Supplementary Methods for
"Systematic identification of mammalian regulatory motifs' target genes and functions"

Supplementary Information is available on the Nature Methods website and on our lab website, http://the_brain.bwh.harvard.edu/.

A. Construction of length-matched background sets against which foreground gene sets are evaluated in Lever

The following procedure is similar to the procedure we described previously in a *Drosophila* context[1]. We first ordered the search regions in each gene set by length. We defined the "foreground regions" to be those regions upstream and downstream of the genes that belong to a given foreground Gene Set, and we defined the "non-foreground regions" to be the collection of all other regions (i.e., regions not upstream or downstream of genes that belong to a given foreground Gene Set). For each foreground region, we took the 2 non-foreground regions occurring directly above and below it in the length-based ranking as background regions. In the event that two or more foreground regions did not have background regions ranked between them, we continued to extend above and below them in the ranking so that the center of this local collection of background regions was the same as the center of their associated foreground regions. Hence, for each foreground region, we were able to initially associate 2 length-matched background regions. We measured the AUC statistic for the lengths of the foreground and the background gene regions accumulated thus far and repeated the procedure of adding more non-foreground regions to the background set of gene regions until this AUC was close to 0.5, and until the background set was at least 10 times as large (and up to 40 times as large) as the foreground set, so that the distribution of the lengths of the foreground set of gene regions is similar to that of the background set of gene regions. The "PhylCRM_preprocess" program that generates the length-matched background sets of gene regions has a user-defined tolerance for what "close" means; in this study, we employed a tolerance of ± 0.02, i.e., for all foreground and background gene sets

considered in this paper, we required an AUC between 0.48 and 0.52 when ranking the foreground and background genes according to their lengths (AUC = 0.5 implies no difference between the distribution of lengths of foreground genes and that of the background genes).

B. Description of PhylCRM scoring scheme

The increasing number of sequenced genomes provides the opportunity for improved identification of regulatory regions by scanning for noncoding loci under negative selective pressure. To accomplish this, the evolutionary conservation must be scored in a way that the evolutionary history of the organisms is appropriately quantified; conservation of a locus between species sharing a recent ancestor should be weighted less than conservation between species that diverged long ago.

1. Scoring scheme and algorithm, one motif

In this section, we develop the scoring scheme for the case of only of one motif; in Section 2 we extend the scoring scheme to incorporate multiple motifs.

We begin with some notation. Given a base sequence g of length L from the genome being searched for TF binding site motif matches, let $a^{(i)}$, $i \in \{1,..,n\}$ denote the sequences aligned to g from each of the n organisms under consideration. We use $(g_j \ldots g_{j+k-1})$ to denote the subsequence of g beginning at position j and of length k, and we use $\left(a_j^{(i)} \mathrm{K}\ a_{j+k-1}^{(i)}\right)$ to denote the corresponding subsequence in the i'th alignment to g. Similarly, let H denote the (n + 1) × $|L|$-dimensional matrix storing both g and the $a^{(i)}$; thus, $H_{0,\bullet} = g$, $H_{i,\bullet} = a^{(i)}$, and $H_{\bullet,j}$ denotes the alignment column at position j (note that the $\bullet$ is used here to denote the collection of all values for that index position; see **Supplementary Fig. 1a** online). Finally, let T be the tree indicating the phylogeny of g and the $a^{(i)}$, let $\{v_\delta\}$ denote the ancestral vertices in T, and let $\{\tau_\varepsilon\}$ denote the branch lengths (see **Supplementary Fig. 1b** online).

For a given TF binding site motif of length m, let $M(\alpha,j)$ be the $4\times m$ matrix indicating the probability of observing the letter $\alpha \in \{A,C,G,T\}$ at position $j = 1,..,m$ of the motif (i.e., M is the frequency-derived probability matrix[2]), and let $Q(\alpha)$ denote the genomic frequency of letter α. For each position $j \in \{1,\ldots,L\}$ of g, we evaluate the degree to which $(g_j,\ldots g_{j+m-1})$ matches M with the quantity[2]:

$$\textbf{Eqn. 1)} \qquad \lambda(j) = \sum_{k=j}^{j+m-1} \log_2\left(\frac{M(g_k,k)}{Q(g_k)}\right)$$

This quantity is the commonly used position weight matrix score[2]. If $\lambda(j)$ is greater than a user-specified cutoff c, which is usually set to 1 or 2 standard deviations below the motif mean for the standard likelihood ratio score of the PWM model M and the genomic frequencies given by Q, we evaluate the degree to which this motif match is conserved throughout the phylogeny using an evolutionary model first suggested by Halpern and Bruno[3] and developed by Moses, Eisen and colleagues[4,5] (henceforth referred to as the MEHB model). In their approach, the degree of evolutionary conservation for the match to the TF binding site motif is scored by taking the log-likelihood ratio of observing the given collection of sequences throughout the phylogeny under the MEHB model as compared to a neutral model of evolutionary change:

$$\textbf{Eqn. 2)} \qquad \varphi(j) = \sum_{k=j}^{j+m-1} \log_2\left(\frac{P_{MEHB}\left(H_{\bullet,k}\middle|T,M,Q\right)}{P_{neutral}\left(H_{\bullet,k}\middle|T,Q\right)}\right) - c$$

Here, $P_{MEHB}\left(H_{\bullet,k}\middle|T,M,Q\right)$ represents the probability of observing $H_{\bullet,k}$ under the evolutionary model where nucleotide substitutions occur along T with a frequency specified by the MEHB proportionality (i.e., with fewer changes expected at the most

conserved positions of the motif; see **Supplementary Fig. 1c** online), and $P_{neutral}\left(H_{\bullet,k}\middle|T,Q\right)$ represents the probability of observing $H_{\bullet,k}$ under a neutral evolutionary model (either Jukes-Cantor[6] or Hasegawa-Kishino-Yano[7]). We have schematized how these probabilities are computed for a small phylogenetic tree in **Supplementary Fig. 1c** online.

Let ξ be an array of length L (i.e., the same length as g) and initialized so that, for all j, $\xi(j) = 0$. When a match to the motif M is made (i.e., $\lambda(j) > c$) in g beginning at position j, then, for $k = j,.., j+m-1$, ξ is updated according to:

Eqn. 3) $$\xi(k) = \max(\varphi(j)/m, \xi(k))$$

Here, the max is taken so that, in the event of overlapping motif matches, both matches contribute to the score, but there is no double-counting of scores. This rationale is schematized in **Supplementary Fig. 1d** online, where $\xi(j)$ is schematized for a sequence g and motif M. Note that we shall refer to quantity $\xi(j)$ as the "positional score for M" at j.

We wish to find sub-windows of the base sequence g that have a statistically significant over-representation of high-scoring matches to M. We do this by deriving the probability distribution function of the sub-windows of a fixed size within an *a priori* specified size range that best fits our data. We then use this probability distribution function in order to evaluate the enrichment of better scoring sub-windows of this size as compared to a given query sub-window under consideration. We also use the derived probability distribution

functions in order to combine the scores from several motifs of interest in the Fuzzy Boolean logic framework (see Section **3.** below).

Specifically, for each window size we derive the shape and the parameters of the null distribution. This is done by fitting a mixture model of three probability distribution functions – Delta, Uniform and Gamma – on a collection of sequences g_b of total length L_b that are believed not to be enriched for matches to motif M (we henceforth refer to this as the "background" sequence). Briefly, the Delta function is used to model the jump in score that occurs when a window of genomic sequence contains the initial portion of a motif at its left-most or right-most edge; the Uniform distribution is used to model the increase in score that occurs as the window contains an increasingly greater portion of the motif at either of its edges; finally, the Gamma distribution is then used for the bulk of the distribution to model an increasing number of binding sites and their evolutionary conservation.

Let w_j be a window of sequence in g of length $|w|$ and beginning at position j; we wish to evaluate whether this window is enriched for instances of M. Consider the following quantity:

Eqn. 4) $$\Xi(w_j) = \sum_{j'=j}^{j+|w|-1} \xi(j')$$

For a motif M and fixed *a priori* window size |w|, we wish to model the distribution of scores $\Xi(w_j) = \sum_{j'=j}^{j+|w|-1} \xi(j')$ under the null hypothesis of no motif enrichment. We shall refer to $\Xi(w_j)$ as the "window score" of w_j and, for a given window w_j, we shall determine

whether $\Xi(w_j)$ is statistically significantly large by estimating the p-value with respect to the modeled distribution at $\Xi(w_j)$.

In order to see how well the window scores $\Xi(w_j)$ are modeled by this mixture of three distributions, we considered the four motifs utilized in this paper: MRF, MEF2, SRF and Tead (see **Supplementary Fig. 2** online). For this analysis we utilized the foreground and background 75-kb regions shown in **Supplementary Fig. 4** online, where the foreground sequences contain a collection of 27 CRMs known to drive expression in muscle and background regions are a collection of 1,080 75-kb regions surrounding genes that were not up- or down-regulated during our time-course analysis of myogenesis. In **Supplementary Fig. 2** online, we have plotted the empirical distribution of $\Xi(w_{100})$ (blue curve) for each of these four motifs, as well as the fitted mixture model (red curve). As can be seen, the match between the fitted and empirical curves is very precise (we note that the fit for Tead is somewhat worse, as it is an infrequently occurring motif, and there are thus very few windows of genomic sequence comprising the right tail of the empirical distribution).

We then define the "output score" for the window to be the negative-log of its corresponding p-value:

Eqn. 5) $$\text{output score} = -\log_{10} P(\Xi(w)).$$

In **Supplementary Fig. 2** online, we have plotted the empirical output score (blue curve) for each of the four motifs mentioned above, as well as the output score from the fitted mixture model (red curve).

