Systematic identification of mammalian regulatory motifs' target genes and functions

Jason B Warner[1,7], Anthony A Philippakis[1,3,4,7], Savina A Jaeger[1,7], Fangxue Sherry He[1,6], Jolinta Lin[1,5] & Martha L Bulyk[1–4]

We developed an algorithm, Lever, that systematically maps metazoan DNA regulatory motifs or motif combinations to sets of genes. Lever assesses whether the motifs are enriched in *cis*-regulatory modules (CRMs), predicted by our PhylCRM algorithm, in the noncoding sequences surrounding the genes. Lever analysis allows unbiased inference of functional annotations to regulatory motifs and candidate CRMs. We used human myogenic differentiation as a model system to statistically assess greater than 25,000 pairings of gene sets and motifs or motif combinations. We assigned functional annotations to candidate regulatory motifs predicted previously and identified gene sets that are likely to be co-regulated via shared regulatory motifs. Lever allows moving beyond the identification of putative regulatory motifs in mammalian genomes, toward understanding their biological roles. This approach is general and can be applied readily to any cell type, gene expression pattern or organism of interest.

Of fundamental importance for understanding transcriptional regulatory networks is the functional annotation of DNA regulatory motifs (typically ~6–15 bp in length) in terms of what groups of target genes they regulate in a tissue- or temporal-specific manner in response to environmental perturbations. Although effective computational methods for mapping DNA regulatory motifs exist in the yeast *Saccharomyces cerevisiae*, where the DNA binding sites of regulatory transcription factors typically occur within ~600 bp upstream of genes, these methods cannot be applied to metazoan genomes, where genes in the same expression cluster are not necessarily co-regulated by a common mechanism, and the regulatory elements can be far from the transcription start site[1].

In metazoans, regulatory motifs tend to co-occur in stretches of noncoding sequence, CRMs, that regulate expression of nearby gene(s). Many approaches have resulted in successful identification of CRMs[1–4], but such approaches do not attempt to predict *ab initio* the gene expression patterns or functions of the genes regulated by the CRMs. Although algorithms have been developed recently for evaluating the regulatory importance of CRM binding site composition[5,6], thus far they have been unable to evaluate the vast sequence regions beyond the proximal promoter that must be considered in mammalian genomes.

Because of these complications, analyses of transcriptional regulatory elements in mammals have focused either on the prediction of CRMs starting with a collection of known co-regulatory transcription factors whose DNA binding specificities are known and a set of genes that the transcription factors may regulate[2,3,7,8], or on the computational identification of 'motif dictionaries'[9–12]. However, with the advent of high-throughput methods for assembling motif dictionaries, from either chromatin immunoprecipitations[13] or protein binding microarrays[14–16], the major computational problem to solve will shift from motif prediction to identifying and associating CRMs to both specific genes and biological processes[17].

Therefore, we developed a computational algorithm, Lever, that systematically identifies the target gene sets that are likely to be regulated by a query collection of candidate regulatory motifs. The ability to screen many gene sets with many motifs or motif combinations allows us to tackle the difficulty in *a priori* identification of co-regulated gene sets. Lever does not perform *de novo* motif discovery but rather evaluates an input collection of motifs for enrichment within candidate CRMs in the noncoding sequences flanking various input gene sets (**Fig. 1a**).

In this study we considered 75 kb of noncoding sequence flanking each gene (50 kb upstream to 25 kb downstream of transcription start site). Lever considers a collection of user-defined gene sets; in this study, we considered Gene Ontology (GO) categories and clusters of coexpressed genes as our gene sets of interest. We examined differentiation of human myoblasts into myotubes and considered 101 myogenic gene sets and 174 candidate regulatory motifs. We define a 'GM pair' to be the pairing of an individual gene set with a particular query motif or motif combination. Specifically, for each GM pair, Lever evaluates the degree to

[1]Division of Genetics, Department of Medicine, [2]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, and [3]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology (HST), Harvard Medical School, Harvard Medical School New Research Building, Room 466D, 77 Ave. Louis Pasteur, Boston, Massachusetts 02115, USA. [4]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA. [5]Department of Biology, MIT, 77 Massachusetts Ave., Cambridge, Massachusetts 02139. [6]Present address: Science Applications International Corporation–Frederick Inc., 1700 W. 7th St., Frederick, Maryland 21702, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

RECEIVED 21 DECEMBER 2007; ACCEPTED 4 FEBRUARY 2008; PUBLISHED ONLINE 2 MARCH 2008; DOI:10.1038/NMETH.1188

which the noncoding sequences surrounding the transcription start sites of the genes in the gene set are enriched for candidate CRMs comprising the given motif or motif combination under consideration as compared to a random background set of genes.

To predict candidate CRMs, we developed a tool, termed PhylCRM, which quantifies both motif conservation[18] and site clustering across multiple genomes. We experimentally validated several predicted CRMs from among the most significant ($Q \leq 0.05$) GM pairs in this study. Lever considered only the highest-scoring candidate CRM for each gene (**Fig. 1**). Each such GM pair can be thought of as an individual element of a gene set by motif or motif combination matrix (**Fig. 1b**). In this study, we assessed more than 25,000 GM pairs.

