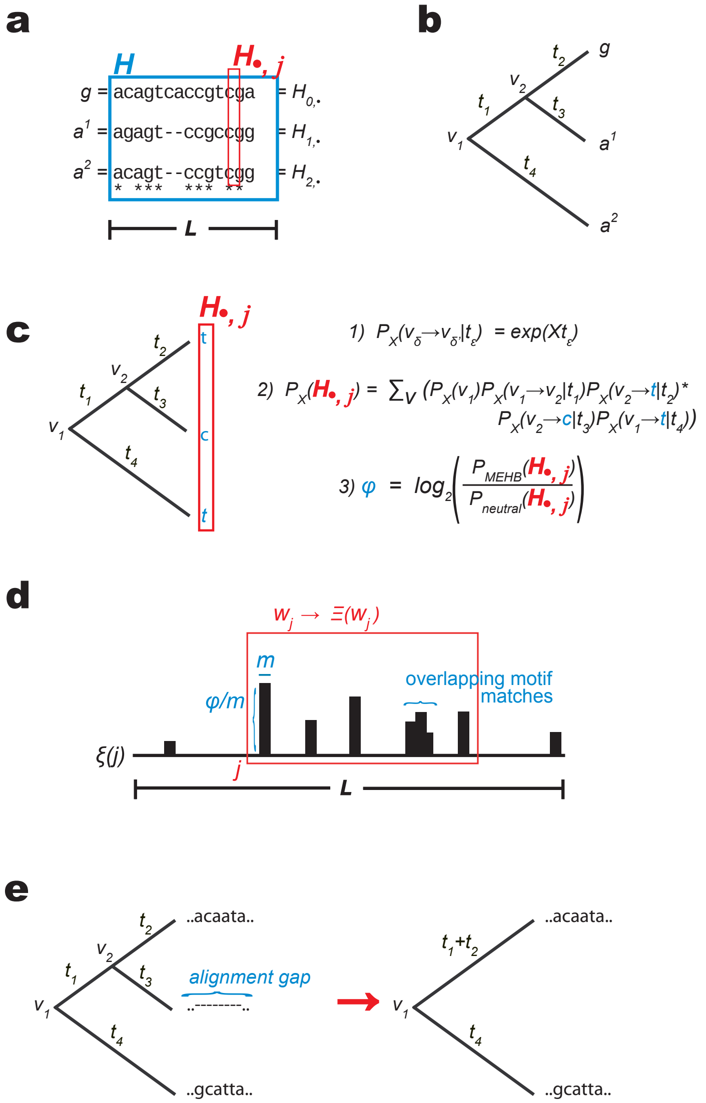
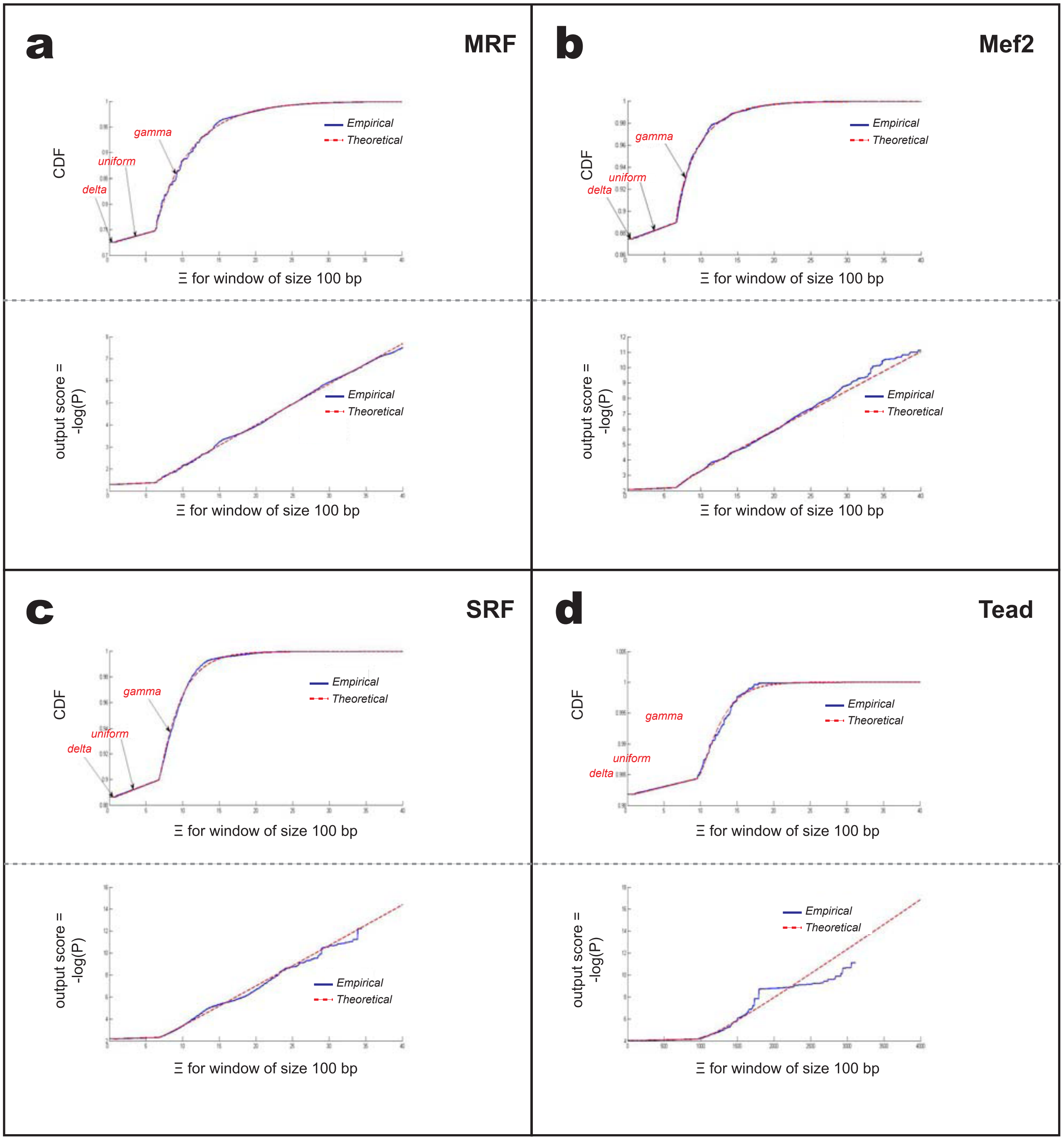
## Supplementary Figure 1



