Limited Time Free Offer. ($3,795 Value)

For additional levels of security during brief power outages, SANYO offers a free, built-in Liquid N2 or Liquid CO2 back-up system with the purchase of any MDF-U500VXC now through June 30, 2010.

Nothing Beats 100% Protection. Except 200% Protection.

If you’re preserving the work of a lifetime, this -86°C freezer is for you.

SANYO introduces Twin Guard Series™, the most reliable dual, redundant cascade freezer for the most critical -86°C, ultra-low applications ever. The V.I.P.™ insulated MDF-U500VXC is powered by two completely separate -86°C DualCool refrigeration systems, each built to run efficiently – alone or even better together in energy-saving EcoMode™ – to safely preserve whatever you put inside.

Learn more. Visit www.twinguardseries.com or call 800-858-8442.

Pictured: The 18.3 cu.ft. V.I.P.™ insulated MDF-U500VXC. Includes integrated LCD performance monitor and digital controller for comprehensive system management, data logging, remote communications, alarms, predictive performance and validation. Maintenance-free, filterless design.
Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites

Savina A. Jaeger a,1, Esther T. Chan d, Michael F. Berger a,e, Rolf Stottmann a, Timothy R. Hughes c,d, Martha L. Bulyk a,b,c,e,f,

Abbreviations: AUC, area under receiver operating characteristic curve; GO, Gene Ontology; MGI, Mouse Genome Informatics; PBM, protein binding microarray; PWM, position weight matrix; TF, transcription factor; TSS, transcription start site; UCE, ultraconserved element; UTR, untranslated region.

⁎ Corresponding author. Division of Genetics, Department of Medicine; Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA.

1 Current address: Computational Sciences Center of Emphasis, Pfizer Global Research and Development, Cambridge, MA 02139, USA.

© 2010 Elsevier Inc. All rights reserved.

Introduction

The interactions between transcription factors (TFs) and their DNA binding sites are an integral part of the regulatory networks within cells. Recent computational studies in Saccharomyces cerevisiae [1] and in Drosophila embryonic development have suggested that low affinity TF binding sites are important in gene regulation [2]. However, those studies utilized position weight matrix (PWM) models of TF binding sites, rather than directly measured preferences of TFs for DNA binding sequence “words” (“k-mers”); PWMs are non-ideal as they typically are based on only a relatively small set of binding site sequences, either inferred or directly measured, and in general do not capture well the full range of binding preferences of TFs [3–6].

The DNA binding specificities of 104 known and predicted mouse TFs from 22 different DNA binding domain (DBD) structural classes found in metazoan TFs were determined recently [6] using the universal [7] protein binding microarray (PBM) technology [8,9]. Those PBM data are the first high-resolution binding specificity data for the vast majority of those 104 TFs; universal PBMs provide a complete look-up table of the relative binding preference of a TF for each gapped and ungapped 8-bp sequence variant. For each of the 8-mers in the dataset for each protein, its PBM enrichment score ($E$-score) [7] was reported [6]. The $E$-score is related to the area under a receiver operating characteristic curve (AUC) and scales from $+0.5$ (best) to $−0.5$ (worst) [7]. Comparisons to $K_{a}$ data [7,10] indicated that 8-mers with higher normalized PBM signal intensities, and thus with higher $E$-scores, are generally bound with higher affinity [6,7]. From the individual 8-mer data, DNA
binding site motifs were derived [6] using the Seed-and-Wobble algorithm [7,11].

In addition, in that study [6] many TFs were found to bind two distinct sets of high-scoring 8-mers. For each such TF, a secondary motif was identified that captured DNA binding preferences not represented well by the primary motif. A linear regression approach was used to learn weighted combinations of PWMs generated from several different motif finding algorithms; the binding profiles for the vast majority of the proteins were represented best by more than one PWM. PWMs are of practical utility because they are employed by numerous genome scanning tools to identify candidate regulatory elements [12].

While those PWM data provided rich, high-resolution profiles of the in vitro DNA binding specificities of 104 mouse TFs, the in vivo functional relevance of those binding site sequences was unexplored for all but 2 of the TFs [6]. Here, we analyzed those DNA binding specificity data to determine the potential utility of in vitro binding data in computational genome analyses aimed at identifying candidate in vivo regulatory associations [13]. In particular, we examined whether TF-binding 8-mers are evolutionarily conserved and whether they are enriched within regions of the genome that are likely to have cis-regulatory function. We report that:

- **PWM “bound” 8-mers are enriched within various classes of putative regulatory regions, including CpG islands and non-exonic ultra-conserved elements (UCEs)** [14].
- **Non-exonic UCEs exhibit the greatest fold-enrichment of high affinity 8-mers for TFs of the homeodomain and BRIGHT classes. Lower affinity homeodomain and BRIGHT class 8-mers are also enriched within non-exonic UCEs.**
- **Of TFs whose PWM “bound” 8-mers are enriched within sets of tissue-specific UCEs, many are expressed in the same tissue(s) as the UCEs.**
- **Both high and lower affinity 8-mers appear to be under negative selective pressure across mammalian genomes as compared to putatively non-binding 8-mers.**
- **Overall, secondary motif 8-mers are conserved as strongly as primary motif 8-mers in both low and high CpG regions.**
- **Putative target genes of most TFs are enriched for particular Gene Ontology functional categories or for tissue-specific gene expression clusters.**

Taken together, our results provide evidence supporting that lower affinity TF binding sites, as determined from PWMs, serve evolutionarily conserved, in vivo regulatory functions.