Identification of significant GM pairs from Lever analysis allows one to assign functional annotation to motifs at the level of GO categories and gene expression patterns. Although prior studies attempted to broadly annotate motifs at the level of tissue specificity[9], Lever assigns specific functional annotation to metazoan motifs and thus provides an entrée into targeted experimentation aimed at understanding the logic of *cis*-regulatory elements. Lever can be applied to any cell type, gene expression pattern or organism of interest to connect regulatory motifs to their biological functions and to gain insight into the architecture of transcriptional regulatory networks.

RESULTS

Identification of CRMs by PhylCRM

Candidate CRMs are first identified and scored with PhylCRM (**Supplementary Figs. 1–3** online), which scans the genomes of interest for matches to an input set of regulatory motifs. PhylCRM combines data for individual motif occurrences scored on an alignment using the previously described MONKEY scoring scheme[18] into a single CRM prediction. PhylCRM can scan very long (here, 75-kb) genomic sequences for candidate CRMs by quantifying both motif clustering and conservation across arbitrarily many genomes using an evolutionary model consistent with the phylogeny of the genomes. In the Lever analyses described in this study, we used the phylogenetic tree containing all 8 sequenced mammalian genomes (human, chimp, macaque, mouse, rat, dog, cow and opossum; **Supplementary Fig. 4** online[19]). PhylCRM identified significantly scoring candidate CRMs of varying lengths, ranging from 20 to 500 bp, and scored them to identify the maximum scoring window for each gene (**Fig. 1b**). PhylCRM can also be used as a stand-alone program for CRM prediction.

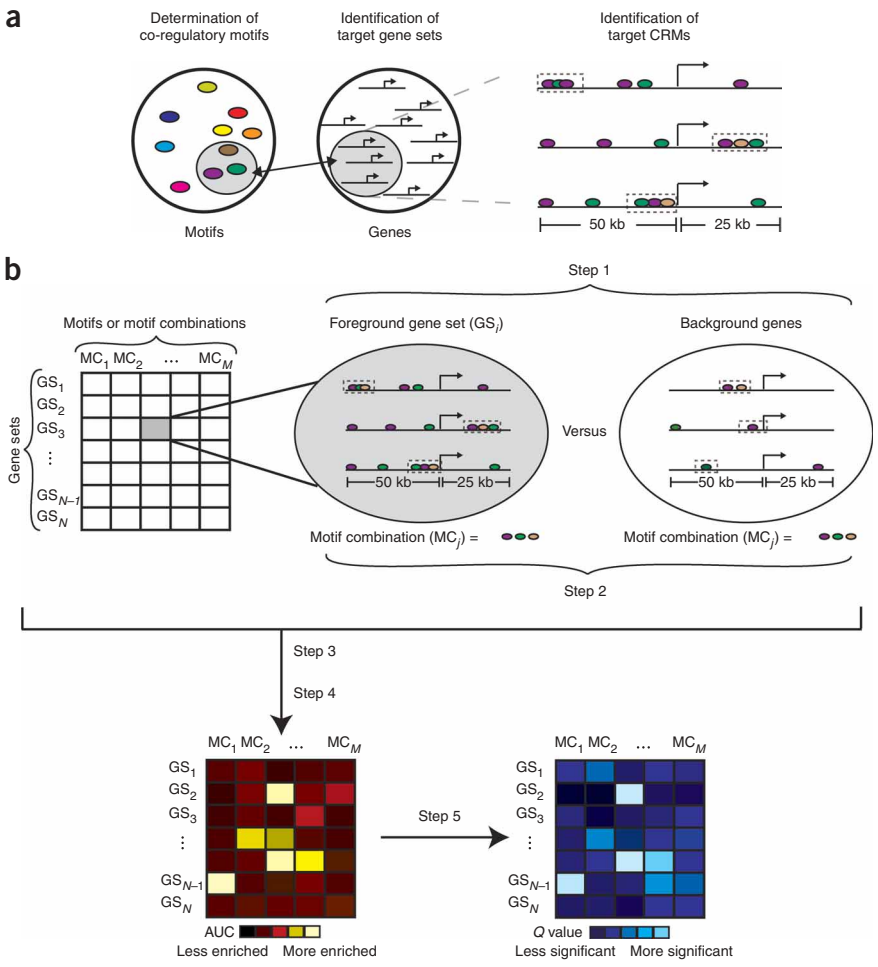

Figure 1 | Lever schema. (**a**) Lever simultaneously identifies: (i) motifs or motif combinations, (ii) their sets of co-regulated genes and (iii) *cis*-regulatory modules containing the enriched motifs or motif combinations. (**b**) Schematic depiction of the Lever scoring scheme. For each GM pair Lever searches for candidate CRMs (step 1) and, for each GM pair and all corresponding background genes, ranks the genes according to the PhylCRM score of each gene's single best-scoring candidate CRM (step 2). Lever evaluates the enrichment (AUC statistics) of a given GM pair (step 3) and repeats this for all other GM pairs (red and yellow matrix; step 4). The statistical significance of each AUC (indicated by a *Q* value, blue matrix) is calculated by permutation approach for multiple hypothesis correction (step 5).