**Supplementary Figure 1: PhylCRM scoring scheme for a single motif**
(**a**) *g* represents the sequence being searched for CRMs and $a^1$ and $a^2$ are sequences from another organism aligned to it. *L* represents the length of the sequence, $H_{0,\bullet}$=*g*, $H_{i,\bullet}$ =$a^{(i)}$, and $H_{\bullet,j}$ denotes the alignment column at position *j*. (**b**) Tree indicating the phylogeny of *g*, $a^1$, and $a^2$. (**c**) Scoring motif matches using the MEHB model. Here, the probability that a given nucleotide *a* turns into *b* during time *t* is given by a matrix exponential, for a suitably chosen rate-matrix *R*. This probability is then used to compute the probability of observing the set of nucleotides $H_{\bullet,j}$ under both the MEHB rate-matrix and the neutral matrix. The score of the motif φ is then taken to be the log-likelihood of the ratio of these probabilities. (**d**) Graphical image of scores for a motif *M* along *g*, where the height of the bars is φ/*m*. These scores are stored in an array ξ and the score of a window $w_j$ (represented by Ξ($w_j$)) is then given by summing ξ in $w_j$. **(e)** When there is no alignable sequence at a given position (or if there is no motif match there), the branch containing that sequence is removed and the pruned tree is used to compute φ.

Supplementary Figure 2



**Supplementary Figure 2. Comparisons between the empirical and the fitted mixture of Delta, Uniform and Gamma distributions.**

(**a**) **Upper panel** shows empirical cumulative distribution function (CDF) for MRF (in blue) and the corresponding CDF for the fitted mixture model (in red).

(**a**) **Lower panel** shows empirical output score for MRF (in blue) and the corresponding output score for the fitted mixture model (in red).

(**b**) **Upper panel** shows empirical CDF for MEF2 (in blue) and the corresponding CDF for the fitted mixture model (in red).

(**b**) **Lower panel** shows empirical output score for MEF2 (in blue) and the corresponding output score for the fitted mixture model (in red);

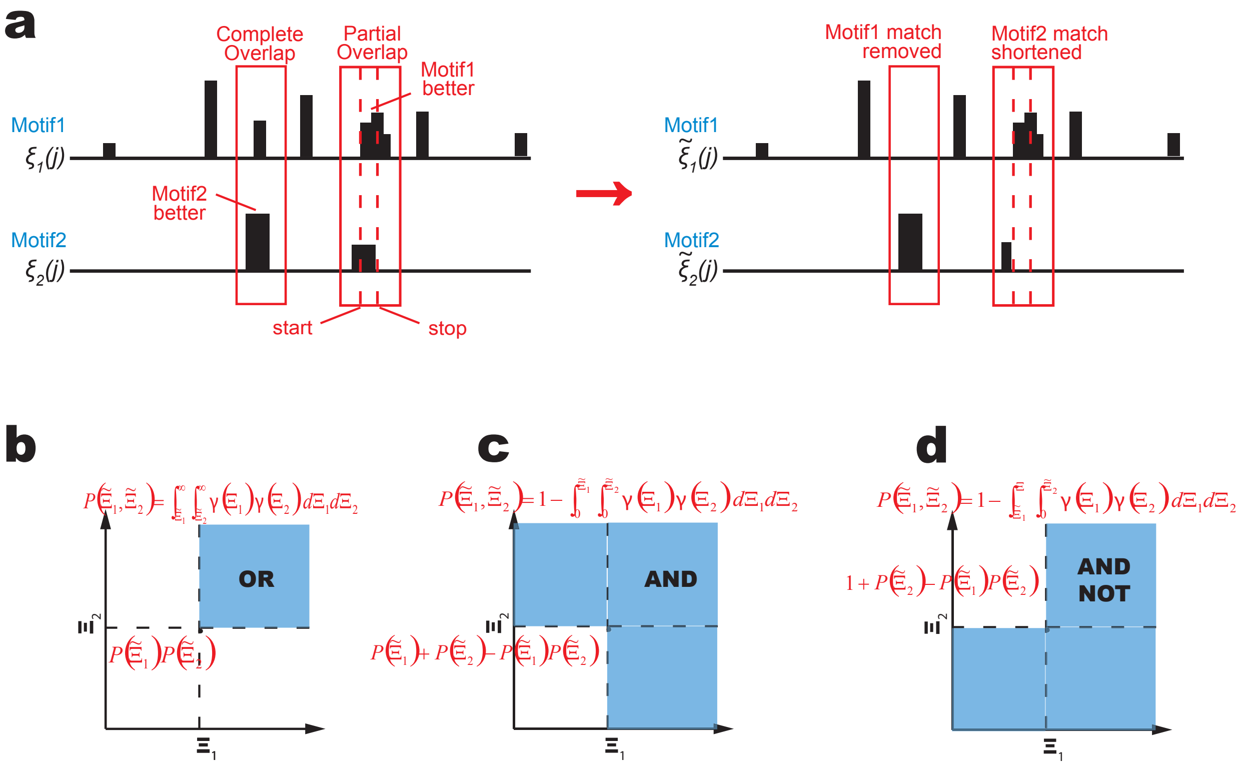
(**c**) **Upper panel** shows empirical CDF for SRF (in blue) and the corresponding CDF for the fitted mixture model (in red).