**Results**

**PWM “bound” 8-mers are enriched within putative regulatory regions, including CpG islands and non-exonic UCEs**

We first examined whether 8-mers bound strongly in vitro in PWMs by any of the 104 previously characterized mouse TFs [6] are enriched in putative regulatory regions of the mouse genome, including CpG islands, UTRs, non-exonic UCEs, and intergenic or intronic regions, considering ancestral repeats as a negative control. Considering 8-mers bound (E ≥ 0.40, 0.43, or 0.45) by any of the 104 TFs, the absolute overall enrichment or depletion of TF-binding 8-mer occurrences in these regions was quite minor, with enrichment of these 8-mers in noncoding regulatory regions in general (P < 10⁻⁸, Wilcoxon–Mann–Whitney test) and depletion in UTRs (P < 10⁻⁸) and ancestral repeats (P < 10⁻¹²) (Supplementary Figure S1). We note that these P-values may be inexact because the overlapping 8-mers are not necessarily independent.

CpG islands were particularly enriched for binding sites for E2F (1.12-fold at E ≥ 0.40) and ETS (1.26-fold at E ≥ 0.40) proteins (Fig. 1A, Supplementary Figure S2); see below). Chromatin immunoprecipitation analysis has shown that in both normal and tumor cell lines, most in vivo binding sites for E2F1, E2F4, and E2F6 are located within 2 kb of a transcription start sites [15], where CpG dinucleotides are enriched [16], and CpG methylation has been shown to differentially regulate the response of distinct E2F elements at various promoters to different E2F family members [17].

Within non-exonic UCEs, we found the greatest fold-enrichment of “high” affinity 8-mers (E ≥ 0.45) for TFs of the homeodomain (P < 10⁻²²) and BRIGHT (P < 10⁻¹²) classes (Fig. 1B); this enrichment is observed even when we considered only the “moderate” affinity 8-mers (0.40 ≤ E < 0.45) (Fig. 1B; Supplementary Table S1). Our finding for the homeodomain class supports prior observations of enrichment of matches to a generalized homeodomain motif within UCE-spanning sequences [18,19].

Previous large-scale testing of these UCEs [20,21] and human-pufferfish conserved noncoding elements [20] in a transgenic mouse enhancer assay showed that many of them drive tissue-specific gene expression, primarily in the embryonic nervous system [20,21]. One of these studies also found that noncoding UCEs and extremely constrained human–rodent elements are highly enriched for neighboring genes involved in regulation of transcription, development, and nervous system development [21]. Taken together, these results suggest that UCEs may be fine-tuned to respond to the transcriptional regulatory state of cells, in particular to the levels of homeodomain and BRIGHT class TFs utilized in development and differentiation.

**Evidence for functional roles of PWM “bound” 8-mers enriched within highly conserved, putative neuronal regulatory regions**

To begin to address the hypothesis that PWM “bound” 8-mers in UCEs have an in vivo regulatory relevance, we examined whether the corresponding TFs are expressed in the same tissue(s) where expression studies of UCE-driven transgenes suggest they may be important. For example, several studies have shown that the Hox genes control motor neuron identity [22]. Consistent with anatomic locations of motor neurons, we observed enrichment (p ≤ 1.2 × 10⁻⁵) Fisher’s Exact test) of Hoxa3 PWM “bound” 8-mers (E ≥ 0.45) within UCEs and highly constrained genomic regions [20] driving expression in the cranial nerve, hindbrain, midbrain, heart, limbs, neural tube, trigeminal nerve and dorsal root ganglion, as compared to 10 independent sets of sequences generated by a 1st-order Markov model (i.e., 10 sets of sequences corresponding to the same UCEs shuffled at the dinucleotide level).

To search systematically for correlations between TF and UCE-driven gene expression, for PWM “bound” 8-mers (E ≥ 0.45) enriched within UCEs and highly constrained genomic elements [20], we examined the expression patterns of the 104 TFs, as determined by recent in situ hybridization studies in the mouse embryonic brain [23,24], within the same tissues as those in which the above constrained genomic elements drove reporter gene expression at E11.5 [20,21]. We manually annotated the expression of TFs in the developing mouse brain in the Allen Institute’s publicly available developing mouse brain in situ hybridization images [24] and the MGI database (http://www.informatics.jax.org/), where such published annotations were not available. The correlations we found between TF and UCE-driven gene expression, for PWM “bound” 8-mers (E ≥ 0.45) enriched within UCEs and highly constrained genomic elements, are described below. A full listing of PWM 8-mer enrichment results for all 104 TFs and all examined UCE expression patterns is provided in Supplementary Table S2.