Scoring GM pairs by Lever

We then input into Lever CRM scores for all genes in the genome (predicted by PhylCRM) and a collection of gene sets. To evaluate GM pairs, Lever first assigns to each gene in the 'foreground' gene set of interest input by the user, and to each gene in the automatically created length-matched 'background', the PhylCRM score of the best-scoring CRM. Considering all the genes in the foreground gene set and all background genes, Lever then ranks the genes according to the PhylCRM score of each gene's single best scoring candidate CRM (**Fig. 1b**). Then, for each entry in the GM pair matrix, Lever calculates both the value of the corresponding area under the curve in a receiver operator characteristic plot (AUC score) and its corresponding *Q* value (**Fig. 1b**). The AUC score indicates the probability that a randomly chosen member of the foreground gene set will rank higher than a randomly chosen background gene, whereas the *Q* value indicates the false discovery rate. In initial

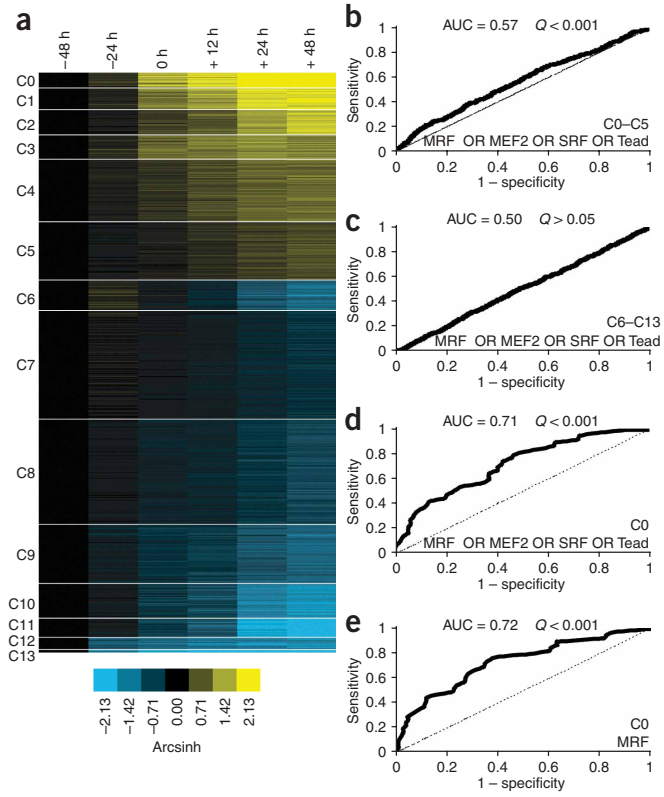

Figure 2 | Analysis of the time course of human skeletal muscle differentiation. (**a**) Expression clusters from gene expression profiling data for human adult primary skeletal muscle cells at the indicated time points with respect to stimulation of differentiation. Arcsinh values are relative to the −48 h time point. Shown here are the genes that are differentially expressed at a false discovery rate of 5%. (**b**–**e**) Evaluation of enrichment using as a foreground sequence set the 75-kb regions surrounding transcription start site for the indicated combinations of motifs for all genes in the indicated clusters. Dashed lines are receiver operator characteristic curves for a completely random ranking of genes into class 1 and class 0 and corresponds to AUC = 0.5.

positive-control analyses, we considered the four well-known myogenic transcription factor binding site motifs for the transcriptional activators[20] MEF2, serum response factor (SRF), Tead and the myogenic regulatory factors (MRFs) MyoD, Myogenin, Myf5 and Myf6, and showed that significant ($P < 10^{-7}$) motif enrichment can be detected when scanning 75-kb regions of genomic sequence (**Supplementary Fig. 4**).

Identification of myogenic gene sets to be examined by Lever

We considered two sources of gene sets: (i) clusters of coexpressed genes and (ii) GO categories. To identify appropriate gene expression clusters for examining the functions of motifs during myogenic differentiation, we first performed expression profiling over a time course of the differentiation of primary human skeletal myoblasts into myotubes at –24, –12, 0, +12, +24 and +48 h relative to stimulation of differentiation. We discovered 591 upregulated and 1,070 downregulated genes at a false discovery rate of 5%. Using *k*-means clustering, we partitioned these genes into 14 expression clusters, C0–C13 (**Fig. 2a** and **Supplementary Table 1** online), many of which showed enrichment for GO annotation terms consistent with myogenic differentiation (**Supplementary Table 2** online). We excluded cluster C13 from Lever analyses because it contained only 12 genes. As additional gene sets to be examined by Lever, we identified the GO categories that were significantly enriched within either the up- or downregulated genes during the timecourse of myogenic differentiation and took their intersection with either the up- or downregulated genes, yielding a final total of 101 gene sets. We did not use GO categories alone as gene sets in this study.

Lever evaluation on expression data and four myogenic motifs

We first applied Lever to systematically analyze each of the myogenic differentiation expression clusters considering all four of the myogenic motifs MRF, MEF2, SRF and Tead individually and also in Boolean (AND, OR and NOT) combinations. In evaluating the degree of enrichment for motifs within gene sets, we simultaneously considered the AUC and *Q* value. For example, when we examined the collection of all ~500 upregulated genes (C0–C5; **Fig. 2a**) using all four myogenic motifs, we observed only slight but significant enrichment (AUC = 0.57 ± 0.01, $Q \leq 0.001$; **Fig. 2b**). Thus, we can be highly confident that targets of these four motifs exist within the set of all upregulated genes, but finding specific target genes within this set would be difficult. Conversely, when we examined the set of all downregulated genes (C6–C13), we observed no enrichment at all with the four-way OR combination of these four motifs (AUC = 0.50 ± 0.01, $Q > 0.05$; **Fig. 2c**). We observed strongest enrichment for these four motifs among the most upregulated genes (C0; AUC = 0.71 ± 0.05, $Q \leq 0.001$; **Fig. 2d**). Within C0, the MRF motif alone showed slightly greater enrichment (AUC = 0.72 ± 0.04, $Q \leq 0.001$; **Fig. 2e**) than all four motifs together, indicating that most of the enrichment from the four-way Boolean OR combination of motifs was likely owing to the MRF motif.