(**c**) **Lower panel** shows empirical output score for SRF (in blue) and the corresponding output score for the fitted mixture model (in red).

(**d**) **Upper panel** shows empirical CDF for TEAD (in blue) and the corresponding CDF for the fitted mixture model (in red).

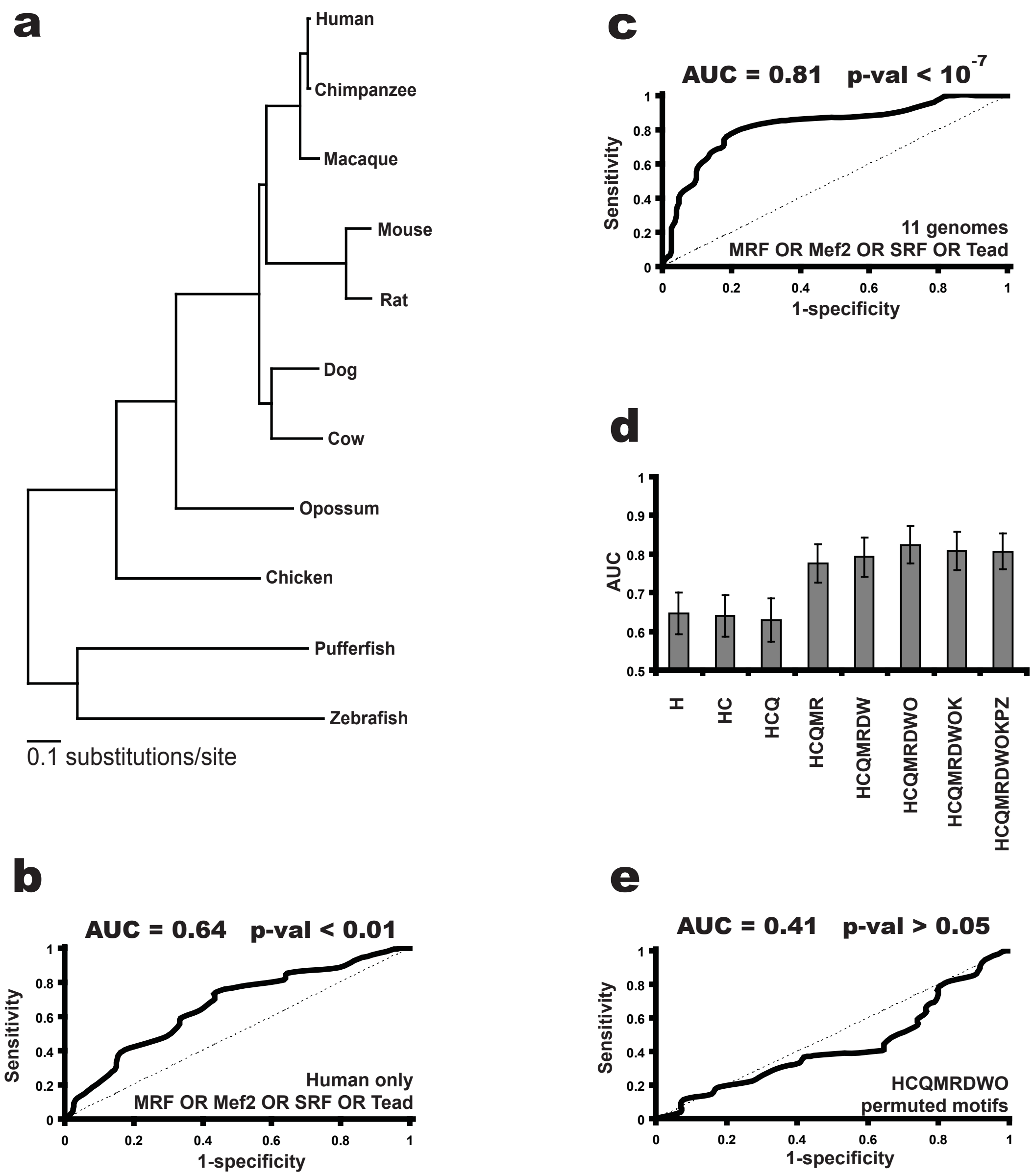
(**d**) **Lower panel** shows empirical output score for TEAD (in blue) and the corresponding output score for the fitted mixture model (in red).

## Supplementary Figure 3


a
Complete Overlap
Partial Overlap
Motif1 better
Motif1
$\xi_1(j)$
Motif2 better
Motif2
$\xi_2(j)$
start
stop
Motif1 match removed
Motif2 match shortened
Motif1
$\tilde{\xi}_1(j)$
Motif2
$\tilde{\xi}_2(j)$
b
$P(\tilde{\Xi}_1,\tilde{\Xi}_2)=\int_{\tilde{\Xi}_1}^{\infty}\int_{\tilde{\Xi}_2}^{\infty}\gamma(\Xi_1)\gamma(\Xi_2)\,d\Xi_1 d\Xi_2$
OR
$P(\tilde{\Xi}_1)P(\tilde{\Xi}_2)$
$\Xi_2$
$\Xi_1$
c
$P(\tilde{\Xi}_1,\tilde{\Xi}_2)=1-\int_0^{\tilde{\Xi}_1}\int_0^{\tilde{\Xi}_2}\gamma(\Xi_1)\gamma(\Xi_2)\,d\Xi_1 d\Xi_2$
AND
$P(\tilde{\Xi}_1)+P(\tilde{\Xi}_2)-P(\tilde{\Xi}_1)P(\tilde{\Xi}_2)$
$\Xi_2$
$\Xi_1$
d
$P(\tilde{\Xi}_1,\tilde{\Xi}_2)=1-\int_{\tilde{\Xi}_1}\int_0^{\tilde{\Xi}_2}\gamma(\Xi_1)\gamma(\Xi_2)\,d\Xi_1 d\Xi_2$
$1+P(\tilde{\Xi}_2)-P(\tilde{\Xi}_1)P(\tilde{\Xi}_2)$
AND NOT
$\Xi_2$
$\Xi_1$