Gata3 is expressed in the lens at E10.5 [23], and its “bound” 8-mers are significantly enriched (p ≤ 0.05, Fisher’s Exact test) in enhancers driving expression in eyes at E11.5 [21] (Fig. 2). Nr2f2 and Six6 are expressed in the optic vesicle at E10.5 [23], and their “bound” 8-mers are likewise enriched in enhancers driving expression in eyes at E11.5 [21]. Nr2f2 is also expressed in all neural tissues at E11.5, including the
neural tube, dorsal root ganglion, and optic nerve [25], and its “bound” 8-mers are likewise enriched in enhancers driving expression in the neural tube, dorsal root ganglion, forebrain, midbrain, hindbrain, and eye at E11.5. Numerous TFs—E2f2, Foxj1, Foxa2, Foxk1, Gata3, Klf7, Mafb, Mtf1, Nr2f2, Rfx3, Rfx4, Sox4, Sox5, Sox14, Sox21, Tcf3, Tcf7, Tcf7l2, and Zic3—are expressed in the midbrain at E13.5 [23], and their “bound” 8-mers are enriched in enhancers driving expression in the midbrain at E11.5 (Fig. 2). A number of these and other TFs—Foxa2, Foxj1, Gata3, Hoxa3, Klf7, Mafb, Mtf1, Myb11, Rara, Rfx3, Rfx4, Sox4, Sox5, Sox7, Sox21, Tcf3, Tcf7l2, and Zic3—are expressed in the hindbrain at E13.5 [23], and their “bound” 8-mers are enriched in enhancers driving expression in the hindbrain at E11.5 [21] (Fig. 2). The Allen Institute’s in situ hybridization data support these findings, with expression of Tcf7l2 [26,27] and Foxa2 in the midbrain at E11.5, and of Foxa2, Hoxa3, and Tcf7l2 in the hindbrain at E11.5 [24]; Allen Institute in situ hybridization images were not available at E11.5 for the other TFs. Prior studies support these correlations for enhancer activity at E11.5 for a number of the other TFs, including Gata3 [28,29], Sox1 [30], and Zic3 [31] in the midbrain and hindbrain, and for Klf7, which is ubiquitous at E10.5 [32], in the hindbrain.

While the enhancer expression data [20,21] and the TF expression data from Gray et al. [23] were obtained at different embryonic time points (E11.5, and E10.5 or E13.5, respectively), these data suggest that these TFs may regulate gene expression through these UCEs in these embryonic tissues. There is literature evidence supporting at least some of these hypotheses, as follows. Foxa2 homozygous null zebrafish exhibit severe reduction of prospective oligodendrocytes in the midbrain and hindbrain [33], and Foxa2 regulates midbrain dopaminergic neuron development in mouse embryos [34]. Mafb is involved in both segmentation and specification of anteroposterior identity in the hindbrain [35]. AP-2 (Tcfap2a) knockout mice exhibit failure in cranial neural tube closure and defects in cranial ganglia

---

**Fig. 1.** Enrichment of particular TFs’ 8-mers within putative regulatory regions. (A) CpG islands are enriched for PBM “bound” 8-mers for E2F and ETS proteins. Results for 8-mers bound at $E \geq 0.43$ or $E \geq 0.45$ are shown in Supplementary Figure S2. (B) “Moderate” ($0.40 \leq E < 0.45$) and “high” ($E \geq 0.45$) affinity 8-mers of TFs in the BRIGHT and homeodomain classes are enriched within non-exonic UCEs as compared to shuffled sequences generated to have the same dinucleotide content. In both panels, “score” and $P$-values are as described in Supplementary Fig. S1.

---
Fig. 2. TFs expressed in same tissues as the UCEs in which their PBM “bound” 8-mers are enriched. TF binding site sequence logos are presented for graphical convenience; binding sequence enrichment analysis was based on PBM 8-mer data. Fold-enrichment of PBM “bound” 8-mers ($E \geq 0.45$) within UCEs is calculated as compared to 10 sets of UCE sequences shuffled at the dinucleotide level. Because of space limitations, only a subset of the expression patterns and TFs are shown. A full listing of PBM 8-mer enrichment results for all TFs and all examined UCE expression patterns is provided in Supplementary Table S2. (UCE reporter assay whole-mount in situ images at mouse E11.5 adapted by permission from Macmillan Publishers Ltd: [Nature] [20], copyright (2006).)
development [36]. Thus, PBM data suggest putative links between transcriptional enhancers and specific TFs, including some with known regulatory functions as described above, that may mediate gene expression by binding within those enhancers. Analysis of over-representation of combinations of TFs' binding site sequences in the future may suggest combinatorial, co-regulatory TF interactions and associated cis regulatory grammars within UCEs. Further experiments will be required to confirm direct binding of the TFs to their putative target regulatory regions.