We generally observed greatest enrichment of these four motifs in upregulated expression clusters (**Supplementary Fig. 5** and **Supplementary Table 3a** online), with the notable exception of C12 of downregulated genes, which contains many genes involved in cell-cycle function (**Supplementary Table 2**). The enrichment we observed here was consistent with an observation from another group suggesting the existence of MRF targets involved in cell-cycle progression and proliferation[21]. Results of additional Lever analysis controls are available in **Supplementary Results** online. Results of the Lever analysis of 101 myogenic gene sets using all four myogenic motifs are available in **Supplementary Figure 6** and **Supplementary Table 3b** online.

Lever screen of 174 motifs across 101 myogenic gene sets

To identify additional motifs that might be involved in the regulation of myogenic gene sets, we performed a Lever analysis of the 101 myogenic gene sets (**Fig. 3a**) using a dictionary of 174 candidate human regulatory motifs that were previously computationally predicted from 4-kb proximal promoter regions[9]. Out of these 17,574 GM pairs, we observed a total of 173 significant ($Q \leq 0.05$) GM pairs, involving a total of 45 distinct motifs and 61 distinct gene sets (**Fig. 3b**,**c** and **Supplementary Table 3c**). These 45 motifs could be broadly classified into 3 categories: (i) 21 motifs enriched among only upregulated gene sets, (ii) 10 motifs enriched among both upregulated and downregulated gene sets and (iii) 14 motifs enriched among only downregulated gene sets (**Fig. 3b**,**c**).

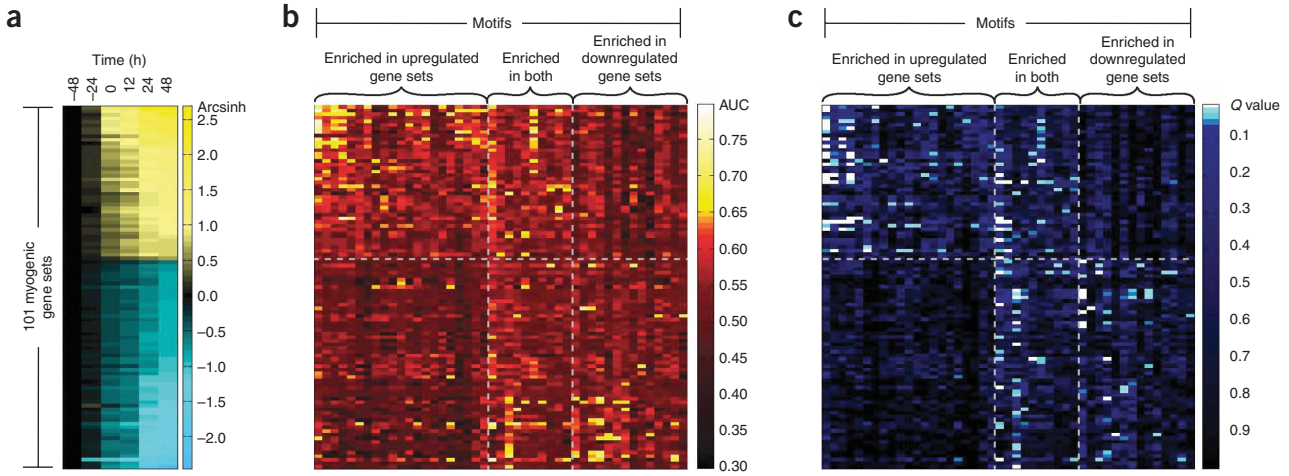

Figure 3 | Lever screen of 101 myogenic gene sets using a dictionary of 174 motifs. (**a**) Median signal intensity throughout the timecourse of gene expression profiling for each of 101 gene sets. (**b**) AUC scores for each GM pair when considering each of the 174 motifs from reference 9. (**c**) False discovery rate *Q* value for each GM pair. In the heat maps in **b** and **c**, there were only the 45 motifs with significant enrichment ($Q \leq 0.05$) in at least one of the 101 myogenic gene sets. The columns of matrices in **b** and **c** were sorted by decreasing overall correlation with gene expression at time +48 h. The rows of the heat maps in **a**–**c** were sorted by decreasing median expression arcsinh values at time point +48 h (relative to −48 h).

Several of the motifs that were part of significant GM pairs resulting from this Lever analysis correspond to the DNA binding site motifs of transcription factors known to function during myogenesis, including AP-1 (ref. 22), Elk-1 (ref. 23) and Pitx2 (ref. 24). This dictionary of candidate regulatory motifs contained matches to the MRF, MEF2 and Tead motifs, all of which were significantly enriched in various gene sets (**Supplementary Table 3**). For example, all of the motifs that we observed to be enriched within the sarcomeric gene set corresponded to discretized versions of either the MRF, MEF2 or Tead motifs.