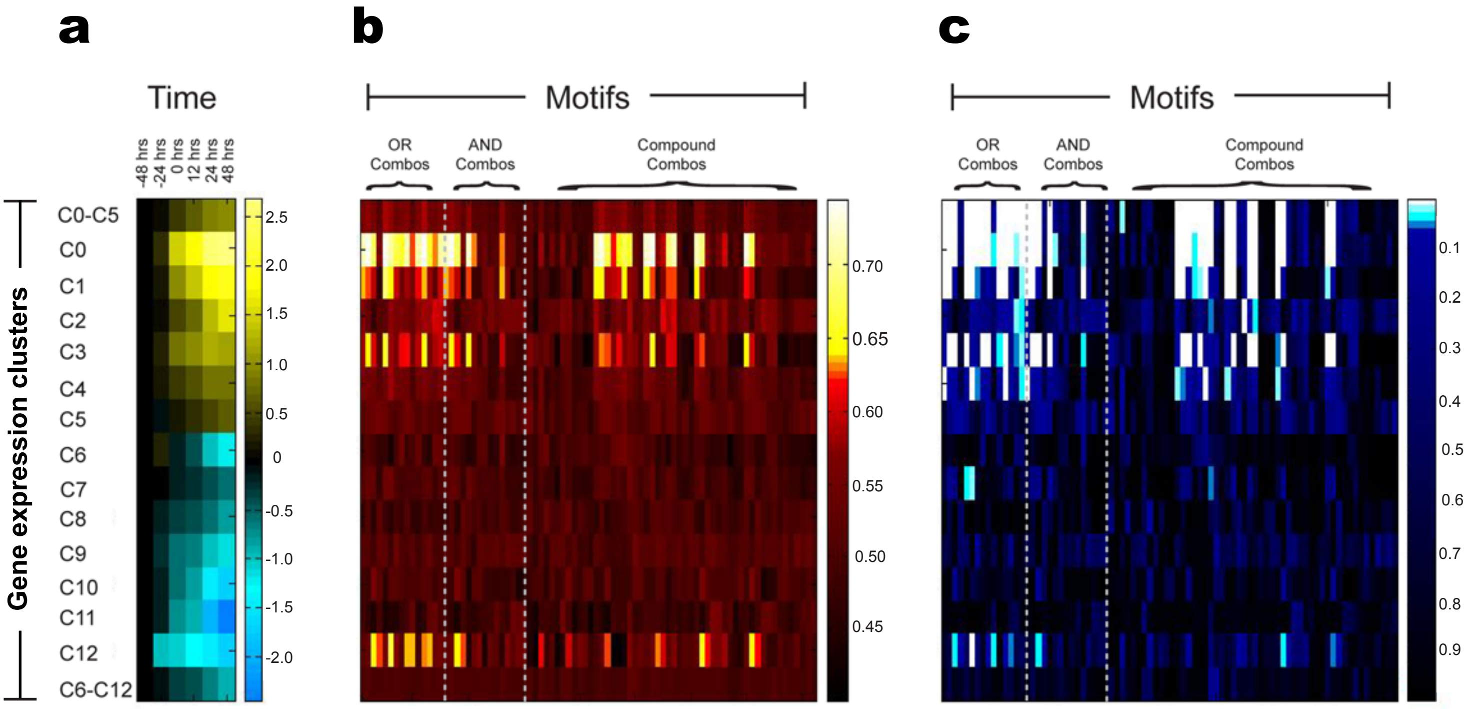
**Supplementary Figure 3: Schema of scoring scheme for PhylCRM, for case of multiple motifs.** (**a**) For two potentially overlapping motifs with positional scores $\xi_1$ and $\xi_2$, a de-overlapping step is performed (see text) where $\xi_i(j)$ =0 if $\xi_i(j) \neq \max\{\xi_1(j), \xi_2(j)\}$, $i \in \{1,2\}$. This step prevents motif-matches from being double-counted. (**b**) A restrictively-defined tail for the joint distribution of window scores $P(\Xi_1,\Xi_2)$. Here, a window can receive a good score (i.e., low $P(\Xi_1,\Xi_2)$) if it is enriched for either of the motifs, and thus this tail can be interpreted as an OR. (**c**) A generously-defined tail for the joint distribution of window scores $P(\Xi_1,\Xi_2)$. Here, a window must be enriched for both motifs in order to score well, and thus this tail can be interpreted as an AND. (**d**) A tail that is analogous to an “AND NOT” Boolean combination. Here, a window must be enriched for motif 1, but not enriched for motif 2 in order to score highly (i.e., low $P(\Xi_1,\Xi_2)$).

## Supplementary Figure 4



**Supplementary Figure 4: Evaluation of PhylCRM and the effect of phylogeny**
(**a**) Phylogenetic tree of 11 vertebrates utilized in this study. (**b**) Sensitivity and specificity of PhylCRM on a collection of 27 sequences of length 75 kb containing a CRM, as compared to a collection of length-matched sequences. Sequences were scanned with the OR combination of MRF, Mef2, SRF and Tead, and using only human sequence. (**c**) Similar to (**b**) but using all 11 vertebrate genomes. (**d**) AUC values when using progressively larger phylogenies. H=Human, C=Chimpanzee, Q=Macaque, M=Mouse, R=Rat, D=Dog, W=Cow, O= Opposum, K=Chicken, P=Pufferfish, Z=Zebrafish. (**e**) Sensitivity and specificity when using the phylogeny HCQMRDWO and a permuted form of these motifs.