**Evolutionary conservation properties of PBM "bound" 8-mers**

8-mers that score highly in a PBM experiment represent the preferred in vitro DNA binding sites for that TF. We reasoned that if genomic occurrences of those 8-mers are functional in vivo, and if the TFs in our dataset capture a substantial fraction of the diversity in TF binding sites, then selective pressure would result in those 8-mers being evolutionarily conserved as compared to other 8-mers. We therefore examined the conservation properties of 8-mers within the non-protein-coding regions 10 kb upstream to 10 kb downstream of the transcription start sites of RefSeq genes considering sequence alignments of 12 mammalian genomes. We further subdivided the regions into those with high versus low CpG content in their sequence, after inspecting the bimodal distribution generated by the scoring function over all of the RefSeq promoters, as defined by Mikkelsen et al. [37]. We utilized SCONE scores for the substitution rates of individual nucleotides in the genome [38] to calculate substitution rates for all 8-mers within the examined genomic regions separated into low CpG and high CpG regions (Supplementary Figure S3). An 8-mer with a low average substitution rate can be considered to be more conserved within these regions than an 8-mer with a higher average substitution rate.

Considering as a whole the set of 8-mers with a maximum E-score above 0.45 for any of the 104 TFs, represented by 111 protein constructs, to be "high" affinity binding sites for at least one TF (i.e., 8-mers that have low E-scores for some TFs but that have $E \geq 0.45$ for at least one TF were concerned as "high" affinity binding sites), we found that these 8-mers on average exhibited lower substitution rates than the average of all other 8-mers ($P \leq 10^{-10}$ for low CpG regions; $P \leq 10^{-71}$ for high CpG regions, two-sample t-test for samples with unequal variances and Satterthwaite's approximation for the effective degrees of freedom).

Moreover, we found that the average substitution rates for 8-mers within the high CpG regions are uniformly lower than the ones within the low CpG regions (Supplementary Figure S3), which suggests that despite the higher mutation rate of CpG dinucleotides [39], on average there has been stronger selective pressure to maintain TF binding sites within high CpG regions. This is consistent with recent observations that CpG islands display high levels of sequence conservation as compared to various other putative regulatory regions [38].

In addition, we observed a trend between 8-mer conservation rates and the relative binding affinities of TFs to those 8-mers. When we binned the 8-mers into "high" (maximum $E \geq 0.45$ over all 104 TFs; 4,787 8-mers fell into this bin), "moderate" ($0.40 \leq E < 0.45$), low ($E < 0.40$), and nonspecific ($E < 0.30$) categories within (A) high CpG regions and (B) low CpG regions, two-sample t-test for samples with unequal variances and Satterthwaite's approximation for the effective degrees of freedom led to the most extreme data points not considered outliers.

**Fig. 3.** The median 8-mer substitution rate decreases monotonically with increasing binding site affinity. All bins of 8-mers—"high" affinity ($0.45 \leq E \leq 0.50$), "moderate" affinity ($0.40 \leq E < 0.45$), "low" affinity ($0.35 \leq E < 0.40$), "very low" affinity ($0.30 \leq E < 0.35$), and nonspecific ($E < 0.30$) categories—within (A) high CpG regions and (B) low CpG regions exhibited mean substitution rates significantly different from each other ($P \leq 0.05$, Tukey's Honestly Significant Differences test). In each box plot, the central bar indicates the median, the edges of the box indicate the 25th and 75th percentiles, and whiskers extend to the most extreme data points not considered outliers.
secondary motif (seed 8-mer had $E \geq 0.45$) distinctly different from the primary motif [6] [Supplementary Table S3], here we examined the evolutionary conservation of the top 100 8-mers belonging to the primary versus the secondary Seed-and-Wobble motifs. 8-mers that matched both the primary and the secondary motifs were removed from consideration.

Although the primary motif by definition corresponds to the motif generated from the highest ranking 8-mer in the PBM data, the set of 8-mers represented by the secondary motif can rank quite similarly to those belonging to the primary motif or can be of distinctly lower affinity [6]. To control for potential affinity bias in the set of 8-mers corresponding to the primary versus secondary motifs, we categorized all primary and secondary motif 8-mers into “high,” “moderate,” “low,” and “very low” categories, as described above. Considering 8-mer data for all 43 of these proteins, we found that the secondary motif 8-mers are just as evolutionarily conserved as the primary motif 8-mers in all 4 of these affinity categories (Table 1). Thus, secondary motif 8-mers may be just as biologically important as primary motif 8-mers. In the future, site-directed mutagenesis studies of primary versus secondary motif 8-mers will need to be performed in vivo to dissect the significance and gene regulatory roles of primary versus secondary motif 8-mers.