Finally, there are two related technical issues that must be addressed in building the array of positional scores ξ. First, due to the difficulties in aligning distant genomes, as well as the presence of sequencing gaps resulting from a genome being incompletely sequenced, there may not be any alignment to g at position j in genome $a^{(i)}$. Thus, it is not clear how to evaluate **Eqn. 2** in the presence of such missing data. Second, there is the possibility that a binding site may be truly present in g but lost (due to evolution) in $a^{(i)}$, particularly if $a^{(i)}$ and g are greatly diverged. In such a situation, it is possible that the quantity φ of **Eqn. 2** will be negative, which is undesirable since it is reasonable to assume that observing the presence of a motif match in a window w_j should increase (not decrease) the window score $\Xi(w_j)$, even if this match is not well-conserved. We handle these issues in a similar fashion by restricting to an appropriate sub-tree of the original tree. In the first scenario, the branches corresponding to genomes with missing alignments are removed; in the second scenario, any binding sites not scoring above the user-specified cutoff for determining a motif match are removed (**Supplementary Fig. 1e** online). We note, however, that for the second scenario it is also possible to run the program so that the entire phylogeny for which alignments are available is considered, even if there is not a motif match in some genomes (such a mode might be used, for example, in attempting to identify exclusively those TF binding sites conserved throughout the phylogeny, as was done in the original work by Moses *et al.*[5]).

2. Flexible scoring scheme and algorithm, multiple motifs

In this section, we assume the case of multiple motifs M_n, $n=1,..,N$. Let $\xi_n(j)$ hold the positional scores of motif M_n. We desire a means of measuring whether a given window w_j is enriched for motif matches. We allow flexibility in the scoring scheme by allowing the user to address the situation of potentially overlapping motifs (refer to the "-DEOVERLAP" option in the algorithms). A naïve approach would be to first define the array:

Eqn. 6) $$\hat{\xi}(j) = \max_n \{\xi_n(j)\}.$$

The score for a window w_j could then be obtained by calculating the significance of:

Eqn. 7) $$\hat{\Xi}(w_j) = \sum_{j'=j}^{j+|w|-1} \hat{\xi}(j').$$

This method has the advantage of appropriately handling overlapping motifs. Unfortunately, it has the disadvantage that the behavior of the score is dominated by the degree of enrichment for the most frequently occurring motifs. For example, assuming similar degrees of degeneracy, a motif of width 6 occurs more than twice as frequently as a motif of width 12, but the contribution of each match of the 6-mer motif to $\hat{\Xi}$ is half that of the motif of width 12.

Therefore, we describe an alternative means of scoring multiple motifs when the "-DEOVERLAP" option is specified (which is the option we employed in this Warner *et al.* paper). First, define:

Eqn. 8) $$\widetilde{\xi}_n(j) = \begin{cases} \xi_n(j) \; if \; \xi_n(j) = \max_{n'}\{\xi_{n'}(j)\} \\ 0 \; otherwise \end{cases}$$

Similar to the case of one motif, this step removes the possibility that the score for different motifs could be double-counted at position *j*, but also ensures that each position receives the score of the motif that best matches it. We shall refer to the $\widetilde{\xi}_n$ as the "de-overlapped" positional score; this de-overlapping step is schematized in **Supplementary Fig. 3a** online. The de-overlapping step is also performed for the background sequences g_b.

From now on, let $\widetilde{\Xi}_n(w_j)$ be the window score of w_j (with or without the "-DEOVERLAP" option specified), and let $\gamma_n(\widetilde{\Xi}_n;|w|)$ be the corresponding mixture distribution of scores $\widetilde{\Xi}_n$ (see **Eqn. 7**) for a motif M_n for a given window length $|w|$ under the null hypothesis of no enrichment.

3. Combinations of several motifs in Fuzzy logic framework

We wish to utilize the mixture distributions $\gamma_n(\widetilde{\Xi}_n;|w|)$ for a motif M_n in order to determine the statistical significance of observing a given degree of clustering and evolutionary conservation for the set of motifs. In the case of one motif, this computation was straightforward, as the statistical significance was directly obtainable from the tail of the appropriate mixture of Delta, Uniform and Gamma distributions. For many motifs we have developed a rich vocabulary of scoring schemes, in order to model the combinatorial interactions between the TFs under consideration.

For simplicity, take the case of two motifs M_n and M_m. It is possible to calculate statistical significance using a "restrictively-defined tail" (**Supplementary Fig. 3b** online):

Eqn. 9) $$P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right) = P_n\left(\tilde{\Xi}_n\right) P_m\left(\tilde{\Xi}_m\right) = \left(\int_{\tilde{\Xi}_n}^{\infty} \gamma_n(\Xi;|w|)d\Xi\right)\left(\int_{\tilde{\Xi}_m}^{\infty} \gamma_m(\Xi;|w|)d\Xi\right)$$

(note: $P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right)$ does not refer to the joint distribution of the random variables $\tilde{\Xi}_n$ and $\tilde{\Xi}_m$).

We take the "output score" to be $-\log\left(P_n\left(\tilde{\Xi}_n\right)P_m\left(\tilde{\Xi}_m\right)\right) = -\log\left(P_n\left(\tilde{\Xi}_n\right)\right) - \log\left(P_m\left(\tilde{\Xi}_m\right)\right)$, and so the output score is additive in the number of motifs. Hence, a given window can achieve significance if it is greatly enriched for matches to *either* motif one or motif two (OR combination).

Conversely, it is also possible to calculate statistical significance of a combination of distributions using a "generously defined tail" (**Supplementary Fig. 3c** online):

Eqn. 10) $$\begin{aligned} P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right) &= 1 - \left(\int_0^{\tilde{\Xi}_n} \gamma_n(\Xi;|w|)d\Xi\right)\left(\int_0^{\tilde{\Xi}_m} \gamma_m(\Xi;|w|)d\Xi\right) \\ &= P_n\left(\tilde{\Xi}_n\right) + P_m\left(\tilde{\Xi}_m\right) - P_n\left(\tilde{\Xi}_n\right)P_m\left(\tilde{\Xi}_m\right) \end{aligned}$$

Here, if $\tilde{\Xi}_n = 0$ (the window score is zero), then $P_n\left(\tilde{\Xi}_n\right) = 1$ and so $P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right) = 1$ and so the window score $-\log\left(P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right)\right) = 0$ (and similarly for the case where $\tilde{\Xi}_m = 0$). Thus, using this tail, a window must be enriched for *both* motifs (AND combination) under consideration in order to be statistically significant.

Finally, it is possible to define the combination of the distributions in more complicated ways. For example, the following combination would assign a high score to windows of sequence that are enriched for the first motif but specifically not enriched for the second (NOT combination; **Supplementary Fig. 3d** online):

Eqn. 11)

$$P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right) = 1 - \left(\int_0^{\tilde{\Xi}_n} \gamma_n(\Xi;|w|)d\Xi\right)\left(\int_{\tilde{\Xi}_m}^{\infty} \gamma_m(\Xi;|w|)d\Xi\right)$$
$$= 1 - P_m\left(\tilde{\Xi}_m\right) + P_n\left(\tilde{\Xi}_n\right)P_m\left(\tilde{\Xi}_m\right)$$

The cases we have described, Eqns. 9-11, can be thought of as Fuzzy logic rules for the discrete Boolean logical functions (M_n OR M_m), (M_n AND M_m), and (M_n AND NOT M_m). In general, we define the "output score" for a Fuzzy logic combination of multiple motifs to be the negative-log of the corresponding P (see Eqns 9-11):

Eqn. 12) $$\text{output score} = -\log_{10}\left(P\left(\tilde{\Xi}_n, \tilde{\Xi}_m\right)\right).$$

We have implemented PhylCRM so that a variety of different tails are possible, in order to allow the evaluation of a more nuanced view of *cis* regulatory logic. A summary of all Fuzzy logic combinations considered is listed below:

a. OR combinations of arbitrarily many motifs

b. AND combinations of arbitrarily many motifs

c. The following four classes of compound combinations involving up to 4 motifs:

1) (M_1 AND NOT M_2) (two motifs)

2) ((M_1 AND M_2) OR M_3) (three motifs)

3) ((M_1 OR M_2) AND M_3) (three motifs)

4) (M_1 AND M_2) AND NOT M_3) (three motifs)

5) ((M_1 AND M_2 AND M_3) AND NOT M_4) (four motifs)

Thus, if one would like to find CRMs enriched for any subset of the motifs under consideration, the OR mode is more appropriate; conversely, if one wishes to specifically identify CRMs enriched for matches to all the motifs under consideration, the AND mode is more appropriate.

C. Evaluation of ability of PhylCRM to identify CRMs

We obtained a phylogenetic tree of 11 vertebrate genomes from the ENCODE multiple sequence alignment working group[8] (**Supplementary Figure 4a** online) and a set of 27 CRMs previously compiled by Wasserman *et al.*[9] that are known to drive expression in muscle and to be regulated by at least one of the four well known myogenic TFs: a) MEF2, b) Serum Response Factor (SRF), c) Tead, and d) the myogenic regulatory factors (MRFs) MyoD, Myogenin, Myf5 and Myf6 (note that the motifs for the MRFs are currently indistinguishable and thus are encompassed by a single, general MRF motif)[9]. Here, we examined windows ranging between 50 and 500 bp (increment size of 50 bp), and utilized the phylogenetic tree derived by the ENCODE multiple sequence alignment working group[8]. The tree is input to PhylCRM in Newick format:

(((((((human:0.006690,chimp:0.007571):0.024272,
macaque:0.059256):0.107134,(mouse:0.077017,rat:0.081728):0.252613):0.023026,(dog:
0.147731,cow:0.159182):0.03945):0.262899,opossum:0.371073):0.189124,chicken:0.454
691):0.279364,(fugu:0.732855,zebrafish:0.782561):0.156067)

The versions of the genomes that we used are:

- human (hg 17)
- chimp (Nov 2003, panTro1)
- macaque (Jan 2006, rheMac2)
- mouse (May 2004, mm7)
- rat (Jun 2003, rn3)
- dog (May 2005, canFam2)
- cow (Mar 2005, bosTau2)

- opossum (Jun 2005, monDom2)
- chicken (Feb 2004, galGal2)
- zebrafish (May 2005, danRer3)
- Fugu (Aug 2002, fr1)

We compiled a "foreground" human gene set consisting of the 75-kb sequence regions surrounding each of these 27 known CRMs, and also a length-matched random "background" set of genomic regions not believed to contain muscle CRMs. We first masked out any coding regions and repetitive elements, and then searched the foreground and background gene sets with PhylCRM in order to identify windows of sequence significantly enriched for clusters of high-scoring, evolutionarily conserved matches to these four myogenic motifs. We assigned to each foreground and background region the score of its highest scoring PhylCRM window ranging between 10 bp and 500 bp, and then determined whether the foreground gene set scored higher than the background gene set by evaluating the AUC.