In examining the results from this Lever analysis, we identified some interesting connections between gene sets. For example, the NF-Y motif was enriched among the upregulated lipid biosynthesis genes, the downregulated chromatin genes, various downregulated organelle gene sets and several downregulated gene sets involved in the cell cycle. Likewise, the downregulated plasma membrane genes appeared to be co-regulated via the AP-1 motif with several gene sets including response to stress, cell proliferation and regulation of cell proliferation, and the upregulated plasma membrane genes appeared to be co-regulated via the MEF2 motif with several upregulated gene sets involving structural properties of muscle cells, including cytoskeletal protein binding, contractile fiber, structural constituent of muscle and actin cytoskeleton.

Certain motifs appeared to regulate a large cohort of gene sets. For example, the NF-Y motif co-regulated many gene sets involved in the cell cycle. The suppression of NF-Y function has been shown previously to be important for the inhibition of several cell-cycle genes and the induction of the early muscle-specific program in post-mitotic muscle cells[25]. Similarly, the motif TGAnTCA (annotated in ref. 9 as the AP-1 motif) co-regulates several gene sets pertaining to cell proliferation and the plasma membrane. AP-1 complexes previously have been shown to be involved in the control of duration of myoblast proliferation and fusion efficiency[22].

Experimental validation of computationally predicted CRMs

We experimentally tested six CRMs predicted by PhylCRM (**Supplementary Fig. 7** online) and consisting of the MRF AND MEF2 motif combination (**Fig. 4a**). We sampled CRMs from various genomic locations relative to transcriptional start site and with a range of PhylCRM scores. Four of these six candidate CRMs were adjacent to genes with known or predicted sarcomeric function; two of these predicted CRMs (the predicted CRM next to *ACTA1* and the predicted CRM between PDLIM3 and SORBS2) are more than 17 kb away from their predicted target transcripts. Since Lever analysis identified significant enrichment (AUC = 0.82 ± 0.04, $Q \leq 0.001$) for the Boolean motif combination MRF AND MEF2 in the set of sarcomeric genes (**Supplementary Fig. 6**), choosing two of the six candidate CRMs to be adjacent to genes not involved in sarcomeric function also allowed us to explore whether CRMs containing this particular motif combination might function for non-sarcomeric genes.

The seven genes adjacent to these six predicted CRMs were upregulated during differentiation (**Supplementary Fig. 8** online), and myogenic transcription factors were differentially expressed at the protein level during differentiation (**Supplementary Fig. 9** online). Chromatin immunoprecipitation (ChIP) assays followed by region-specific quantitative PCR showed that 4/6 candidate CRMs were significantly enriched for binding by MEF2 ($P \leq 0.05$), MyoD ($P \leq 0.05$) and myogenin ($P \leq 0.005$) (**Fig. 4b**). Notably, of the six tested CRMs, the four that showed significant binding by MEF2, MyoD and myogenin were the ones that are located next to genes involved in sarcomeric function, whereas the two that did not show significant binding by these factors are not. Although this does not tell us what sequence features distinguish the active from the inactive CRMs, it does suggest that the choice of the likely target gene sets is important in predicting CRMs that are active in a given condition (here, myogenic differentiation).

We performed luciferase assays for the four new, candidate CRMs that were enriched for *in vivo* transcription factor binding. For these candidate CRMs we observed significant ($P \leq 0.05$)