# Supplementary Figure 5



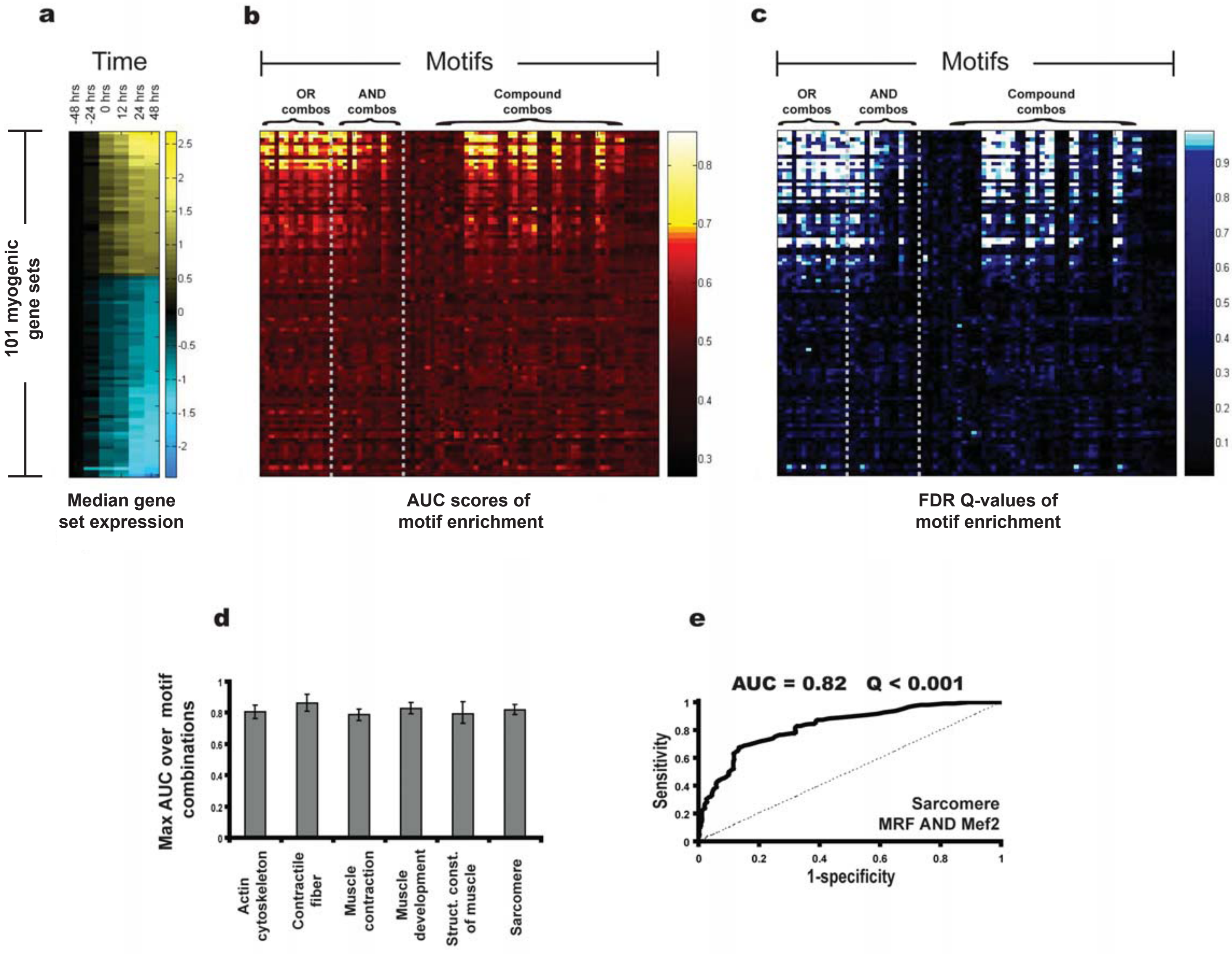
**Median gene set expression**

**AUC scores of motif enrichment**

**FDR Q-values of motif enrichment**

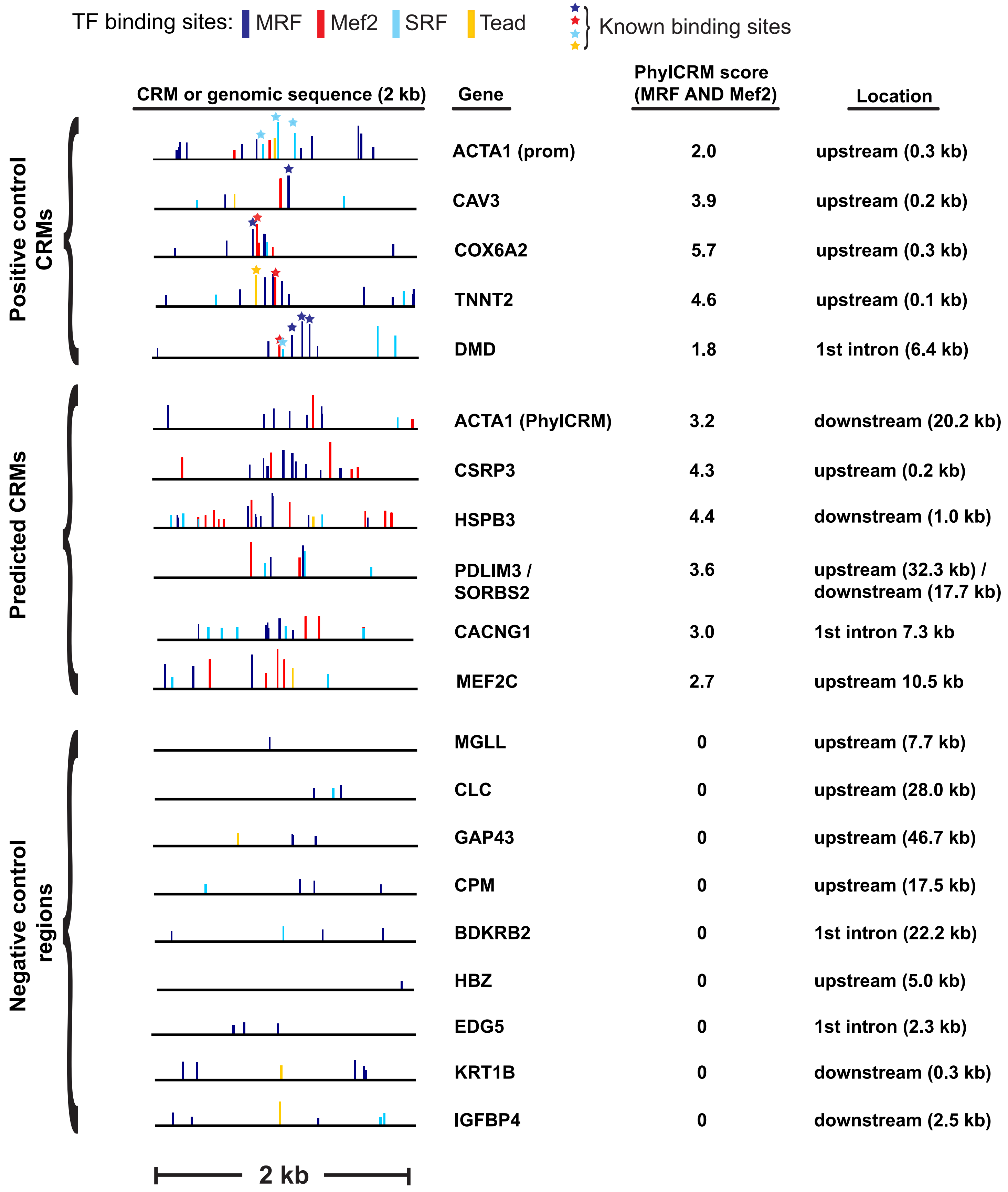
**Supplementary Figure 5: Lever screen of time course of human skeletal muscle differentiation.** (**a**) Median arcsinh value (relative to –48 hrs) of each considered expression cluster or combination of clusters. (**b**) AUC values for each TF binding site motif combination and gene set (GM-pair). (**c**) FDR q-value for each GM pair computed by Lever using a permutation test.

Supplementary Figure 6



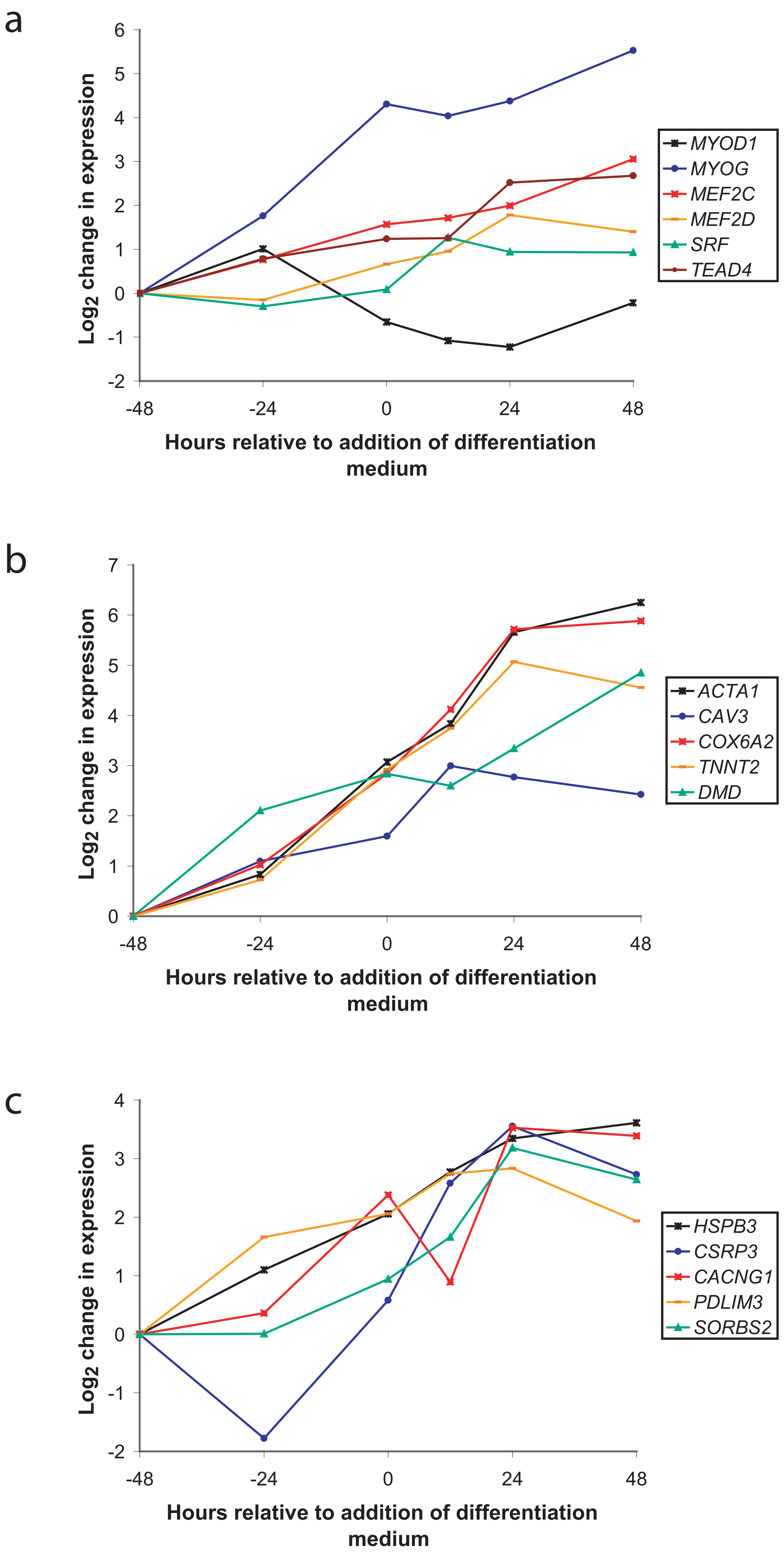
**Supplementary Figure 6: Lever screen of 101 myogenic gene sets using Boolean combinations of MRF/Mef2/SRF/Tead myogenic motifs.**
(**a**) Median signal intensity throughout the time-course of gene expression profiling for each of the 101 gene sets derived from GO categories and expression clusters.
(**b**) AUC values for each GM-pair using 75-kb regions surrounding transcription start.
(**c**) FDR Q-value for each GM- pair.
(**d**) Bar graphs indicating the maximum AUCs across all considered Boolean combinations of the motifs for these gene sets
(**e**) Sensitivity vs.specificity curves for the MRF AND MEF2 combination on the sarcomere gene set.