Without the use of phylogenetic conservation, we observed statistically significant enrichment for these motifs within this positive control foreground gene set (AUC = 0.64 ± 0.05; $P < 0.01$ calculated by the Wilcoxon-Mann-Whitney[10] (WMW) statistic; **Supplementary Figure 4b** online). When utilizing all 11 available vertebrate genomes, the degree of foreground enrichment increased significantly (AUC = 0.81 ± 0.05; $P < 10^{-7}$ by WMW; **Supplementary Figure 4c** online), demonstrating that the use of evolutionary conservation can increase discriminatory power.

Next, we evaluated whether the use of a subset of species in PhylCRM might yield higher foreground enrichment than the use of all available vertebrate genomes for this positive control set of myogenic CRMs. To evaluate such subsets, we systematically added those branches extending from each preceding common ancestor of human (**Supplementary Figure 4d** online). We observed the greatest degree of enrichment when using all available vertebrate genomes except those of chicken, pufferfish and zebrafish (AUC = 0.82 ± 0.05; $P < 10^{-8}$ by WMW), indicating that a judicious choice of sub-tree could yield improved performance. Finally, as a negative control we scanned the foreground and background regions with a permuted form of the four considered motifs and observed no enrichment (AUC = 0.41 ± 0.06; $P > 0.05$ by WMW; **Supplementary Figure 4e** online).

From this analysis, we concluded that PhylCRM can detect enrichment of motifs within 75-kb regions of genomic sequence within an appropriate gene set, and that the utilization of many aligned genomes increases the power of PhylCRM.

D. Comparison of PhylCRM to other CRM prediction methods

There are many available computational tools for CRM identification, and a full comparison of PhylCRM against each of them is beyond the scope of this present study. Therefore, we have selected two computational tools against which to compare PhylCRM, as they have similar goals of taking as input a collection of TF binding site motifs and outputting target CRMs.

We compared the performance of PhylCRM to two other algorithms: Comet (which utilizes a hidden Markov model (HMM) based approach and does not utilize information on the evolutionary conservation of the TF binding site motifs) and Stubb (which also utilizes an HMM-based approach and incorporates information on evolutionary conservation across up to two species of interest – one base genome plus one alignment genome). We selected two data sets for comparison: 1) the collection of 27 known muscle CRMs previously compiled by Wasserman *et al.*[9] (the results of PhylCRM analysis for this collection of CRMs is shown in **Supplementary Fig. 4** online), and 2) the collection of "sarcomeric genes" highlighted in **Fig. 3** and **Supplementary Fig. 6**. Thus, these were the two sequence sets that were most carefully examined in this paper.

First, we took as a "foreground" set of sequences the 27 75-kb regions containing each of the known muscle CRMs (i.e., we considered the 75-kb regions within which the CRMs were located) as well as a length-matched background set of sequences (data #1). Next, we took as a "foreground" set of sequences the set of the 75-kb regions around transcription start of the 46 known sarcomeric genes, as well as a length-matched set of

background sequences (data #2). Because of computational limitations of the Stubb algorithm in handling large amounts of sequence, we had to reduce the size of the background data sets from what we used to generate the results shown in the main body of this paper (in this comparison, we used the same background to evaluate the results from all three programs – PhylCRM, Comet, and Stubb – in order to ensure that they were compared in a fair and systematic way). Consequently, the performance of PhylCRM shown below is slightly different from the results shown in **Supplementary Figure 4** online.

We ran the three programs by varying the input parameters in order to obtain the best performance from each program. We compared Comet, PhylCRM and Stubb by utilizing the same measure of performance as that utilized in the main text, namely, the AUC statistic that indicates the degree to which foreground sequences are ranked higher than background sequences (see the table below for a summary of the results). First, we observed that when no phylogeny was utilized the performance of PhylCRM on data #1 was AUC = 0.70 ± 0.06 (error represents 1 standard deviation determined by applying bootstrap) ($P < 10^{-3}$); this is within the margin of error of the performance observed for Comet (AUC = 0.70 ± 0.05, $P < 10^{-4}$) and for Stubb (AUC = 0.68 ± 0.05, $P < 10^{-3}$) on data #1. On the sarcomeric gene set (data #2), without utilizing phylogeny, PhylCRM (AUC = 0.64 ± 0.05, $P < 10^{-2}$) was within the margin of error of Comet (AUC = 0.60 ± 0.05, $P > 0.01$), but better than Stubb (AUC = 0.49 ± 0.05, $P > 0.1$).

We then examined how PhylCRM compares with Stubb in the case when information on the evolutionary conservation of the binding sites is utilized. We note that Stubb currently can consider conservation between only two species, while PhylCRM can utilize arbitrarily many genomes. On data #1, using the phylogenetic tree: Human/Chimp/Macaque/Mouse/Rat/Dog/Cow/Opossum, the performance of PhylCRM (AUC = 0.81 ± 0.06, $P < 10^{-6}$) was within the margin of error of Stubb when using human and mouse (AUC = 0.80 ± 0.05, $P < 10^{-6}$). We note that many of the CRMs in data set #1 were originally discovered in mouse and other non-human species[11], and this bias in the creation of this positive control data set may have resulted in their being better conserved in mouse. Using the same phylogenetic tree (Human/Chimp/Macaque/Mouse/ Rat/Dog/Cow/Opossum) but now considering the Sarcomeric gene set (data #2), PhylCRM (AUC = 0.74 ± 0.05, $P < 10^{-6}$) performed significantly better than Stubb (AUC = 0.59 ± 0.04, $P > 0.01$) when Stubb was run utilizing human and mouse.

Table S1: Summary of algorithm comparison

Algorithm:	Wasserman data: (without utilizing phylogeny)	Wasserman data: (with phylogeny)	Sarcomeric data: (without utilizing phylogeny)	Sarcomeric data: (with phylogeny
Stubb	AUC = 0.68 ± 0.05	AUC = 0.80 ± 0.05	AUC = 0.49 ± 0.05	AUC = 0.59 ± 0.04
PhylCRM	AUC = 0.70 ± 0.06	AUC = 0.81 ± 0.06	AUC = 0.64 ± 0.05	AUC = 0.74 ± 0.05
Comet	AUC=0.70±0.05	N/A	AUC = 0.60 ± 0.05	N/A

From these comparisons we conclude that that PhylCRM performs comparably to the other algorithms on the collection of 27 known CRMs, and better on the Sarcomeric gene

set. Additionally, PhylCRM has the added feature of being able to score CRMs using a rich vocabulary of Fuzzy Boolean logic rules in order to discover nuanced *cis* regulatory codes (in the preceding comparisons, we utilized the OR combination for simplicity, although the performance could possibly be improved with a different combination of TF binding site motifs but would have complicated a direct comparison with the other algorithms). We show that in all of the datasets considered, using phylogeny information helps to improve the performance (this is also shown in **Supplementary Figure 4** online). Also, we expect that the performance of PhylCRM will continue to improve on these data sets (and other data sets) as more mammalian genomes are sequenced.

E. Lever

The statistical framework of Lever is based upon principles used by other groups for gene set enrichment analysis[12,13] and utilizes permutation-based corrections for multiple hypothesis testing[14]. However, in contrast to gene set enrichment analysis[12,13], 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 (i.e., GM-pairs).

1. Statistical assessment for enrichment

Let $g_1, g_2, \ldots g_G$ be a collection of G genes whose upstream/downstream/intronic regions are being searched for CRMs, and let $GS_1, GS_2, \ldots, GS_N$ be a collection of subsets of these genes. Within each subset GS_j, the genes g_i which are elements of it will be labeled as either being "foreground" or "background". To denote this labeling, we use the matrix Y where:

Eqn. 1)

$$Y_{i,j} = \begin{cases} 1 & \text{if } g_i \text{ is a foreground gene in set } GS_j \\ 0 & \text{if } g_i \text{ is a background gene in set } GS_j \\ \bullet & \text{if } g_i \notin GS_j \end{cases}$$

The final value (•) of the above equation serves as a set membership indicator, which is used for efficient processing in order to assemble all of the required sets of genes. Specifically, information on set membership is required in a later permutation-based approach for evaluation of statistical significance, during which the assignment of genes

to the various gene sets changes. Let $F_{S_j} \subset GS_j$ and $B_{S_j} \subset GS_j$ be the sets of all foreground and background genes, respectively, within GS_j, and let $\left|F_{GS_j}\right|$ and $\left|B_{GS_j}\right|$ be the number of foreground and background genes, respectively, within GS_j. Finally, let MC_k , $k = 1,..,M$ denote a given collection of combinations of motifs, and let the matrix $X=(X_{i,k})$, $i=1,..,G$, $k = 1,..,M$, where $X_{i,k}$ denote the PhylCRM score (see **Supplementary Figures 1-4** online) of the maximum scoring window within the flanking sequence of g_i when scanning it with a motif combination MC_k.