To investigate further the conservation properties of “high” affinity 8-mers, we considered each TF’s “high” affinity 8-mers as compared to those bound most weakly ($E \leq 0$) for their conservation within low versus high CpG regions. Although we observed an overall trend that higher affinity 8-mers were more highly conserved within both low and high CpG regions (Fig. 3), on an individual TF basis we found that not all “high” affinity 8-mers ($E \geq 0.45$) were

### Table 1

<table>
<thead>
<tr>
<th>Affinity category</th>
<th>Low CpG regions (AUC)</th>
<th>High CpG regions (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High ($E \geq 0.45$)</td>
<td>0.512</td>
<td>0.511</td>
</tr>
<tr>
<td>Moderate ($0.40 \leq E &lt; 0.45$)</td>
<td>0.467</td>
<td>0.496</td>
</tr>
<tr>
<td>Low ($0.35 \leq E &lt; 0.40$)</td>
<td>0.555</td>
<td>0.596</td>
</tr>
<tr>
<td>Very low ($0.30 \leq E &lt; 0.35$)</td>
<td>0.537</td>
<td>0.536</td>
</tr>
<tr>
<td>Nonspecific ($E &lt; 0.30$)</td>
<td>0.521</td>
<td>0.483</td>
</tr>
</tbody>
</table>
conserved more highly than their weakest associated 8-mers \((E \leq 0)\). Within the low CpG regions, for only 51 of the 104 TFs did the “high” affinity 8-mers exhibit greater conservation overall \((\text{AUC} \leq 0.5\) and \(P < 0.05\), Wilcoxon–Mann–Whitney test) than their weakest 8-mers (see E2F2 example in Fig. 4A). For 40 of the 104 TFs there was no significant difference between the “high” affinity and the weakest 8-mers, and for the remaining 13 TFs the “high” affinity 8-mers were actually less conserved than their weakest 8-mers (see Zic2 example in Fig. 4C). This trend was similar within the high CpG regions (see E2F2 example in Fig. 4B, and Zic2 example in Fig. 4D). Only 34 TFs, including E2F2, exhibited greater conservation of high affinity 8-mers, as compared to their weakest 8-mers, within both the low and high CpG regions.

Since the presence of 8-mers bound by other TFs among weak 8-mers for the TF in question has the potential to account for a portion of the unexpected high conservation of weak 8-mers, for each TF we next removed from consideration any of its weak 8-mers that had \(E \geq 0.45\) for any of the other 103 TFs. We then found that within the high CpG regions, for 68 TFs high affinity 8-mers were more conserved than the non-binding 8-mers on average; within the low CpG regions, for 62 TFs high affinity 8-mers were more conserved than the non-binding 8-mers on average. Within the high CpG regions, for 19 TFs the non-binding 8-mers exhibited significantly higher conservation rates on average than the high affinity binding 8-mers; within the low CpG regions, these were 10 TFs. For 25 TFs there was no significant difference \((P > 0.05)\) of conservation rates between high affinity and non-binding 8-mers determined within the high CpG regions; there were 40 such TFs within the low CpG regions. 44 TFs exhibited greater conservation of high affinity 8-mers, as compared to their weakest 8-mers, within both the low and high CpG regions. Despite the diversity of TFs present in this collection of 104 TFs, it remains to be seen whether the rest of this signal is due to other, as yet uncharacterized TFs. Thus, analysis of TF binding site conservation as compared to non-binding genomic sequence may underweight the significance of conservation of some TFs’ binding sites.

Association of PBM-derived motifs with specific functional categories of genes

We next sought to further annotate these TFs and to predict their functional categories of target genes, albeit coarsely, by searching for enrichment of phylogenetically conserved motif occurrences in the noncoding sequence surrounding numerous sets of candidate target genes. We used the PhyICRM and Lever algorithms, which predict cis regulatory modules, and infer cis regulatory codes, respectively (i.e., Lever searches for TF binding site motifs enriched within PhyICRM-predicted cis regulatory modules for various input gene sets) \([40]\). In analysis of the PBM data for the 104 nonredundant TFs we analyzed here, it was found that multiple PWMs frequently capture the binding preferences of a TF better than does a single PWM \([6]\).

Therefore, for Lever analysis this study, we used the PWMs previously selected as best representing the DNA binding profiles of each of these proteins \([6]\).