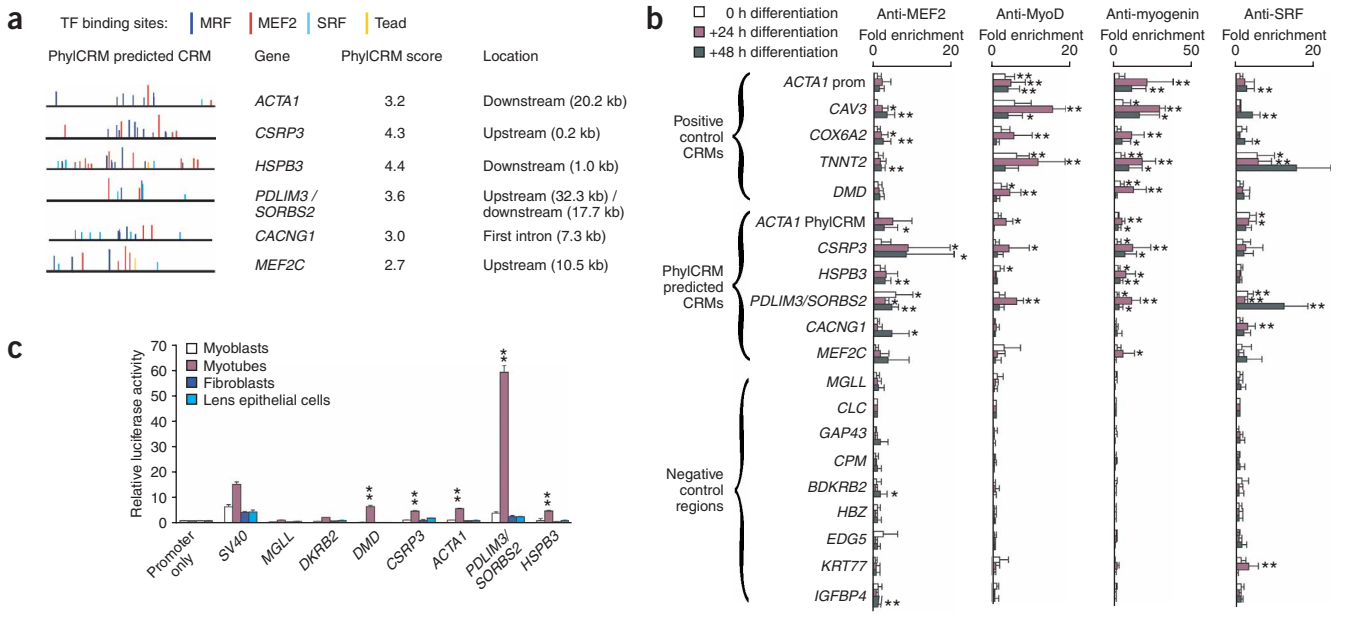

Figure 4 | Experimental validation of computationally predicted CRMs. (**a**) Predicted human CRMs. PhylCRM scores are $-\log_{10}$(PhylCRM *P*-value) of the given sequence window and the MRF AND MEF2 motif combination, which showed greatest enrichment among the sarcomeric gene set. Window locations are relative to transcriptional start or transcriptional stop of the nearest gene(s); intronic window locations are relative to transcription start sites. (**b**) Predicted CRMs were enriched for transcription factor occupancy during myogenic differentiation. Indicated antibodies were used in biological triplicate ChIP assays. Fold enrichment was calculated relative to mock ChIPs using anti–immunoglobulin gamma. $*P \leq 0.05$; $**P \leq 0.005$. "*ACTA1* prom" is a previously described muscle CRM; "*ACTA1* PhylCRM" was newly predicted. (**c**) Luciferase reporter assays for predicted CRMs indicate activity in myotubes. *MGLL* and *BDKRB2* are negative control regions; *DMD* is a positive control muscle CRM; *CSRP3*, *ACTA1*, *PDLIM3/SORBS2* and *HSPB3* are four predicted CRMs. $**P \leq 0.005$, increase in luciferase activity relative to the empty vector negative control. Error bars indicate s.d.

activation of luciferase expression during myogenic differentiation, but not in either fibroblasts or lens epithelial cells (**Fig. 4c**). ShRNA knockdowns of *MEF2D*, *MYOG* or *SRF* (**Supplementary Fig. 10** online) confirmed that these four candidate CRMs drive expression specifically in response to myogenic differentiation (**Supplementary Fig. 11** online). Results for a synthetic CRM suggest that there are further sequence requirements aside from the MRF and MEF2 motifs (**Supplementary Fig. 12** online). A detailed description of these experimental validations is available in **Supplementary Results**.

Functional annotation of regulatory motifs

Identification of significant GM pairs involving GO categories allowed us to assign to a regulatory motif the functional annotation of the GO categories within which it showed significant enrichment. For example, a discretized form of the MEF2 DNA binding site motif was enriched among many GO categories related to muscle contraction, including contractile fiber, muscle contraction and actin cytoskeleton, consistent with recently published ChIP-chip results[21]. Notably, Lever was able to identify these regulatory associations using only sequence data and gene expression data. In addition, although the published ChIP-chip study[21] identified surprisingly few MEF2, MyoD and myogenin targets as being involved in cell-cycle progression, our Lever results not only agreed with these findings, but also identified additional motifs, including AP-1, that are likely to be involved in the downregulation of the cell cycle during myogenesis. We missed several known regulatory interactions because of the stringency of our statistical analyses, primarily because of our need to correct for the many hypotheses tested (over 17,500 GM pairs) in our large Lever analysis of 174 motifs across 101 gene sets (**Supplementary Table 3**).

We can also apply this annotation method to the 13 motifs belonging to 30 significant GM pairs, for which the *trans* factors that may bind them have not yet been discovered (**Supplementary Table 3**). For example, we found that the putative regulatory motif TGACATY can be annotated as being involved in the regulation of plasma membrane genes. This level of functional annotation is much more specific than just indicating the tissue specificity of the genes upstream of which the motif is found[9]. We note that these annotations indicate the functions of the motifs during myogenic differentiation, and that the motifs may serve other functions in other cell types or in response to other environmental stimuli.

DISCUSSION

Our approach went beyond recent efforts at metazoan CRM identification by identifying motifs or motif combinations and their target gene sets in an automated manner. The level of functional annotation we achieved is an important step in moving from a listing of candidate regulatory motifs toward a functional understanding of the biological roles of such motifs. Our approach also allows for *de novo* reconstruction of transcriptional regulatory networks, without any prior knowledge of the functions of the examined regulatory motifs. We anticipate that this method will also be useful for the analysis of candidate regulatory motifs and gene sets from other biological systems, including other metazoans. Indeed, with motif dictionaries being derived either computationally or experimentally by high-throughput methods for identifying transcription factor DNA binding sites, the next major challenges

are identification of CRMs that contain those motifs and mapping those motifs and CRMs to the biological processes that they regulate. Lever analyses could be performed using any gene sets of interest. The utility of our computational framework will greatly increase in the coming years as expanded genome-wide motif dictionaries will be both predicted computationally[11] and derived experimentally[13,16] using genome-scale techniques.

Here we chose an appropriate subset of species to consider in scoring phylogenetic conservation, based on the evaluation of Lever on a positive control set of myogenic CRMs. However, the choice of the most suitable set of species to use will not always be determined as readily, particularly in the absence of a positive control set of CRMs. Future work on identifying the gene expression patterns of orthologous transcription factors will provide useful data for choosing the appropriate set of species to consider in evaluating phylogenetic conservation of their corresponding DNA binding site motifs. However, even with conservation of expression of the orthologous transcription factors, the binding site composition and locations of CRMs may still diverge rapidly[26].