Our goal is to determine which combinations of motifs MC_k are significantly enriched within the various gene sets GS_j. We consider the ranked PhylCRM scores for each combination of motifs and utilize the AUC statistic of the ranked scores in order to evaluate this enrichment. The AUC statistic is broadly applied for bipartite ranking problems and for comparisons of performance of binary classifiers[15]:

Eqn. 2) $$AUC\left(GS_j, MC_k\right) = \left(\frac{1}{\left|F_{GS_j}\right|\left|B_{GS_j}\right|}\right)\left(\sum_{i:Y_{i,j}=1}\sum_{i':Y_{i',j}=0} I_{[X_{i,k}>X_{i',k}]} + \frac{1}{2} I_{[X_{i,k}=X_{i',k}]}\right)$$

where I is the indicator function taking the value of "1" if the statement in brackets is true and "0" otherwise. The AUC of a ranking function takes values in the range [0,1], and is the probability that a randomly chosen positive instance (a member of the foreground set) will rank higher than a randomly chosen negative instance (a member of the background). It will take the value "1" if all of the genes in the foreground rank higher than genes in the background, the value "0" if all of the genes in the foreground rank lower than genes

in the background, and a value close to 0.5 if the ordering of foreground genes is not biased toward higher or lower ranks.

2. Adjustment for multiple hypothesis testing

An explicit goal of Lever is to evaluate many pairings of gene sets and motifs or motif combinations simultaneously, in order to identify motif combinations exhibiting statistically significant enrichment in specific gene sets (we refer to a matching of a gene set and a motif combination (GS_j, MC_k) as a GM-pair). The evaluation of so many GM-pairs, however, necessitates a mechanism to correct for multiple hypothesis testing. Observe that AUC scores of distinct pairings (GS_j, MC_k) and ($GS_{j'}$, $MC_{k'}$) are not independent under the null hypothesis of no enrichment, since GS_j and $GS_{j'}$ may contain common genes and MC_k and $MC_{k'}$ may contain common motifs. Consequently, a simple Bonferroni correction for multiple hypothesis testing is overly conservative and would cause many biologically relevant pairings (GS_j, MC_k) to be missed. Therefore, we applied a permutation-based approach for evaluation of statistical significance that takes into account the non-independence of the hypotheses.

For a given gene g_i let $\vec{Y}_i = \left(Y_{i,1}, \quad Y_{i,2}, \quad \text{K} \quad Y_{i,S-1}, \quad Y_{i,N}\right)$ be the row vector of Y indicating membership of g_i in each of the sets GS_j, $j = 1,..,N$ and let $\vec{X}_i = \left(X_{i,1}, \quad X_{i,2}, \quad \text{K} \quad X_{i,M-1}, \quad X_{i,M}\right)$ be the row vector of X indicating the PhylCRM score of g_i for each combination of motifs MC_k $k = 1,..,M$. Let π be a fixed permutation of $\{1,..,G\}$ (where G is the total number of genes).

Next, let:

Eqn. 3) $$AUC\left(GS_j, MC_k, \pi\right) = \left(\frac{1}{\left|F_{GS_j}\right|\left|B_{GS_j}\right|}\right)\left(\sum_{i: Y_{i,j}=1}\ \sum_{i': Y_{i',j}=0} I_{[X_{\pi(i),k} > X_{\pi(i'),k}]} + \frac{1}{2} I_{[X_{\pi(i),k} = X_{\pi(i'),k}]}\right)$$

This is the AUC computed for the GM-pair (GS_j, MC_k) when the class labels are permuted. Observe that, as desired, the definition of this permutation preserves all correlations in values of AUC statistics between pairings (GS_j, MC_k) and ($GS_{j'}$, $MC_{k'}$) resulting from genes being elements of both GS_j and GS_k and motifs being elements of both MC_k and $MC_{k'}$.

We use the permutation approach in order to evaluate the significance of the values AUC(GS_j, MC_k) when controlling for false discovery rate (FDR) and family-wise error rate for multiple comparisons. Let $\{\pi_l\}_{l=1}^{P}$ be a collection of P randomly chosen permutations over the gene labels. Because different gene sets GS_j , $j = 1,..,N$ contain different numbers of genes, and because different motif combinations can result in more or fewer ties in PhylCRM scores between distinct genes (for example, AND combinations involving many motifs may result in many genes having a PhylCRM score of "0"), the variance of AUC(GS_j, MC_k) is not constant across pairings (GS_j, MC_k). Let:

Eqns. 4 and 5)

$$\mu_{j,k} = \frac{1}{P}\sum_{l=1}^{P} AUC\left(GS_j, MC_k, \pi_l\right)$$

$$\sigma_{j,k} = \left(\frac{1}{P-1}\sum_{l=1}^{P}\left(AUC\left(GS_j, MC_k, \pi_l\right) - \mu_{j,k}\right)^2\right)^{1/2}$$

We normalize the AUC(GS_j, MC_k) value by applying the z-transformation:

Eqn. 6)

$$AUC\left(GS_j, MC_k\right)' = \frac{AUC\left(GS_j, MC_k\right) - \mu_{j,k}}{\sigma_{j,k}}$$

Following the method of Subramanian *et al.*[13], for family-wise error rate estimation of significance for each value AUC(GS_j, MC_k)′, we take the maximum of the normalized AUC statistics across all gene set and motif combination pairings within a given permutation:

Eqn. 7)

$$U_{\pi_l} = \max_{j,k}\left\{AUC\left(GS_j, MC_k, \pi_l\right)\right\}$$

The family-wise error rate estimate of statistical significance of a specific value AUC(GS_j, MC_k)′ is then given by:

Eqn. 8)

$$P\left[AUC\left(GS_j, MC_k\right)'\right]_{FWER} = percentage\ of\ (l)\ s.t. U_{\pi_l} \geq AUC\left(GS_j, MC_k\right)$$

Similarly, the FDR estimate of statistical significance is obtained by utilizing the entire distribution of AUC(GS_j, MC_k, π_l)′ values and by calculating the FDR *Q*-values, denoted as *Q* in the main text and in the figures:

Eqn. 9)

$$Q\left[AUC\left(GS_j, MC_k\right)'\right] = \frac{percentage\ of\ \left(j', k', l\right)\ s.t.\ AUC\left(GC_{j'}, MS_{k'}, \pi_l\right) \geq AUC\left(GS_j, MC_k\right)}{percentage\ of\ \left(j', k'\right)\ s.t.\ AUC\left(GC_{j'}, MS_{k'}\right) \geq AUC\left(GS_j, MC_k\right)}$$

In this paper, we report AUCs along with an error term that corresponds to one standard deviation of the bootstrap confidence interval[14].

3. Correction for AT/GC-rich motifs

We have observed that many genes of interest have G/C-rich flanking sequences; consequently, many gene sets will show artificially high enrichment for G/C-rich motifs. For the Lever screens shown in **Figure 3** and **Supplementary Figures 5-6**, we controlled for this by first generating many permuted forms of each motif (50 for analyses involving the Xie *et al.*[16] motifs, and 100 for analyses involving the four motifs MRF/MEF2/SRF/Tead). For each gene of interest, we scored its 75-kb flanking noncoding sequence with permuted forms of the motifs. For each gene and each motif or combination of motifs, we *z*-transformed the PhylCRM scores (similarly to **Eqns. 4** and **5**) after calculating the mean and variance from the permuted forms of the motifs. This approach showed reduction of the artifacts described above.

PhylCRM and Lever software parameter settings

For all runs and all motifs considered in this study, as the threshold cutoff used by Lever and PhylCRM for calling a motif match, we used 2 standard deviations (SDs) below the motif average[17] and the "-THRESHOLD" setting in both of these programs. For the PhylCRM results shown in **Supplementary Figure 4**, we used the "-DEOVERLAP" option for the OR combination of the MRF/MEF2/SRF/Tead motifs. We observed very similar trends without the "-DEOVERLAP" option, *i.e.* without removing the overlaps between different motifs. In the rest of this study, we applied PhylCRM and Lever without the "-DEOVERLAP" option. For PhylCRM and Lever runs involving the MRF/MEF2/SRF/Tead motifs, we used windows ranging between 10 and 500 bp, and for

runs involving the Xie *et al.*[16] motifs we used a window range of 25 to 500 bp since some of those motifs can be wider than 10 bp.

Gene sets examined in this study

For the Lever scans shown in **Supplementary Figure 5**, we examined each of the *k*-means expression clusters as an input library of foreground gene sets (we excluded cluster **C13** because it contained only 12 genes). For those shown in **Supplementary Figures 5** and **Figure 3,** we added to this collection by additionally considering gene sets based upon shared GO annotation terms (we considered the Biological Process, Molecular Function and Cellular Component terms). Specifically, significantly over-represented GO categories among the up- and down-regulated genes were determined using FuncAssociate[18]. Only the significantly (FDR $\leq$ 0.05) up- or down-regulated genes belonging to each of those GO categories were considered in constructing the corresponding reduced GO category gene sets. Nonredundant gene lists were created by matching Refseq sequences to common gene names using DAVID[19] and removing redundancies. Finally, we considered only those gene sets that contained at least 15 members. Also, if two gene sets were found to contain identical genes, one of the gene sets was dropped.

We noticed that numerous sarcomere-related GO categories, such as actin cytoskeleton, contractile fiber, structural constituent of muscle, muscle contraction, and muscle development, were enriched among the up-regulated genes. Sarcomeric genes might be especially likely to be co-regulated, as they are all components of a single protein

complex utilized in muscle. However, the GO category "sarcomere" contained only 12 genes observed to be up-regulated in our study. Therefore, knowing that GO annotation of mammalian genes can be quite incomplete, we manually compiled from the literature a list of 46 sarcomeric genes that were up-regulated during the differentiation of myoblasts into myotubes. This list of 46 genes included two genes (*ACTA1* and *CSRP3*) for which probes were not included on the microarrays utilized studying gene expression profiling, but for which RT-PCR experiments confirmed their up-regulation (**Supplementary Figure 8**).