We considered a broad range of mouse tissue-specific gene expression clusters \([41]\) and Gene Ontology (GO) annotation terms \([42]\) in assembling gene sets, which we sorted into low versus high CpG content gene sets, yielding a total of 1371 gene sets. We considered 20 kb of noncoding sequence surrounding the transcription start site of each gene and 12 mammalian genomes in scoring motif occurrences. Out of the 152,181 pairings of these gene sets with each particular protein, we observed a total of 283 significant pairings, involving a total of 45 gene sets and 88 TFs, at stringent significance thresholds \((\text{AUC} \geq 0.8\) and \(P < 0.01\)) (Fig. 5, Supplementary Table S4). As expected \([40]\), motifs for the myogenic TFs Myf6 (an HLH factor) and Srf are enriched in various muscle-related GO categories such as “sarcomere” and “muscle cell differentiation” (Fig. 5). The Irf proteins are known regulators of interferon response and other aspects of immune response and hematopoiesis \([43]\); our Lever results revealed Irf3 to be associated with the GO category “cytokine biosynthetic process” (Fig. 5), and Irf6 associated with a gene expression cluster appearing in tissues that contain immune cells (e.g. spleen, lymph

---

**Fig. 5.** Lever screen of GO categories. Heatmap color gradient indicates Lever AUC values. The columns have been restricted such that only those GO categories that exhibit significant enrichment \((\text{AUC} \geq 0.8, P \leq 0.01)\) for at least one of the TFs’ binding sites are shown. The columns (TFs) were sorted according to TF structural class, while the rows (GO categories) were clustered hierarchically according to the AUC values calculated by Lever \([40]\).
node, bone marrow, lung, etc.) (Supplementary Figure S4). Other associations represent potentially new findings. For example, the zinc finger protein Sp4 is associated with a gene expression cluster expressed primarily in brain, and reports in the literature [44] indicate that mutants have defects in memory and that the gene itself is expressed primarily in brain. TcF3, which is best known for its involvement in Wnt signaling [45], is associated with the GO category “segmentation,” consistent with a recently described role in restricting induction of the anterior-posterior axis [46]. Zfp161 (ZFS), a repressor of the human fragile X mental retardation 1 (FMR1) gene [47], is associated with gene expression clusters primarily in the embryo (Supplementary Figure S4).

Other independent lines of evidence also support the biological relevance of our Lever results. First, we observed a tendency for the TF itself to be co-expressed with the gene expression clusters for which its binding sequences are enriched (P<0.001 at AUC≥0.6, Q≤0.05, for both low and high CpG gene sets; Supplementary Figure S5), likely because transcriptional activation of the TF gene itself is required for transcriptional activation of its targets. There is also a tendency for the GO categories that Lever associates with each TF to be among the categories in which the TF itself is annotated (P<0.04 and P<0.005 at AUC≥0.7, Q≤0.05, for low and high CpG gene sets, respectively). Some of the Lever results, however, may correspond to correct associations of motifs to target gene sets, but for related TFs not present in our set of 104 examined mouse TFs but with highly similar DNA binding site motifs; for example, the association of the Myf6 motif with various GO categories pertaining to neuron development may be due to neuronally expressed bHLH factors such as NeuroD (Neurod1), Mash1 (Ascl1), Neurog1 or Neurog2 [48].

Discussion

Our analysis of the substitution rates of 8-mers across mammalian genomes suggests that not only high affinity binding sites, but also numerous moderate and low affinity binding sites, are more evolutionarily conserved than nonspecific binding sites. Despite this overall trend, the 8-mers bound most strongly by 27 of the 104 TFs we examined did not exhibit significantly greater conservation in either the low or high CpG regions. Many genomic occurrences of the binding site sequences for such TFs may not be functional, or if functional, may have exhibited significant binding site turnover. Indeed, a recent study of the occupancies of four tissue-specific TFs in human and mouse promoters found that as few as 41% of the binding sites are conserved [49]. Algorithms capable of scoring the conservation of “k-mers” binding site affinities despite sequence differences need to be developed. In addition, the sequences not bound by one TF potentially could be genuine binding sites for another TF.

Our analysis of UCEs revealed that groups of UCEs that drive similar gene expression patterns in mouse embryos at E11.5 are enriched for particular TFs’ PBM “bound” 8-mers, and that those TFs are expressed in the same tissue(s) as those in which the UCEs drive gene expression. These results can be utilized as predictions of which TFs may regulate gene expression through particular sites within those UCEs. The enrichment of both high and lower affinity binding sites in UCEs suggests that UCEs may have differential regulatory outputs specific to the nuclear levels of those TFs.

Prior studies typically have focused on high Affinity TF binding sites or higher-scoring motif matches. The advent of high-resolution PBM data for these 104 mouse [6] and many other TFs [13,50,51] opens a wide avenue for investigation of the regulatory importance of lower affinity sites. In addition, as more TFs’ DNA binding specificities are discovered, much of the genome may appear to be a reservoir of potential regulatory sequences from which binding sites for different TFs may arise or erode.