This method represents a major step in moving from a genome-wide motif dictionary to understanding the language of *cis* regulation. Although Lever analysis does not directly inform us what sequence features in candidate CRMs distinguish the active from the inactive CRMs, it does suggest that the choice of the likely target gene sets is important in predicting CRMs that are active in a given condition (here, myogenic differentiation). Improved computational methods and experimental testing of both native and synthetic CRMs will be important for deciphering the 'grammar' of how regulatory motifs must be organized within sequence windows to construct CRMs that are active in a given cellular and environmental context.

METHODS

Genomic sequences used in this study. We obtained all genomic sequences for scans used in this work from the University of California Santa Cruz (UCSC) Genome Browser Hg17 assembly. For alignments, we used all genomes and alignments available at the time we began our study, corresponding to the "Multiple alignments of 8 vertebrate genomes with Human," along with pairwise alignments for macaque, cow and opossum. For annotation of gene coordinates, we used the UCSC "refGene" and "all_mrna" files. We repeat masked all sequences using the RepeatMasking provided by UCSC. We also masked out all exonic regions (exon coordinates were obtained from the refGene files).

We obtained a previously described[27] collection of 27 muscle CRMs containing matches to at least one of the MRF, MEF2, SRF or Tead DNA binding site motifs (we note that our "Tead" motif is the same as "Tef" motif in ref. 27). Genomic coordinates of positive control CRMs, negative control regions and PhylCRM predicted CRMs are available in **Supplementary Table 4** online.

PhylCRM: a computational approach for finding CRMs by quantifying motif clustering and evolutionary conservation. Briefly, PhylCRM takes as input a set of pre-defined DNA motifs, a set of aligned genomic sequences within which to search for candidate CRMs comprising a particular group of motifs and a tree indicating the phylogeny of the genomes. PhylCRM scans for the presence of transcription factor binding site motifs using sliding windows of continuously varying sizes, since CRMs span a wide range of lengths. For each motif, it scans the aligned sequences and quantifies the degree to which each position is a phylogenetically conserved motif match, using the MONKEY scoring model[18] to evaluate the degree to which that position is both a conserved and a high-affinity match to the transcription factor binding site motif (**Fig. 4a**). Then, for each transcription factor binding site motif and for each window within a user-defined size range, it computes the summation of these motif match scores and evaluates its statistical significance using an empirically derived probability distribution of the window scores to give a motif output score. This probability distribution depends on the transcription factor binding site motif and on the window size and is generated by inspecting all of the genomic sequences (here, 50-kb upstream and 25-kb downstream of transcription start site) with a sliding window of fixed size (**Supplementary Methods** online). The motif output scores from all of the motifs are combined into one output score. This output score is computed differently depending on the Boolean motif combination that is considered. This score simultaneously reflects motif over-representation and evolutionary conservation when scoring entire windows of sequence containing multiple transcription factor binding site motifs. Because PhylCRM provides a continuous (non-binary) measure of motif enrichment within a flanking region, we sought a similarly continuous set of logical AND, OR and NOT logical operations when combining several motifs. Therefore we used concepts from Fuzzy logic[28], where statements have a gradual assessment of being either 'true' or 'false'. A complete description of the PhylCRM scoring scheme is available in **Supplementary Methods** and **Supplementary Figures 1–3**.

Lever. The statistical framework of Lever is based on principles used by other groups for gene-set enrichment analysis[29] and uses permutation-based adjustment for multiple hypothesis testing. However, in contrast to gene-set enrichment analysis, in the Lever framework genes are ranked by a sequence-based, rather than an expression-based, scoring function, and each combination of motifs gives rise to a distinct scoring function. For each gene set and scoring function, the ranking power of the function is statistically assessed by calculating the enrichment for highly scoring genes within the gene set. Thus, Lever simultaneously calculates and assesses the enrichment for many gene sets across many motif combinations (that is, GM pairs).

Noncoding foreground and background sequence regions examined by Lever. For each gene in each of these foreground gene sets, we obtained 75 kb of genomic sequence overlapping the transcription start site. As a background set, we obtained a collection of non-overlapping, 75-kb genomic sequences for genes that were observed to be 'present' in the expression microarray data but not up- or downregulated at a false discovery rate of less than 0.1. For each foreground gene set we selected a length-matched background set[17] to remove the possibility that any observed enrichment for high-scoring candidate CRMs could be solely due to a larger search space. For each foreground gene set, a background gene set was automatically built that is as large as possible (usually 10–40 times as large as the foreground) so that the overall distribution of lengths in the foreground and background sets is well matched (**Supplementary Methods**).

Statistical analyses in data processing. We determined over-representation of GO annotation terms in various gene sets using FuncAssociate, a web-based program that corrects for multiple hypothesis testing[30]. Significant changes in luciferase reporter array and ChIP data were determined by Student's unpaired two-tailed *t*-tests.

Additional methods. Detailed descriptions of the construction of length-matched background sets against which foreground gene sets were evaluated in Lever; description of PhylCRM scoring scheme; evaluation of ability of PhylCRM to identify CRMs; comparison of PhylCRM to other CRM prediction methods; Lever; further discussion of interpretation of CRM enrichment results from Lever; position weight matrices used in this study; and details of all experimental protocols, including primer sequences, are available in **Supplementary Methods**.