F. Further discussion of interpretation of CRM enrichment results from Lever

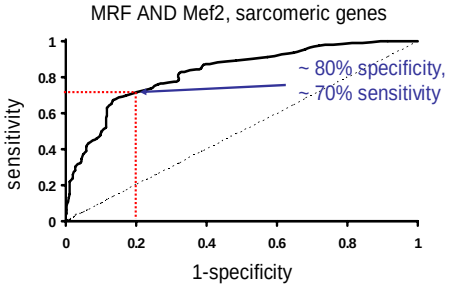
We note that Lever identifies CRM enrichment within a given gene set. Of the six tested CRMs, the four that showed significant binding by MEF2, MyoD, and myogenin were the ones that are located next to genes involved in sarcomeric function, whereas the two that did not show significant binding by these factors are not. The MEF2 AND MRF motif combination within the up-regulated sarcomeric gene set was one of our top 10 GM-pairs in terms of AUC and *Q*-value from our Lever screen of 101 myogenic gene sets and the four known myogenic motifs MRF, MEF2, SRF and Tead (data provided in **Supplementary Table 3c**). Ranking by AUC values, the top 10 GM-pairs from that screen were:

Gene set	Boolean Motif combination	FDR Q-value (*Q*)	AUC
CONTRACTILE FIBER_up	OR(MRF,MEF2)	0	0.864706
CONTRACTILE FIBER_up	AND(MRF,MEF2)	0	0.856747
CONTRACTILE FIBER_up	COMPOUND(MRF AND (MEF2MEF2 OR SRF))	0.000028	0.846021
CONTRACTILE FIBER_up	OR(MRF,MEF2,SRF)	0.000028	0.842907
CONTRACTILE FIBER_up	COMPOUND(MRF AND (MEF2MEF2 OR Tead))	0.000037	0.8391
CONTRACTILE FIBER_up	COMPOUND(MEF2 OR (MRF AND SRF))	0.000033	0.828893
MUSCLE DEVELOPMENT_up	COMPOUND(MEF2 OR (MRF AND SRF))	0	0.828668
CONTRACTILE FIBER_up	COMPOUND(MEF2 AND (MRF OR SRF))	0.000077	0.828374
sarcomere_up	AND(MRF,MEF2)	0	0.821739
CONTRACTILE FIBER_up	COMPOUND(MEF2 AND (MRF OR Tead))	0.000035	0.819377

For experimental validation, we chose to examine simple Boolean motif combinations instead of compound Boolean combinations, because simple Boolean motif combinations would be easier to test in subsequent construction and analysis of synthetic CRMs. We also expected an AND motif combination to confer greater specificity of gene expression regulation than an OR motif combination. The MRF AND MEF2 motif combination for

the sarcomere_up gene set (FDR 0, AUC 0.822) scored slightly less well than the MRF AND MEF2 motif combination for the CONTRACTILE FIBER_up gene set (FDR *Q*-value = 0, AUC 0.857). One of our positive control CRMs was for the gene *ACTA1*, which belongs to the sarcomere_up gene set and not to the CONTRACTILE FIBER_up gene set, and we were interested to see if there might be more than 1 functional CRM per gene at a given time point in a given cell type. It would be interesting to see if the predicted CRMs containing the MRF AND MEF2 motif combination for the CONTRACTILE FIBER_up gene set work with just as high a success rate.

To try to estimate what the anticipated CRM success rate might be for a given gene set, consider the following example. The figure below shows the degree to which all 46 sarcomeric genes are enriched for the MRF AND MEF2 TF binding site motif combination, as compared to a background set of 1840 (= 46*40) length-matched background genes that were not observed to be up- or down-regulated in this cell-type:



Sensitivity and specificity of MRF AND MEF2 for sarcomeric genes.

In looking at this figure, we see that at a given PhylCRM score threshold where 20% of background genes have a positive hit (i.e., a maximum-scoring window that we predict as

being a CRM) somewhere within their 75 kb regions around transcription start (80% specificity), 70% of sarcomeric genes (foreground) have such a positive hit within their 75 kb regions (i.e., 70% sensitivity). We note that sensitivity values for any given specificity can immediately be read off of the ROC curve, although for simplicity we use the 80% specificity / 70% sensitivity point for the following discussion. At this threshold, we can compile the following table of summary statistics indicating the fraction of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and also the positive predictive value (PPV) and misclassification error:

Estimation of summary statistics at a given score cutoff			
	Predicted Positive	Predicted Negative	
Foreground = 46	TP = 32	FN = 14	Sensitivity = TP/(TP + FN) = 70%
Background = 1840	FP = 368	TN = 1472	Specificity = TN/(FP + TN) = 80%
	PPV = TP/(TP + FP) = 8%		Misclassification error = (FP + FN)/(TP + FP + TN + FN) = 20%

Using the cutoff mentioned above, 20% of the background genes have a positive PhylCRM hit (i.e., predicted CRM) somewhere within 75 kb of transcription start, and 30% of the foreground genes do not have a predicted CRM, giving a misclassification error of 20% and positive predictive value (PPV) of 8%. We see three possible explanations for these results. First, some background genes containing a PhylCRM hit might be located close to a gene that is expressed in muscle and regulated by MRF AND MEF2; such PhylCRM hits would correspond to *bona fide* myogenic CRMs that were incorrectly placed into the background. Second, many of these PhylCRM hits might

represent CRMs that are targeted by TFs binding to the MRF AND MEF2 motifs but that do not drive expression in muscle. For example, MEF2 is known to regulate gene expression in the brain, and there are several bHLH TFs that are crucial for neuronal cell fate specification and are likely to have a binding site motif similar to the MRF motif bound by the myogenic bHLH TFs (MyoD, myogenin); thus, many of these hits could be true CRMs that drive expression in the brain rather than the muscle. Finally, it is possible that many of the PhylCRM hits are simply false predictions and are not actually CRMs. We have given this issue extensive thought, and we do not presently see a reliable means of estimating what fraction of MRF AND MEF2 hits adjacent to background genes fall into each of these three potential classes. We expect that prioritizing for experimental testing those significant PhylCRM hits that contain MRF AND MEF2 motifs and that are directly adjacent to sarcomeric genes, will lead to a greatly increased success rate in experimental validation of predicted CRMs functional in myogenic differentiation. In general, we believe that the results of Lever can be used to prioritize predicted CRMs for experimental testing, by picking for testing those candidate CRMs which lie next to genes that belong to significant scoring GM-pairs.

G. Position Weight Matrices utilized in this study:

We obtained from the supplementary data of Wasserman *et al.*[9] DNA binding site sequences corresponding to these 4 motifs from the supplementary data of that study, although we added additional myogenic MEF2 sites obtained from a SELEX experiment[20].

MRF:

$$\begin{bmatrix} A \\ C \\ G \\ T \end{bmatrix} \begin{bmatrix} 24 & 17 & 0 & 39 & 0 & 3 & 0 & 0 \\ 2 & 4 & 39 & 0 & 5 & 13 & 0 & 1 \\ 12 & 13 & 0 & 0 & 34 & 14 & 0 & 38 \\ 1 & 5 & 0 & 0 & 0 & 9 & 39 & 0 \end{bmatrix}$$

MEF2:

$$\begin{bmatrix} A \\ C \\ G \\ T \end{bmatrix} \begin{bmatrix} 6 & 3 & 107 & 73 & 113 & 117 & 114 & 1 & 125 & 17 \\ 97 & 9 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 \\ 4 & 1 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 103 \\ 18 & 112 & 18 & 52 & 10 & 8 & 10 & 123 & 0 & 3 \end{bmatrix}$$

SRF:

$$\begin{bmatrix} A \\ C \\ G \\ T \end{bmatrix} \begin{bmatrix} 0 & 0 & 13 & 14 & 10 & 4 & 14 & 8 & 6 & 0 \\ 20 & 20 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 2 & 0 & 14 & 18 \\ 0 & 0 & 7 & 6 & 6 & 16 & 4 & 11 & 0 & 2 \end{bmatrix}$$

Tead:

$$\begin{bmatrix} A \\ C \\ G \\ T \end{bmatrix} \begin{bmatrix} 4 & 9 & 0 & 12 & 0 & 0 & 0 & 0 & 4 & 0 & 1 \\ 6 & 0 & 12 & 0 & 0 & 0 & 12 & 11 & 1 & 6 & 4 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 6 \\ 1 & 0 & 0 & 0 & 12 & 12 & 0 & 1 & 7 & 0 & 1 \end{bmatrix}$$

H. Detailed experimental protocols, including primer sequences

Cell culture

Adult human skeletal myoblasts (Cambrex) were grown in SkGM2 medium (Cambrex) for optimal growth and differentiation potential. Myogenic differentiation was stimulated by switching the culture medium to DMEM with 2% horse serum (Sigma) when the cells reached about 70% confluence. All time points referred to in this study are with respect to the time of switching to differentiation medium. Mouse C2C12 cells (ATCC), mouse 3T3 cells (ATCC), and human lens epithelial cells (gift from Amy Donner) were cultured in DMEM (Invitrogen) with 10% fetal bovine serum (Sigma), respectively. HEK293T cells were a gift from Karen Cichowski.

RNA purification

Total RNA was isolated from primary human skeletal muscle cells using TRIzol reagent (Invitrogen) according to the manufacturer's protocols. For microarray experiments, total RNA was further purified with RNeasy columns (Qiagen).

Gene expression profiling microarray experiments

Microarrays were synthesized and hybridized by the Harvard Partners Center for Genetics and Genomics. Briefly, each glass slide was spotted with the Human OligoLibrary^{TM} Release 1.0 that was designed by Compugen, Inc. and manufactured by Sigma-Genosys, Inc. This oligonucleotide library consists of 18,864 60-mers representing 18,664 unique genes. We extracted mRNA at 6 time points (-48 hrs, -24 hrs, 0 hrs, 12 hrs, 24 hrs, and 48 hrs relative to serum withdrawal). These time points were selected

since prior studies in a related cell type (mouse C2C12 cells) demonstrated their effectiveness for capturing key transcriptional events during myogenic differentiation[21,22]. For each time point, four hybridizations, consisting of duplicate hybridizations with Cy3 and Cy5 dye-reversal, were performed essentially as described previously[23].