In our Lever analysis of 1,371 gene sets, we observed a total of 285 significant pairings, involving a total of 45 gene sets and 88 TFs, at stringent significance thresholds (AUC ≥ 0.8 and Q ≤ 0.01). Numerous associations of TFs with GO categories are supported by the literature or by the TF being expressed in the same tissue(s) as the gene set(s) with which its binding sites are enriched, while many others are novel, putative regulatory associations. Evaluation of motif combinations in the future is likely to identify more specific associations of TFs to gene sets [40]. In this analysis we examined 20 kb of noncoding sequence surrounding the transcription start site of each gene for motif occurrences. Our choice of 20 kb was both practical in terms of computational time requirements and also likely somewhat conservative for many TFs. TF binding events in vivo have been observed to be enriched within ∼5 kb surrounding transcription start sites, as evidenced by studies employing chromatin immunoprecipitation coupled with microarray readout (ChIP-chip) for a handful of TFs in human cell lines [52,53]. However, given the limited number of TFs analyzed in those ChIP-chip studies [52,53] and the abundance of distally located TF binding events in ChIP-chip [52] and numerous validated, distantly located transcriptional enhancers [54], it may be informative for the prediction of cis regulatory modules and codes to expand these Lever analyses in future studies to consider genomic sequences located further upstream or downstream of the genes [40].

To our knowledge, the Lever analysis we presented here is the first study to employ a multiple motifs model for individual TFs in searching genomic sequence for TF binding site motif matches. This is important, since it was recently shown that multiple motifs often capture the DNA binding preferences of proteins more accurately than does any single motif [6]. Algorithms that can predict cis regulatory modules while scoring the evolutionary conservation of k-mers need to be developed, in particular since TF-specific differences in DNA binding, which may not be captured well by PWMs, are being revealed [50,51]. Our Lever analysis was limited by potential divergence in TF expression or DNA binding specificity, and potential divergence in binding site composition and locations of orthologous cis regulatory modules [55]. Such k-mer analyses will be important for understanding the “grammar” of how cis regulatory modules drive particular gene expression patterns, as well as for understanding the evolution of transcriptional regulatory networks. Future studies addressing these issues and considering motif combinations and additional gene expression datasets focused on particular cell types of interest [56], should further specify the context-dependent regulatory roles of these motifs.

Methods

PBM data

We obtained PBM-derived 8-mer data and position weight matrices (PWMs) for 104 nonredundant TFs from a recent study presenting those data [6].

Data collection and generation of control data sets

Human-based ultraconserved conserved sequences [14] and Phastcons syntenic conserved elements [57] were downloaded from the authors’ supplementary websites (http://www.soe.ucsc.edu/%7Enjill/ultra.html and http://compgen.bscb.cornell.edu/~acs/conservation/ , respectively). Human CpG island sequences were downloaded from the CpG island track in the hg17 genome build using the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTable?command=start) [58]. Control sequences with similar dinucleotide frequencies were generated using 1st order Markov models for every sequence within the data sets, using the program GenKGenS [59]. Two sets of control sequences were generated by: (1) selecting a length-matched set of random sequences across the mm8 build of the mouse genome, and (2) using a 1st and 2nd order Markov model to control for di- and tri-nucleotide frequencies in overlapping 8-mer windows with a step size of 1 bp.
Scanning of genomic regions and promoters

Each sequence in each set of genomic regions and control sequences was scanned in overlapping 8-mer windows with a step size of 1 bp using custom Perl scripts. The fraction of windows with matches to 8-mers with E-scores equal or greater than a threshold value for any TF, or for any TF of a particular structural class, for each sequence was plotted using MATLAB. Enrichment and statistical significance of scores in real versus control sequence sets were determined using the Wilcoxon–Mann–Whitney test implemented in MATLAB.

Each genomic sequence and the matching control sequences were scanned in overlapping 8-mer windows with a step size of 1 bp for hits against 8-mers with \( E \geq 0.40 \) for each TF using custom Perl scripts. The distribution of the positive 8-mer hits relative to the TSS was plotted as a normalized fraction for each sequence set and each TF using MATLAB.

Evolutionary conservation properties of PBM 8-mer data

We examined the conservation properties of 8-mers of various PBM E-scores by using the SCONE algorithm [38], which calculates a substitution rate for each nucleotide in the whole genome. Using SCONE, the MultiZ 17-way sequence alignments of the vertebrate genomes performed by Genome Bioinformatics Group of UC Santa Cruz and the phylogenetic tree provided by the ENCODE multiple sequence alignment working group [60] that we trimmed down to 12 mammals, we computed the nucleotide substitution rate at every position within the non-protein coding regions from 10 kb upstream to 10 kb downstream of annotated transcription start sites for RefSeq genes with annotated 5’ UTRs. Then, for each sliding window of size 8 nt within the regions for each RefSeq gene, we summed the individual substitution rates for consecutive positions within the 8-nt window. For every possible 8-mer we calculated the average substitution rate over all genomic occurrences within these regions. An 8-mer with a low average substitution rate can be considered to be more conserved genome-wide than an 8-mer with a higher average substitution rate.