Accession numbers. Gene Expression Omnibus (GEO): GSE4460.

Software. Software and manuals for PhylCRM and Lever are available at our laboratory website (http://the_brain.bwh.harvard.edu/).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank E. Margulies and the ENCODE Multiple Sequence Alignment working group for generously allowing use of their phylogenetic tree before its publication; S. Asthana, S. Sunyaev, G. Kryukov, M. Berger, T. Siggers and A. Aboukhalil for helpful discussions; J. Chee, E. Mathewson and T. Sierra for technical assistance; S. Elledge, A. Friedman, T. Siggers, M. Berger and F. De Masi for critical reading of the manuscript; A. Donner (Brigham & Women's Hospital) for the generous gift of human lens epithelial cells; and K. Cichowski (Brigham & Women's Hospital) for kindly providing lentiviral reagents. This work was funded in part by a PhRMA Foundation Informatics Research Starter Grant (M.L.B.), a William F. Milton Fund Award (M.L.B.), a Harvard-MIT Division of Health Sciences & Technology (HST) Taplin Award (M.L.B.) and US National Institutes of Health (NIH) National Human Genome Research Institute (R01 HG002966 to M.L.B.). J.B.W. was supported in part by an NIH Training Grant T32 HL07627 and NIH Individual National Research Service Award F32 AR051287. A.A.P. was supported in part by a National Defense Science and Engineering Graduate Fellowship from the Department of Defense and an Athinoula Martinos Fellowship from HST. S.A.J. was supported in part by a US National Science Foundation Postdoctoral Research Fellowship in Biological Informatics.

AUTHOR CONTRIBUTIONS

J.B.W. participated in the experimental design, performed the experiments and participated in analysis of the results and drafting of the manuscript. A.A.P. conceived of the PhylCRM scoring algorithm, participated in programming PhylCRM and running PhylCRM analyses, the development of Lever, programming Lever, running Lever analyses and analyzing the results and drafting of the manuscript. S.A.J. optimized the performance and participated in programming PhylCRM, running PhylCRM analyses, development of Lever, programming Lever and running Lever analyses and in analysis of the results and drafting of the manuscript. F.S.H. assisted with programming PhylCRM and running PhylCRM analyses. J.L. assisted with the experiments. M.L.B. conceived of the study and participated in the study design, analysis of the results and drafting of the manuscript.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201 (2003).
2. Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**, 656–668 (2006).
3. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
4. Pennacchio, L.A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
5. Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S. & Lawrence, C.E. Decoding human regulatory circuits. *Genome Res.* **14**, 1967–1974 (2004).
6. Zhou, Q. & Wong, W.H. CisModule: *de novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* **101**, 12114–12119 (2004).
7. Wasserman, W.W. & Fickett, J. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181 (1998).
8. Philippakis, A.A., He, F.S. & Bulyk, M.L. Modulefinder: a tool for computational discovery of *cis* regulatory modules. *Pac. Symp. Biocomput.* **10**, 519–530 (2005).
9. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
10. Elemento, O. & Tavazoie, S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.* **6**, R18 (2005).
11. Huber, B.R. & Bulyk, M.L. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics* **7**, 229 (2006).
12. Ettwiller, L. *et al.* The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol.* **6**, R104 (2005).
13. Bulyk, M.L. DNA microarray technologies for measuring protein-DNA interactions. *Curr. Opin. Biotechnol.* **17**, 422–430 (2006).
14. Bulyk, M.L., Huang, X., Choo, Y. & Church, G.M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. USA* **98**, 7158–7163 (2001).
15. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339 (2004).
16. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
17. Philippakis, A.A. *et al.* Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLOS Comput. Biol.* **2**, e53 (2006).
18. Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N. & Eisen, M.B. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* **5**, R98 (2004).
19. Margulies, E.H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**, 760–774 (2007).
20. Messenguy, F. & Dubois, E. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene* **316**, 1–21 (2003).
21. Blais, A. *et al.* An initial blueprint for myogenic differentiation. *Genes Dev.* **19**, 553–569 (2005).
22. Daury, L. *et al.* Opposing functions of ATF2 and Fos-like transcription factors in c-Jun-mediated myogenin expression and terminal differentiation of avian myoblasts. *Oncogene* **20**, 7998–8008 (2001).
23. Wang, Z. *et al.* Myocardin and ternary complex factors compete for SRF to control smooth muscle gene expression. *Nature* **428**, 185–189 (2004).
24. Martinez-Fernandez, S. *et al.* Pitx2c overexpression promotes cell proliferation and arrests differentiation in myoblasts. *Dev. Dyn.* **235**, 2930–2939 (2006).
25. Gurtner, A. *et al.* Requirement for down-regulation of the CCAAT-binding activity of the NF-Y transcription factor during skeletal muscle differentiation. *Mol. Biol. Cell* **14**, 2706–2715 (2003).
26. Ludwig, M.Z., Bergman, C., Patel, N.H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
27. Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J. & Lawrence, C. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26**, 225–228 (2000).
28. Kasabov, N.K. *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering* (MIT Press, Cambridge, Massachusetts, 1998).
29. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
30. Berriz, G.F., King, O.D., Bryant, B., Sander, C. & Roth, F.P. Characterizing gene sets with FuncAssociate. *Bioinformatics* **19**, 2502–2504 (2003).