer modules} = 25{,}812$$

For the TF-family subsets in Figures 3D, 3E, and S4, the total number of ChIP-seq-module combinations varied between 108 (MYB) and 3,780 (bHLH).

To compute the AUROC for sequences analyzed using DeepBind models (Alipanahi et al., 2015), each sequence in the foreground and background sets was scanned using the open source C-code "DeepBind.c" provided in the DeepBind webpage (http://tools.genes.toronto.edu/deepbind/). We used the code's standard mode, which returns the DeepBind score of the best match. We also developed a "position" option in the "DeepBind.c" code that returns the distance between the best match and the peak summit. Similarly to the PWM case, the AUROC statistic was obtained by calculating sensitivity and specificity values as the threshold of the DeepBind score was varied between 0 and the max score.

In the case of Figure S5, since just 13 PBM datasets representing modules were available as DeepBind motif models, the total number of ChIP-seq-module combinations for each was (239*13) = 3,107.

Building Background Sets for ChIP-seq Peaks

From the ''foreground'' dataset of genomic regions defined as a BED file, we obtained the actual sequences with the ''getfasta'' command from BEDTools (Quinlan, 2014) using the unmasked genome as input. Using the BED file and fasta file as required, we created the following background sets:

Dinucleotide Shuffle

We obtained the permuted sequences with identical dinucleotide frequencies and length using uShuffle software (Jiang et al., 2008).

Flanking Regions

Genomic regions 1000 bp upstream from the ChIP-seq peak were obtained using the ''getfasta'' command as for the foreground. We decided to use 1000 bases as an intermediate flanking distance among several proposed in similar analysis (Bailey and Machanick, 2012; Orenstein and Shamir, 2014; Siggers et al., 2011a; Wang et al., 2012a), ranging from 300 to 2000 bp. Notably, the ChIP-seq peaks in the foreground datasets have a median size of 305 bp (+/- 110), ensuring that the flanking regions do not overlap with the actual peaks.

GC-BiasAway

We downloaded BiasAway (Worsley Hunt et al., 2014) from https://github.com/wassermanlab/BiasAway/ and the human 200 bp background repository from http://cisreg.cmmt.ubc.ca/BiasAway_background/. After learning that the program was not equipped to handle a directory as a background, we randomly sampled 5000 sequences from each repository 1% bin file and put them into a single fasta file to be used as the background pool file. The BiasAway subcommand ‘g’ (%GC distribution-based background chooser) was used for each foreground ChIP set.

GC-HOMER

We used HOMER (Heinz et al., 2010) version 4.6 subprogram findMotifsGenome.pl with the arguments hg19, -size 200, -N 500 to coincide with our foreground set; -dumpFasta to obtain the background fasta file; -nomotif and -noknown to avoid any *de novo* motif searching.

GENRE

We used the hg19 uniquely mappable regions (ENCODE Project Consortium, 2012) subtracted by hg19 blacklisted regions (ENCODE Project Consortium, 2012) as the BED file to be tiled for possible background regions (https://personal.broadinstitute.org/anshul/projects/umap/encodeHg19Female/globalmap_k20tok54.tgz; http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz, respectively). The chosen matched features from the UCSC hg19 genome version (Speir et al., 2016) were: promoters (defined as 2 kilobases upstream of transcription start sites (TSS) according to strandedness for known genes), repeat regions (Repeat Masker files), GC content and CpG dinucleotides (unmasked genome). The overall binning scheme that we chose was: independently binning promoters and repeats with binary ranges, and then nesting quartiles of GC content and subsequently CpG dinucleotides. The all-but-one analysis seen in Figure S3 utilized the same nested quartile scheme, but removing one feature, *i.e.*, for leaving promoters out, we binned repeats in a binary manner, and for each repeat bin (0/not0), quartiles were nested first for GC content and then for CpG dinucleotides.

GENRE Implementation and Benchmarking

Database Construction

GENRE is a background generation tool that builds an a priori database of possible background regions that are matched to foreground regions via user-defined features. It has been implemented using Python, SQLite (with a common mathematical and string function library extension from R package RSQLite written by Liam Healy), and BEDTools (Quinlan, 2014). Initially, a BED file is tiled into uniformly sized regions and loaded into a table in the database. That table is then extended to include the real values associated with each feature's overlap percentage as calculated by BEDTools (genomic features via ''intersectBed'' with the feature's BED file; compositional features via ''nucBed'' with a fasta file). The user can then decide how to bin features, either by ranges (i.e. binary – ''0,(0-100]''), equal percent size (*e.g.*, 5% bins), or nested quartiles (*e.g.*, GC quartiles stratified by promoters). Each binning scheme is housed in its own table within the database.