Preprocessing and clustering of gene expression microarray data

Scanned TIF images were quantified with GenePix software (Axon Instruments). For each feature, the median pixel intensity of the local background was subtracted from the spot's median pixel intensity. We then applied variance stabilizing normalization[24] to normalize all single channels to each other. False discovery rates (FDRs) were calculated using Significance Analysis of Microarrays[25] (one class time-series and slope parameters) on the four replicate arrays. The arcsinh values of the four replicate arrays for each time point were then combined by taking the arithmetic mean and expressed as the fold-change relative to the first time point (-24 hrs). Changes in arcsinh values correspond to the following approximate ratios (arcsinh = linear): 0 = 1/1; 1 = 2.7/1; 2 = 7.5/1; 3 = 20/1, 4 = 55/1; 5 = 155/1, 6 = 405/1. Genes that were differentially expressed at a 5% FDR were clustered using *k*-means clustering by de Hoon's Cluster 3.0 software[26] (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv). Our choice of 14 clusters was determined empirically.

Western blotting

Cell nuclei extracts and cytoplasmic extracts were obtained from human skeletal myoblasts at -48, -24, 0, +24, and +48 hours with respect to stimulation of differentiation,

according to standard protocols. Equal protein amounts were subjected to standard SDS-PAGE. Western Blots were performed using SuperSignal West Femto Maximum Sensitivity Substrate (Pierce) according to the manufacturer's instructions. Blocking solution consisted of 5% nonfat dry milk in TBS-T (Tris Buffered Saline with 0.1% Tween 20) and washing solution was TBS-T.

The following antibodies used in Western blots were purchased from Santa Cruz: Myf5 (sc-302), MyoD (sc-760), Myogenin (sc-576), Myf6 (sc-784), SRF (sc-335), MEF2C (sc-13266), MEF2 (sc-10794), MEF2A (sc-313), and lamin B1 (sc-20682). Tead1 (or Tef-1) antibody was purchased from BD Biosciences Pharmingen (610923). All antibodies were probed at a 1:1,000 dilution in blocking solution, except for the lamin B1 and MEF2C antibodies which were probed at a 1:2,000 dilution. Anti-rabbit and anti-mouse HRP-conjugated secondary antibodies (as supplied by Pierce) were diluted 1:3,000 in blocking solution. Anti-goat secondary antibodies (Sigma) were diluted 1:300,000.

The Tead or Tef family of transcription factors are comprised of at least four mammalian members, Tead1 (TEF-1), Tead2 (TEF-4), Tead3 (TEF-5), and Tead4 (TEF-3)[27]. Tead4 and Tead2 are the only two members detectable in regenerating mouse skeletal muscle[27,28]. Tead1 is broadly expressed in many different embryonic tissues[29], but Tead1 knockout mice have severe cardiac defects suggesting a major role in cardiac development[30]. Tead3 is detectable in skeletal and cardiac muscle but is preferentially expressed in the developing placenta[31,32]. Since the immunogen used to develop the BD Pharmingen is 53% identical and 66% similar to Tead4 protein, it is possible that the

antibody is cross-reactive with Tead4 or other Tead family members using a sensitive Western blot detection system. At the time of submission of this paper, it was believed that Tef1 was the relevant Tead family member for myogenic differentiation[11], and BD Biosciences Pharmingen had no data for or against the cross-reactivity of their Tead1 antibody.

ChIP

Chromatin immunoprecipitations were carried out using a modified version of the Farnham protocol (http://mcardle.oncology.wisc.edu/Farnham/protocols/chips.html). 5 x 10^8 cells fixed at days 0, 1, and 2 of differentiation.

Cells were fixed with 1% formaldehyde at room temperature for 10 minutes with occasional agitation of the plates. 2.5 M glycine was added to the cell media for 5 minutes to stop the crosslinking reaction. The cells were then washed twice with ice-cold PBS and incubated in PBS with 20% trypsin-EDTA (Cambrex) for 10 min at 37ºC. 0.5 ml of FCS was added to inhibit trypsinization. The cells were then scraped and collected into 50-ml conical tubes and kept on ice. Cells were washed once with ice-cold PBS with PMSF (Sigma, 100 μM) and protease inhibitors (20 μl per ml, Sigma P8340), flash frozen in ethanol/dry ice, and kept @-70ºC until chromatin immunoprecipitation.

Frozen cells were thawed on ice, resuspended in ice-cold cell lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP40, 1:50 protease inhibitor mix [Sigma catalog # P8340]), and incubated on ice for 10 minutes. Nuclei release was aided by dounce

homogenization. Nuclei were pelleted by centrifugation and resuspended in room temperature nuclei lysis buffer (50 mM Tris-Cl pH 8.1, 10 mM EDTA, 1% SDS, 1:50 protease inhibitor mix), followed by incubation on ice for 10 minutes. The nuclei were then sonicated to achieve chromatin fragments with an average length of 1,000 bp. The sonication conditions used were 9 sonications of 15-second pulses separated by 1-minute incubation on ice. Samples were centrifuged at high speed to remove cellular debris. The supernatant containing the sonicated chromatin was transferred to a 50-ml conical tube and diluted 1:10 with ice-cold dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-Cl pH 8.1, 167 mM NaCl, 1:50 protease inhibitor mix). Chromatin was precleared by adding 50 µl of Protein A beads/Salmon Sperm DNA (Upstate Protein A/Salmon Sperm DNA, cat# 16-157) per ml and incubating on a rotating platform at 4ºC. 3 ug of antibody was used for each immunoprecipitation. The following antibodies were purchased from Santa Cruz: MyoD (sc-760), Myogenin (sc-576), SRF (sc-335), and MEF2 (sc-10794). 60 µl of Protein A/salmon sperm DNA beads were added to each sample and incubated on a rotating platform at 4°C for 1-2 hours. Samples were then microcentrifuged for 1 min and placed into fresh microcentrifuge tubes.

Immunoprecipitates were washed twice with ice-cold wash buffer 1 (20 mM Tris, pH 8.1, 150 mM NaCl, 2 mM EDTA, 0.1 % SDS, 1% Triton X-100), once with wash buffer 2 (20 mM Tris, pH 8.1, 500 mM NaCl, 2 mM EDTA, 0.1 % SDS, 1% Triton X-100), once with wash buffer 3 (10 mM Tris, pH 8.1, 250 mM LiCl, 2 mM EDTA, 1% NP-40, 1% deoxycholate), and once with ice-cold 4 M LiCl/TE. After the last wash and spin, all remaining buffer was carefully removed with a sterile 1-ml pipette.

Antibody/protein/DNA complexes were eluted by adding 100 µl of IP elution buffer 1 (1% SDS, 1 mM EDTA, 10 mM Tris, pH 8.1) and incubated @65°C for 15 min. Samples were microcentrifuged for 3 minutes. Supernatants were then transferred to fresh microcentrifuge tubes. Samples were then eluted again with 150 µl of elution buffer 2 and incubated at 65°C for 15 min. Samples were then combined and incubated overnight at 65°C to reverse formaldehyde crosslinks.

To each tube, 250 µl TE and 5 µl of proteinase K (20 mg/ml) were added. The tubes were then incubated at 37ºC for 1 hour. To each tube, 55 µl of 4M LiCl was added. The samples were then extracted twice with 500 µl phenol/chloroform/isoamyl alcohol and once with 500 µl of chloroform. Then, 1 µl (10 mg) of glycogen to each sample and the samples were ethanol precipitated. After drying the pellets, the samples were resuspended in 150 ul of 10 mM Tris 8.5. Each IP was performed in triplicate for each individual chromatin sample.

In our ChIP assays, as positive controls we examined five previously described muscle CRMs, and as negative controls we examined two noncoding regions with no significant matches and eight noncoding regions with only a single significant match, to any of these five motifs. The positive control regions were as follows:

CAV3 (0.2 kb upstream of transcriptional start):

- myotube specific promoter; previously confirmed myogenin (MYF) binding site in mouse C2C12 cells[33]

COX6A2 (0.3 kb upstream of transcriptional start):

- myotube specific promoter; previously confirmed MRF (E-box) and MEF2 binding sites in mouse Sol8 and C2C12 cells[34]

ACTA1 (0.3 kb upstream of transcriptional start):

- promoter region
- 3 previously confirmed SRF sites in primary chicken muscle culture[35]
- previously confirmed Tead1 site in rat cardiomyocytes[36]

TNNT2 (0.1 kb upstream of transcriptional start):

- conserved Tead1 (M-CAT) site in chicken promoter was previously shown to be important for chicken skeletal muscle[37]
- MEF2 site was previously shown to be important for rat cardiac muscle expression[38]
- CArG boxes (SRF sites) were previously confirmed by footprinting in rat cardiomyocytes[38]

DMD (6.4 kb into 1st introns):

- myotube-specific enhancer
- three MRF sites and one MEF2 site required for activation in myotubes[39]

Primer sequences:

Gene Name	Forward Primer	Reverse Primer
ChIP primers		
ACTA1	ACCCTCGCCCCACCCCATCC	GGCCGCTTGTCCCTCTGCTC
BDKRB2	GCCCGGGCTCTTGCTCCAG	CTCCTCAGGGCCTCAGTTTCTTCAT
CAV3	GCCCTCTGCACCCTCTCCTG	CCGGCTGGGGCTGAAAATAC
CLC	TCCAGGGGGCAAATGAGGGTAAT	CATAAGAGACTGGGCGCGGTGGTTC
COX6A2	GCCTGTAATCCCAGCACTGT	AGCTGTTGTCCTGTGCCTCT
CPM	TGTGCCACGTGTCCTTTCATCATCAGTA	GCACCCAAATCCCCATCTCAGTCC
CSRP3	GTGGGGGCCTGGAGAAATGAT	AGCCACAGAACCAACCCACCTC
DMD	CTGCGACAAAATGGGCACTCAATA	CTGCGACAAAATGGGCACTCAATA
GAP43	CTGAGGCGGGGAGAGGAGAG	TGGGAAGTGGTTATTATGGGATTG
HBZ	GGCCTTGTCTGTCTTTTCCTCCATA	GGCAGCTCAGCACCCATCCT
HSPB3	GGACTAGTGCCTTCAACAGC	TAAAACAACGTGGGGGAGTA
EDG5	CTAGCCCATGTCCCCTCCCTGTGTAA	TCCCCCTGGCTGCTTGGTAGAGAAT
KRT2A	GCCCCTCACCGCCCTCTCCT	ATTATGCGCCTTGTCGATGCTCTC
MEF2C	AGGGCAGTCATGGAGAGGTC	TTATGGCAAGGGAGAACTGG
MGLL	CAAGGGGGATGGCACTAAACC	CTCCTACAGCCTGCGATGAAAAG
MTP	TTGGGTACTATCGGTGGAGA	GTGGGCAGAAAGGAGTTGAG
PTHR1	GGGGGTCCAAAGCGGGTCCTGTT	TCCTGGCCCCCTCCTCCCTTCAAA
TNNT2	TCTTTACCCCCAGCATCAGT	GGGACAAGGCTACAGGAACA
TOP2A	AAGTCTGCCCCACGGTCCTGA	CTCTGGGCCCTGCTTGCTCTTC
RT-PCR primers		
DMD	GCGCCTCCTAGACCTCCTC	ACCCGCAGTGCCTTGTTG
ACTA1	GCCCGAGCCGAGAGTAGCAGTTGT	CTCGCGGTTGGCCTTGGGATTG
COX6A2	CCAAAGGAGGCCACGGAGGAGCAG	GGTGGCCCGAGTGGAGATAGGAGTTGA
CAV3	TTGACCTGGTGAACCGAGAC	CGTGGACAACAGACGGTAGC
TNNT2	CTGAGCGGGAAAAGAAGAAGAAGATT	GTGGGGGCAGGCAGGAGTG
MYOD	AGCACTACAGCGGCGACT	GCGACTCAGAAGGCACGTC
MYOG	TAAGGTGTGTAAGAGGAAGTCG	CCACAGACACATCTTCCACTGT
MEF2C	CTCCCAGTCGGCTCAGTCATTG	CGAAGGGGTGGTGGTACGGTCTCTA
MEF2D	AAGCGGAAGTTTGGCCTGATGAAGAA	GCCGCTGGGATTGCTGAACTGC
SRF	ACTGCCTTCAGTAGGAACAA	TTCAAGCACACACACTCACT
TEAD4	TGTGGCAGGCGCAAAATCATCC	GTCCGGGTCCTGCTGCTGCTC
HSPB3	GGGGCTCGCCACTGACTGAA	AGACTGCGCTGCCCTGGTTTT

CSRP3	CTCTTCCCACAGATGGCACA	GAGAAGGTTATGGGAGGTGGC
CACNG1	ATGTCCCAGACCAAAATGCTG	CAGGTAGTGTTGTGGTGCTC
PDLIM3	ACTCCCTCCGGGATTGACTG	AGCTTAGCCGCAACTTTCAAG
ARGBP2	AACACAGGGCGTGATTCTCAG	TGGTCGAACGCTTCTAAAACC
RPS18	GATGGGCGGCGGAAAATAG	GCGTGGATTCTGCATAATGGT
Cloning primers		
DMD_BAM	CACCGGATCCCACGGCCATACAACCTCTACCTC	GGATCCTTCATCTCCACTGTCCCCATTCTA
PDLIM3_BAM	CACCGGATCCCTACCCGCCAGTGCTGTGTTGAG	GGATCCGGGAAGGCCTGGGGGAGAAG
MGLL_BAML	ACGCGGATCCCAAGGGGGATGGCACTAAACC	ACGCGGATCCCTCCTACAGCCTGCGATGAAAAG
CSRP3_BAM	ACGCGGATCCGTGGGGGCCTGGAGAAATGAT	ACGCGGATCCAGCCACAGAACCAACCCACCTC
Primers for cloning into pLKO.1 vector (RNAi)		
MYOG_shRNA_1	CCGGGCCCACAATCTGCACTCCCTTCTCGAGAAG GGAGTGCAGATTGTGGGCTTTTTG	AATTCAAAAAGCCCACAATCTGCACTCCCTTCTCG AGAAGGGAGTGCAGATTGTGGGC
MYOG_shRNA_2	CCGGGCACATCTGTTCTAGTCTCTTCTCGAGAAG AGACTAGAACAGATGTGCTTTTTG	AATTCAAAAAGCACATCTGTTCTAGTCTCTTCTCG AGAAGAGACTAGAACAGATGTGC
MYOG_shRNA_3	CCGGCCCAGACGAAACCATGCCCAACTCGAGTTG GGCATGGTTTCGTCTGGGTTTTTG	AATTCAAAAACCCAGACGAAACCATGCCCAACTC GAGTTGGGCATGGTTTCGTCTGGG
MEF2D_shRNA_1	CCGGCCCTGGTGACATCATCCCTTACTCGAGTAA GGGATGATGTCACCAGGGTTTTTG	AATTCAAAAACCCTGGTGACATCATCCCTTACTCG AGTAAGGGATGATGTCACCAGGG
MEF2D_shRNA_2	CCGGCAATGGCAACAGCCTAAACAACTCGAGTT GTTTAGGCTGTTGCCATTGTTTTTG	AATTCAAAAACAATGGCAACAGCCTAAACAACTC GAGTTGTTTAGGCTGTTGCCATTG
MEF2D_shRNA_3	CCGGCACATCAGCATCAAGTCAGAACTCGAGTTC TGACTTGATGCTGATGTGTTTTTG	AATTCAAAAACACATCAGCATCAAGTCAGAACTC GAGTTCTGACTTGATGCTGATGTG
SRF_3882F	CCGGCCCTTGGTGTATCCCTAATTACTCGAGTAA TTAGGGATACACCAAGGGTTTTTG	AATTCAAAAACCCTTGGTGTATCCCTAATTACTCG AGTAATTAGGGATACACCAAGGG
SRF_2110F	CCGGGCTCAATTTGCTATGAGTATTCTCGAGAAT ACTCATAGCAAATTGAGCTTTTTG	AATTCAAAAAGCTCAATTTGCTATGAGTATTCTCG AGAATACTCATAGCAAATTGAGC
SRF_2934F	CCGGGAGAGGAGATTGATGTCCTTTCTCGAGAAA GGACATCAATCTCCTCTCTTTTTG	AATTCAAAAAGAGAGGAGATTGATGTCCTTTCTCG AGAAAGGACATCAATCTCCTCTC
HNF4alpha	CCGGCCGACAATGTGTGGTAGACAACTCGAGTTG TCTACCACACATTGTCGGTTTTTG	AATTCAAAAACCGACAATGTGTGGTAGACAACTC GAGTTGTCTACCACACATTGTCGG

Quantitative RT-PCR

Total RNA was reverse-transcribed using SuperScript III (Invitrogen) according to the manufacturer's protocols. Quantitative PCRs were performed using iQ™ SYBR® Green Supermix (BioRad) and 0.2 μM primers with an iCycler iQ Real-Time PCR Detection System (BioRad).

Quantitative ChIP-PCR

ChIPs were performed in biological triplicate using a modified version of the Farnham protocol[40]. The following antibodies were used in ChIPs: MyoD (sc-760), myogenin (sc-576), SRF (sc-335), and MEF2 (sc-10794), all from Santa Cruz Biotechnology, Inc. We included SRF since we observed that several of our predicted CRMs contained SRF motif matches. Tead was not included since a suitable antibody was not available. As positive controls, we examined five previously described muscle CRMs. Negative control genomic regions were chosen based on their not having any significant PhylCRM hits when considering the MRF, MEF2, SRF, or Tead motifs, and their being adjacent to genes called "present" in the expression microarray data but not up- or down-regulated at a FDR less than 0.1. Quantitative ChIP-PCRs were performed essentially as described above, except using 6 μl of immunoprecipitated DNA.

Luciferase reporter assays

Putative and control CRMs were cloned either upstream (BglII) or downstream (BamHI) of the luciferase reporter gene into pGL3-Promoter vector (Promega) in their native genomic orientation (i.e., upstream versus downstream of transcription, Watson versus Crick strand). As a positive control, we used one of the five previously known muscle

CRMs used in our ChIPs. A negative control human noncoding genomic region not enriched for matches to these four motifs was indistinguishable from the corresponding enhancer-less empty vector negative control. C2C12 cells were cultured in 6-well plates (9.4 cm^2 per well) 24 hours prior to transfection at 3 x 10^4 cells per well for myoblasts or 1.5 x 10^5 cells per well for myotubes. The cells were then co-transfected in triplicate with 1 μg of experimental vector (pGL3-P with or without inserted region) and 50 ng of the normalization vector (pRL-TK) using FuGENE 6 transfection reagent (Roche) according to the manufacturer's protocols. Cell extracts were obtained from an aliquot of the proliferating myoblasts 24 hours after transfection. The remaining cell cultures were then switched to differentiation medium, and cell extracts were obtained after 96 hours in differentiation medium. Luciferase reporter assays were performed using the Dual-Luciferase® Reporter Assay System (Promega) according to the manufacturer's protocols. Firefly luminescence intensities were normalized by the luminescence intensities of the internal *Renilla* control. We used C2C12 cells in these assays instead of primary adult human skeletal myoblasts because the primary cells failed to differentiate robustly after transfection.

shRNA knockdowns

Short hairpin RNA (shRNA) constructs directed against mouse RNA transcripts were generated essentially as described previously[41]. Lentiviral reagents were kindly provided by Karen Cichowski. For lentiviral production, HEK293T cells were transfected with the Δ8.2 lentiviral construct (encoding *gag, pol, rev)*, VSVG, and either empty pLKO.1 vector or the pLKO.1 vector containing a sequence for a shRNA specific for each of the

muscle genes *MYOD*, *MYOG*, *MEF2D*, *SRF*, and the liver gene *HNF4α.* Three distinct shRNA constructs were created for each gene in order to control for off-targets effects. Lentivirus was titered by serial dilution followed by colony formation assays in medium containing puromycin. C2C12 cells (7×10^4) were plated on 100-mm plates 24 hours prior to infection. After infection at 5 multiplicities of infection of lentivirus, C2C12 cells were grown in growth media for 24 hours and selected in puromycin for 72 hours. Luciferase reporter assays were then performed as described above, except cells were plated onto 12-well plates and transfected with proportionately less of the reagents. Our MEF2C knockdowns resulted in extensive cell death, and thus could not be utilized here.