Lever analysis

The Lever algorithm was developed previously and is described in a separate paper [40]. The February 2006 mouse (Mus musculus) genome data were obtained from the Build 36 "essentially complete" assembly by NCBI and the Mouse Genome Sequencing Consortium. We incorporated the phylogenetic information from 12 mammals: mouse, rat, human, rabbit, chimpanzee, cow, dog, armadillo, tenrec, opossum and elephant. We utilized the MultiZ 17-way alignment of: mouse (Feb 2006, mm8); rat (Nov 2004, r4); rabbit (May 2005, oryCun1); human (Mar. 2006, hg18); chimpanzee (Mar. 2006, panTro2); macaque (Jan. 2006, rheMac2); dog (May 2005, canFam2); cow (Mar 2005, bosTau2); armadillo (May 2005, dasNoV2); elephant (May 2005, loxAfr1); tenrec (Jul 2005, echTol1); opossum (Jan 2006, monDom4); chicken (Feb 2004, galGal2); frog (Aug 2005, xenTro2); zebrafish (Mar 2006, danRer4); Tetraodon (Feb 2004, tetNig1); Fugu (Aug 2002, fr1). The MultiZ alignment was performed by Genome Bioinformatics Group of UC Santa Cruz. The mouse (mm8) sequence and annotation data were downloaded from the Genome Browser FTP server (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm8/). Repeats from RepeatMasker and Tandem Repeats Finder (with period of 12 or less) were masked out (RepeatMasker January 2012 2005 version with RepBase libraries: RepBase Update 9.11, RM database version 20050112). The exon coding regions were also masked out by utilizing the annotation database. For each RefSeq mouse gene and the corresponding 11 alignments, we considered in our Lever analysis the sequence regions from 10 kb upstream through to 10 kb downstream of the annotated transcription start sites for RefSeq genes with annotated 5’ UTRs. Gene sets corresponding to gene expression clusters and GO annotation terms were generated as described below.

Gene expression cluster generation

Probes from Zhang et al. [41] and Su et al. [61] were mapped to MGI gene name identifiers. Expression profiles for probes mapped to the same gene identifier were averaged. Those genes not in the intersection gene set between both studies were excluded from further analysis. The final data contained expression profiles for 12,065 genes. Pair-wise Pearson correlation coefficients between the expression profiles of all genes were obtained for each dataset. Affinity Propagation [62] was used to cluster the data using the sum of the Pearson correlation coefficients between two genes as their similarity metric. This resulted in a total of 392 expression clusters.

GO annotation term gene sets

GO annotations provided by Mouse Genome Informatics (MGI) were downloaded from the Gene Ontology website on Oct. 19, 2007, version 1.686. GO annotations were up-propagated. We used the batch system provided in MGI for gene naming conversions. Since overlaps in the sequence flanking adjacent genes could create artifactual results, we eliminated any overlaps so that our analysis considers only non-overlapping sequences. We sorted the genes into those with high versus low CpG content in their flanking sequence, after inspecting the bimodal distribution generated by the scoring function over all of the RefSeq promoters, as defined by Mikelsen et al. [37]. We then eliminated any gene sets that had fewer than 15 or more than 300 genes. Thus, the following gene sets were used in Lever analysis: (a) 25 gene expression clusters filtered to consist solely of low CpG content genes and containing at least 15 and no more than 300 genes; (b) 166 gene expression clusters filtered to consist solely of high CpG content genes and containing at least 15 and no more than 300 genes; (c) 352 Gene Ontology (GO) Biological Process (BP) annotation term gene categories filtered to consist solely of low CpG content genes and containing at least 15 and no more than 300 genes; (d) 628 GO BP annotation term gene categories filtered to consist solely of high CpG content genes and containing at least 15 and no more than 300 genes; (e) 67 GO Cellular Component (CC) annotation term gene categories filtered to consist solely of low CpG content genes and containing at least 15 and no more than 300 genes; (f) 133 GO CC annotation term gene categories filtered to consist solely of high CpG content genes and containing at least 15 and no more than 300 genes.

Lever [40] requires an input set of PWMs to use in the genome sequence searches. The PWMs and corresponding motif match thresholds utilized in Lever analysis are provided in Supplementary Data.

Acknowledgments

This project was supported by funding from the Canadian Institutes of Health Research (MOP-77721), Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, the Canadian Institute for Advanced Research to T.R.H., and by grant R01 HG003985 from NIH/NHGRI to M.B. We thank Saurabh Asthana, Ivan Adzhubey, and Shamil Sunyaev for helpful discussions of k-mer conservation analysis and SCONE scoring. David Beier for helpful discussion about in situ TF expression analysis, Lourdes Peña-Castillo for assistance with clustering and GO analysis, and Rachel McCord and Steve Gisselbrecht for critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.01.002.

References