Table Binning – Nested Quartiles

GENRE, being able to match for multiple features, can present with a scarcity problem; lesser populated multiplexed background bins may not contain enough sequences from which to build a background set. While background scarcity was only a minor issue for this study, it could potentially become a larger problem for smaller genomes with fewer possible background sequences and analyses that require a background multiple times bigger than the foreground. Giving a nested quartiles option addresses this issue by binning the background to match the overall interdependent feature densities of the foreground with quartiles to modulate bin size (Figures 2C–2E and S2). The more populated multiplexed bins should be the same in both the foreground and the background, resulting in fewer issues finding multiple matches to a foreground region within a proportionally larger background bin. This is done by first finding the independently grouped feature combinations, and then for each group, nesting for loops for every subsequent nesting criterion. Within each for loop, the quartiles of the nesting criterion are found, and the where clause of the query is updated to only include background regions that fall into the independent group combination and individual quartile ranges. The last for loop writes the where clause information into a quartile bin table within the database to be used when mapping foreground regions to background bins.

Database Querying for Background

To match a foreground set, each region's percent overlap is found and binned in the same way as the designated background table. That table's matching multiplexed bin is randomly sampled without replacement per foreground region for the user-requested multiplicity of background sequences. If all foreground regions find a match, the background set is validated, and a background fasta file is made using "fastaFromBed" of the BEDTools suite and a supplied fasta file.

Motif Enrichment of Highly Expressed TFs

We wanted to evaluate whether the enrichment of a motif not bound directly by a ChIPed TF was predictive of the high expression of TFs specific for a cell line. For that, we collected RNA-seq data from ENCODE (Table S5), which profiled the expression of most of the cell lines assayed in our ChIP-seq datasets. In each cell line, we ranked the genes according to their expression level (FPKM), where low rank indicates high expression. If there were replicate RNA-seq datasets, the average FPKM was computed and used to rank the cell line expression profile.

For the TF motifs, we focused on the 12 TF families with strong DNA binding preferences ("tethering motifs" list: POU, GATA, FOX, KLF, CEBP, MAF, IRF, ETS, TEA, AP2, ZIC and E2A) that play a relevant role in transcriptional control and cellular differentiation. To compare the different settings used in this study for motif enrichment analysis, we considered the AUROC values obtained previously (Figures 3 and 4). For each setting, we obtained the AUROC values for the 12 tethering motifs across the ChIP-seq datasets. To include only tethering interactions between the motif and the ChIPed TF, we sorted out AUROC values coming from cases of direct binding (such as GATA motif enrichment in GATA3 peaks in MCF-7 cells). For the other cases of indirect binding, we selected the ChIP-seq motif pair exhibiting a significant AUROC enrichment ($p < 0.05$). For each pair, we then asked whether at least one TF associated to the motif ("tethering TF") was highly expressed in the cell line assayed by ChIP-seq. The TF was marked as highly expressed if its expression rank was greater than a predefined threshold (*i.e.*, top 3000 genes, see below). If this was the case, we recorded the event as positive. For each of the 12 motifs along the whole ChIP-seq dataset, the count of highly expressed tethering TFs was recorded.

A permutation test was performed for 200 randomly reshuffled genomes (Barrera et al., 2016) to determine the null value of highly expressed TFs per motif. The p value was calculated as the proportion of times the reshuffled genome found more highly expressed genes than the real genome. A robust Z-score of finding highly expressed potential tethering factors was calculated as follows where x is the number of highly expressed genes in a ChIP-seq set A:

$$Z_A = \frac{x_A - median_x}{MAD}$$

Expression Rank of Lineage-Specific TFs

We considered well-established lineage-specific TFs for 7 cell types with publicly available transcriptomic profiles (RNA-seq data from ENCODE, Table S5): Sox2 (Wang et al., 2012b; Zhou et al., 2016) and POU5F1 (Wang et al., 2012b) in hESC cells, IRF4 (Mendez and Mendoza, 2016) and PAX5 (Mendez and Mendoza, 2016) in B cells, MYOD1 (Fong et al., 2015) in skeletal muscle cells, GATA1/2 (Bresnick et al., 2012) in K562, HNF4A (Nagy et al., 1994) and FOX2 (Lee et al., 2005) in HepG2, and GATA3 (Takaku et al., 2015) in Mcf-7. For each cell type, we independently ordered the genome by expression level (FPKM), and we found that the lineage-specific TFs were within the top 3,000 genes in their respective cell types, but were of lower expression ranks (>5,000) in lineages they do not specify.

Predicting Binding Type from Motif Enrichment

We identified the subset of glossary's 8mer modules that showed a significant enrichment of their associated TFs' DNA binding specificities in at least one of the their cognate ENCODE TF ChIP-seq datasets used in this study (Table S3). For example, the TEA module was selected since it was enriched in TEAD4 ChIP-seq peaks in HepG2 cells. This sorting ensured that the enrichment of these modules is *bona fide* due to the direct binding of a TF to its cognate motif. These modules are: KLF, AP2, RFX, GATA, ETS, FOX, IRF, NFkBp65, HNF4, CEBP, bZIP, MAF, TEA, POU, Hbox, E2A, EGR, GR, E2F_zfC2H2, zfC4, TCF7, bHLH, ETSI, FOXM, ETSIII, ZFP691, ESRR, ARID3, zfC2H2_EGR_RHD.