Creation of synthetic CRMs

To test the sufficiency of the inferred MRF AND MEF2 *cis* regulatory code for myogenic differentiation, we created a synthetic CRM containing consensus MRF and MEF2 binding sites arranged as in our newly discovered *ACTA1* CRM, but in the context of the *MGLL* negative control flanking sequence. The MGLL negative control region was selected as a template into which to place TF binding sites in order to experimentally test the MRF AND MEF2 *cis* regulatory code for myogenic differentiation. To create synthetic CRMs, we created variants of a shorter 167-bp *MGLL* negative control region by ligating segments of the original *MGLL* region or by ligating modified segments of the *MGLL* region such that the new construct would have two consensus MRF sites and one consensus MEF2 site. The reconstituted *MGLL* region served as a negative control. As positive controls, we used an SV40 enhancer, one of the five previously known muscle CRMs used in our ChIPs (DMD), and a novel CRM that we verified previously CRM

(ACTA1, see **Fig. 4**). The TF binding sites were placed in the modified *MGLL* region such that they mimicked the position and orientation of our newly discovered *ACTA1* CRM. The sense (F) and antisense ® strand of each segment were synthesized as single-stranded DNA oligonucleotides and were then annealed to form double-stranded DNA. The following oligonucleotides were used in the annealing reactions:

MGLL_SEG3_CAGCTG_R	GATCTCAGCTGCAGCCTGCGATGAAAAGGTAAGGGTTGCAATT
MGLL_SEG1_F	CCATGATGCATTCACCTCCCACCAGGCCCCACCTTCAACATTG CAAGTAA
	GGGATTACAGTTCAAAATGAGG
MGLL_SEG1_R	ATTTTGAACTGTAATCCCCAATGTTGAAGGTGGGGCCTGGTGG
	GAGGTGAATGCATCATGGAGCT
MGLL_SEG2_F	TTTGGTGGGGACACAGATCCAAACCATATCAACTTGTAGGGGC
	AGAAAGACGTCACCTTTAC
MGLL_SEG2_R	AGGTGACGTCTTTCTGCCCCTACAAGTTGATATGGTTTGGATC
	TGTGTCCCCACCAAACCTC
MGLL_SEG3_F	TTGAATTGCAACCCTTACCTTTTCATCGCAGGCTGTAGGAGA
MGLL_SEG3_R	GATCTCTCCTACAGCCTGCGATGAAAAGGTAAGGGTTGCAATT
	CAAGTAA
MGLL_SEG1_CAGCTG_F	CCATGATGCATTCACCTCCCACCAGGCCCCACCTTCAACATTG
	GGGCAGCTGGTTCAAAATGAGG
MGLL_SEG1_CAGCTG_R	ATTTTGAACCAGCTGCCCCAATGTTGAAGGTGGGGCCTGGTGG
	GAGGTGAATGCATCATGGAGCT
MGLL_SEG2_ACTA1_PMEF2_F	TTTGGTGGGGACACAGATCCAAACCATATCAACTTGTAGGGGC
	AGAACTAAAAATAGTTTAC
MGLL_SEG2_ACTA1_PMEF2_R	ACTATTTTTAGTTCTGCCCCTACAAGTTGATATGGTTTGGATC
	TGTGTCCCCACCAAACCTC
MGLL_SEG3_CAGCTG_F	TTGAATTGCAACCCTTACCTTTTCATCGCAGGCTGCAGCTGA

Segment 1 was designed to have a *SacI*-compatible end and segment 3 a *NheI*-compatible end such that an entire Seg1-Seg2-Seg3 sequence could be ligated into a pGL3-P vector that was previously digested with *NheI* and *SacI* and treated with alkaline phosphatase. The short MGLL sequence was reconstituted by ligating the following double-stranded segments: MGLL_SEG1, MGLL_SEG2, and MGLL_SEG3. The MGLL region with two MRF sites and one MEF2 site was created by ligating MGLL_SEG1_CAGCTG, MGLL_SEG2_ACTA_PMEF2, and MGLL_SEG3_CAGCTG.

References

1. Philippakis, A.A. et al. Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Computational Biology* 2(2006).
2. Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics* 16, 16-23 (2000).
3. Halpern, A.L. & Bruno, W.J. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* 15, 910-7 (1998).
4. Moses, A.M., Chiang, D.Y., Kellis, M., Lander, E.S. & Eisen, M.B. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol* 3, 19 (2003).
5. 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).
6. Holmquist, R., Cantor, C. & Jukes, T. Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids. *J Mol Biol* 64, 145-61 (1972).
7. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22, 160-74 (1985).
8. Margulies, E.H. et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* 17, 760-74 (2007).
9. Wasserman, W., Palumbo, M., Thompson, W., Fickett, J. & Lawrence, C. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* 26, 225-8 (2000).
10. Toutenburg, H. *Statistical Analysis of Designed Experiments*, (Springer-Verlag, New York, 2002).
11. Wasserman, W. & Fickett, J. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278, 167-181 (1998).
12. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267-73 (2003).
13. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-50 (2005).
14. Good, P.I. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, (Springer, 2005).
15. Agarwal, P. & Graepel, T. Generalization Bounds for the Area under the ROC Curve. *Journal of Machine Learning Research* 6, 393-425 (2005).
16. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338-45 (2005).
17. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 296, 1205-14 (2000).

18. Berriz, G.F., King, O.D., Bryant, B., Sander, C. & Roth, F.P. Characterizing gene sets with FuncAssociate. *Bioinformatics* 19, 2502-4 (2003).
19. Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, P3 (2003).
20. Andres, V., Cervera, M. & Mahdavi, V. Determination of the consensus binding site for MEF2 expressed in muscle and brain reveals tissue-specific sequence constraints. *J. Biol. Chem.* 270, 23246-9 (1995).
21. Shen, X. et al. Genome-wide examination of myoblast cell cycle withdrawal during differentiation. *Dev Dyn* 226, 128-38 (2003).
22. Tomczak, K.K. et al. Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J.* 18, 403-5 (2004).
23. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-70 (1995).
24. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1, S96-104 (2002).
25. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98, 5116-21 (2001).
26. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* 20, 1453-4 (2004).
27. Zhao, P. et al. Fgfr4 is required for effective muscle regeneration in vivo. Delineation of a MyoD-Tead2-Fgfr4 transcriptional pathway. *J Biol Chem* 281, 429-38 (2006).
28. Yockey, C.E., Smith, G., Izumo, S. & Shimizu, N. cDNA cloning and characterization of murine transcriptional enhancer factor-1-related protein 1, a transcription factor that binds to the M-CAT motif. *J Biol Chem* 271, 3727-36 (1996).
29. Jacquemin, P., Hwang, J.J., Martial, J.A., Dolle, P. & Davidson, I. A novel family of developmentally regulated mammalian transcription factors containing the TEA/ATTS DNA binding domain. *J Biol Chem* 271, 21775-85 (1996).
30. Chen, Z., Friedrich, G.A. & Soriano, P. Transcriptional enhancer factor 1 disruption by a retroviral gene trap leads to heart defects and embryonic lethality in mice. *Genes Dev* 8, 2293-301 (1994).
31. Jacquemin, P., Martial, J.A. & Davidson, I. Human TEF-5 is preferentially expressed in placenta and binds to multiple functional elements of the human chorionic somatomammotropin-B gene enhancer. *J Biol Chem* 272, 12928-37 (1997).
32. Jacquemin, P. et al. Differential expression of the TEF family of transcription factors in the murine placenta and during differentiation of primary human trophoblasts in vitro. *Dev Dyn* 212, 423-36 (1998).
33. Biederer, C.H. et al. The basic helix-loop-helix transcription factors myogenin and Id2 mediate specific induction of caveolin-3 gene expression during embryonic development. *J. Biol. Chem.* 275, 26245-51 (2000).

34. Wan, B. & Moreadith, R.W. Structural characterization and regulatory element analysis of the heart isoform of cytochrome c oxidase VIa. *J. Biol. Chem.* 270, 26433-40 (1995).
35. Chow, K.L. & Schwartz, R.J. A combination of closely associated positive and negative cis-acting promoter elements regulates transcription of the skeletal alpha-actin gene. *Mol. Cell. Biol.* 10, 528-38 (1990).
36. MacLellan, W.R., Lee, T.C., Schwartz, R.J. & Schneider, M.D. Transforming growth factor-beta response elements of the skeletal alpha-actin gene. Combinatorial action of serum response factor, YY1, and the SV40 enhancer-binding protein, TEF-1. *J Biol Chem* 269, 16754-60 (1994).
37. Mar, J.H. & Ordahl, C.P. M-CAT binding factor, a novel trans-acting factor governing muscle-specific transcription. *Mol. Cell. Biol.* 10, 4271-83 (1990).
38. Wang, G., Yeh, H.I. & Lin, J.J. Characterization of cis-regulating elements and trans-activating factors of the rat cardiac troponin T gene. *J. Biol. Chem.* 269, 30595-603 (1994).
39. Marshall, P., Chartrand, N. & Worton, R.G. The mouse dystrophin enhancer is regulated by MyoD, E-box-binding factors, and by the serum response factor. *J Biol Chem* 276, 20719-26 (2001).
40. Boyd, K.E. & Farnham, P.J. Coexamination of site-specific transcription factor binding and promoter activity in living cells. *Mol Cell Biol* 19, 8393-9 (1999).
41. Stewart, S.A. et al. Lentivirus-delivered stable gene silencing by RNAi in primary cells. *RNA* 9, 493-501 (2003).