The enrichment of a module to a ChIPed TF dataset through direct binding (*i.e.*, TEA module enrichment in TEAD4 ChIP-seq for K562, HepG2 and ESC cells; green dots in Figure 6A) were further separated out and reported in Table S4; we refer to these modules as "Direct Module A". The remaining combinations, which evaluate the potential indirect binding of the ChIPed TF ("Tethered TF A") through a tethering factor ("Tethering TF B") that is among the TFs that are specific for the interacting module ("Indirect Module B"), were identified as those with AUROC > 0.6 and distance-to-summit < 40 bp as criteria (black dots in Figure 6A). More precisely, we put these limits on the indirect binding to avoid the bulk of non-enrichment unlikely to yield true results, as suggested by Figure 4A. The schematics of these potential indirect binding mechanisms with their naming schemes are shown in Figure 6B. These analyses indicated 48 indirect binding interactions (which sometimes occurred in multiple cell lines): 16 of which have been well-validated in previous experimental reports; 19 have been proposed in some previous analysis and are supported by our findings; and 13 that are to our knowledge newly identified in this study (Table S6). Table S6 also denotes the percentage of the tethered TF A's best 500 peaks that contain its own direct motif (*i.e.*, at least one 8mer of Direct Module A within the peak), its potential indirect motif (*i.e.*, at least one 8mer of Indirect Module B within the peak), and both motifs together. We performed a Fisher Exact Test with correction for false discovery rate (FDR) (fisher.test and p.adjust functions in R) to test two hypotheses: the first is to determine whether the co-occurrence of the two motifs within the same peaks is significantly enriched (fisher.test argument: alternative = "greater"), and the second is to determine whether the co-occurrence of the two motifs within the same peaks was significantly depleted (fisher.test argument: alternative = "less"). An example 2x2 contingency table is shown below for the bZIP TF BACH1 putatively being tethered by the MAF family TF MAFK in H1-hESC cells (row 2 in Table S6), indicating a significantly enriched ($p < 0.05$) co-occurrence of the bZIP and MAF motifs (schematic in Figure 6B).

Peaks with motifs from:	Direct Module A (bZIP)	No Direct Module A
Indirect Module B (MAF)	301	94
No Indirect Module B	46	59

To further investigate the potential interaction between the Tethered TFs A and their Indirect Module(s) B of Table S6, we analyzed the extent to which the set of ChIP-seq peaks of TFs A and B overlap. We considered the top 500 peaks of every Tethered TF A with a potential indirect interaction. We trimmed each peak to 200 bp, as defined above (see "Analysis of motif enrichment and Motif-Summit distance in ChIP-seq peaks"). Of all the analyzed ChIP-seq datasets (Table S3), we then restricted our analysis to those for which a TF specific to Indirect Module B was also profiled by ChIP-seq in the same cell type (candidate Tethering TF). For each candidate Tethering TF, we then trimmed its ChIP-seq peaks to 200 bp and intersected them with the Tethered TF A file ("intersectBed" of BEDTools) to determine how many of the Tethered TF A peaks overlapped the candidate Tethering TF peaks. The Tethering TF B overlapping the highest number of Tethered TF A peaks is listed in Table S7. To ensure robustness, this analysis was repeated varying the cutoff for the number of best peaks of Tethering TF B (All, 20000, 10000, 5000, 2000, 1000, 500 top peaks; Table S7). Typically, the potential tethering overlap occurs primarily within the top 5000 peaks of Tethering TF B, which in many cases overlap with all the peaks of the Tethered TF A.

For interactions with partially overlapping sets of peaks (10% < All Peaks Overlap Percentage < 90% in Table S7), we tested whether the presence of the indirect motif (*i.e.*, 8mer of Indirect Module B) within the peak was discriminative of the peak overlap between Tethered TF A and Tethering TF B. For that, we used a Fisher Exact Test with FDR correction (Table S8); for example, BACH1 peaks (Tethered TF A) potentially tethered by MAFK (Tethering TF B) in H1-hESC cells had the following contingency table:

BACH peaks with:	MAF Motif	No MAF Motif
MAFK peak Overlap	392	50
With No Overlap	3	55

and was statistically significant ($p < 0.05$). For those interactions that were not statistically significant, the indirect interaction may still be valid, but the responsible Tethering TF B that recognized Module B may be different from the one(s) for which ChIP-seq data were available for our analysis.

We repeated the Fisher Exact Test with FDR correction from Table S8 for the entirety of the candidate Tethering TFs, regardless of maximum overlap (data not shown). Those statistically significant interactions ($p < 0.05$) that have not yet been fully validated in the literature are reported in Table 1 as well as any fully overlapping interactions where a Fisher Exact Test is not appropriate due to the lack of a negative case. Representative examples of those putative indirect interactions are shown in Figure 6D.

DATA AND SOFTWARE AVAILABILITY

PBM Data

The new PBM data for 63 TFs have been deposited into the publicly available UniPROBE database of PBM data (publication dataset accession MAR17A).

Software

Both the glossary of TF-8mer modules and the GENRE suite are publicly available at http://thebrain.bwh.harvard.edu/glossary-GENRE/download.html.