Supplementary Figure 7

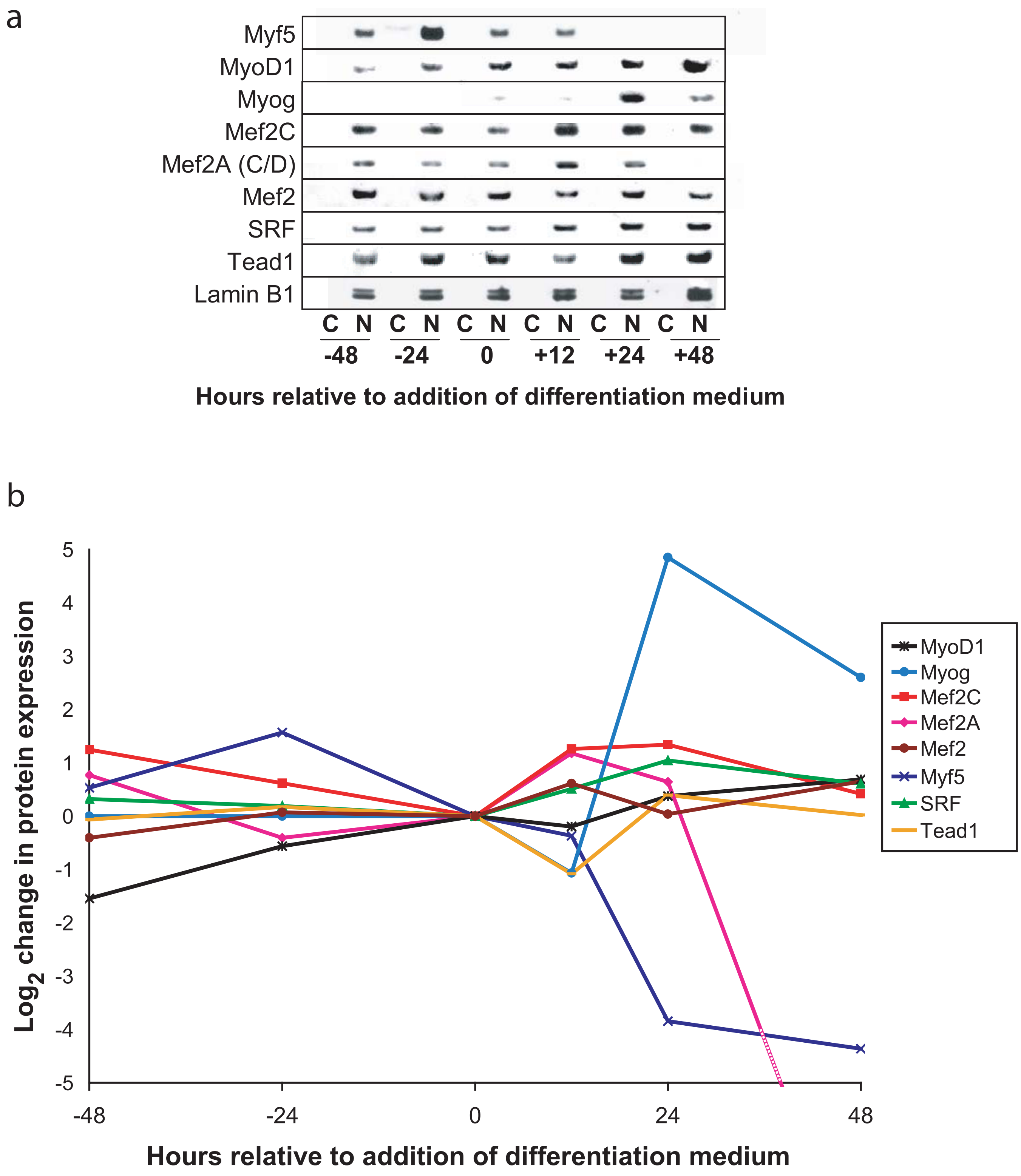


| Gene | PhylCRM score (MRF AND Mef2) | Location |
|---|---|---|
| ACTA1 (prom) | 2.0 | upstream (0.3 kb) |
| CAV3 | 3.9 | upstream (0.2 kb) |
| COX6A2 | 5.7 | upstream (0.3 kb) |
| TNNT2 | 4.6 | upstream (0.1 kb) |
| DMD | 1.8 | 1st intron (6.4 kb) |
| ACTA1 (PhylCRM) | 3.2 | downstream (20.2 kb) |
| CSRP3 | 4.3 | upstream (0.2 kb) |
| HSPB3 | 4.4 | downstream (1.0 kb) |
| PDLIM3 / SORBS2 | 3.6 | upstream (32.3 kb) / downstream (17.7 kb) |
| CACNG1 | 3.0 | 1st intron 7.3 kb |
| MEF2C | 2.7 | upstream 10.5 kb |
| MGLL | 0 | upstream (7.7 kb) |
| CLC | 0 | upstream (28.0 kb) |
| GAP43 | 0 | upstream (46.7 kb) |
| CPM | 0 | upstream (17.5 kb) |
| BDKRB2 | 0 | 1st intron (22.2 kb) |
| HBZ | 0 | upstream (5.0 kb) |
| EDG5 | 0 | 1st intron (2.3 kb) |
| KRT1B | 0 | downstream (0.3 kb) |
| IGFBP4 | 0 | downstream (2.5 kb) |

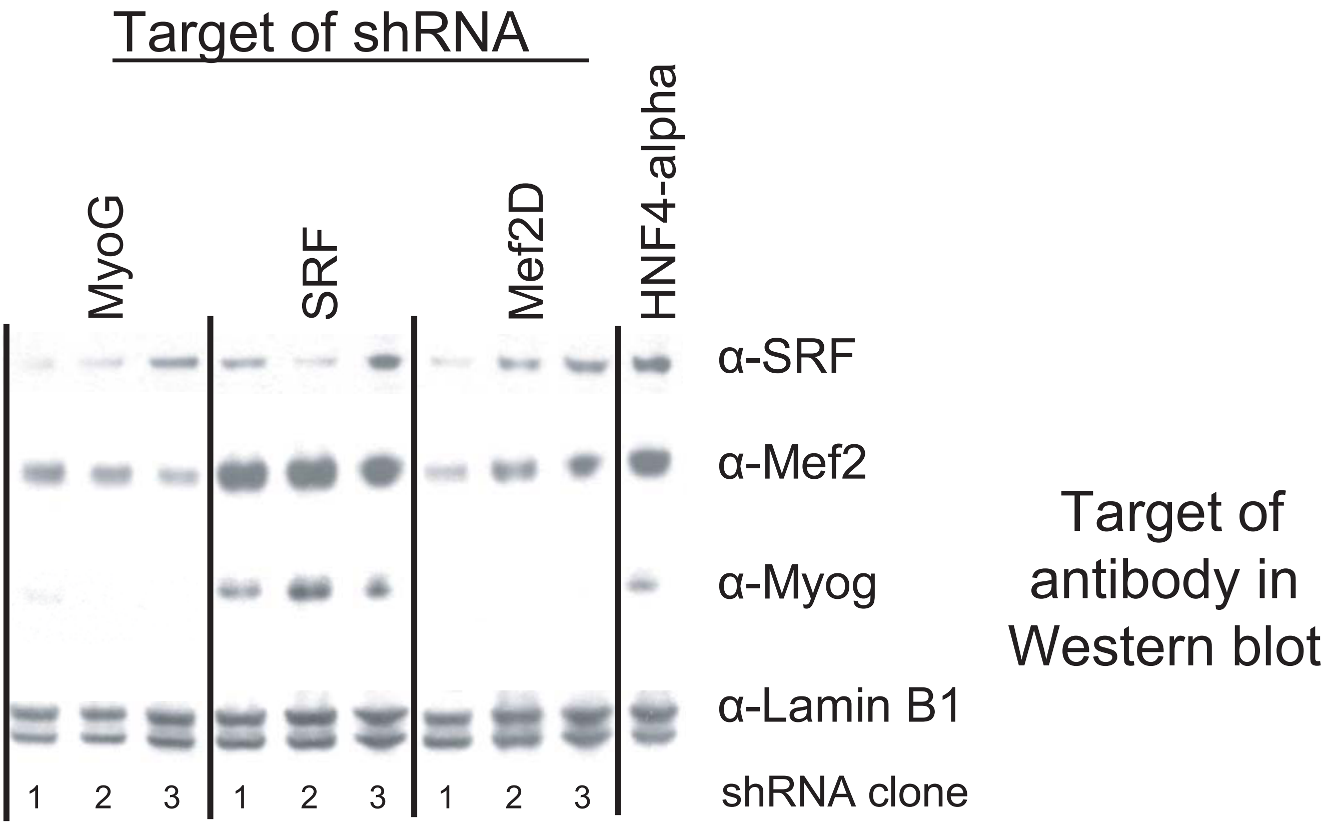
**Supplementary Figure 7. Schematic display of comptutationally predicted human CRMs and control sequences.** Previously described CRMs were used as positive controls in ChIP assays; see Supplementary Methods for full descriptions of the known and candidate CRMs. Negative control regions used in ChIP assays were chosen to not contain matches to the MRF AND Mef2 motif combinations, and to also not be enriched for the other binding sites under consideration (MRF = blue, Mef2 = red, SRF = cyan, Tead = gold), where stars indicate known binding sites. The PhylCRM score of the degree of enrichment for MRF AND Mef2 is shown (see Supplementary Methods for a description of the PhylCRM scoring scheme). Locations of sequence windows in relation to transcriptional start (if upstream or intronic) or stop (if downstream) are shown. We note that the region labeled "PDLIM3/SORBS2" was located between the PDLIM and SORBS2 genes. Also, we note that "ACTA 1 (prom)" refers to a previously known CRM located at transcriptional start, while "ACTA 1 (PhylCRM)" refers to a novel PhylCRM prediction.



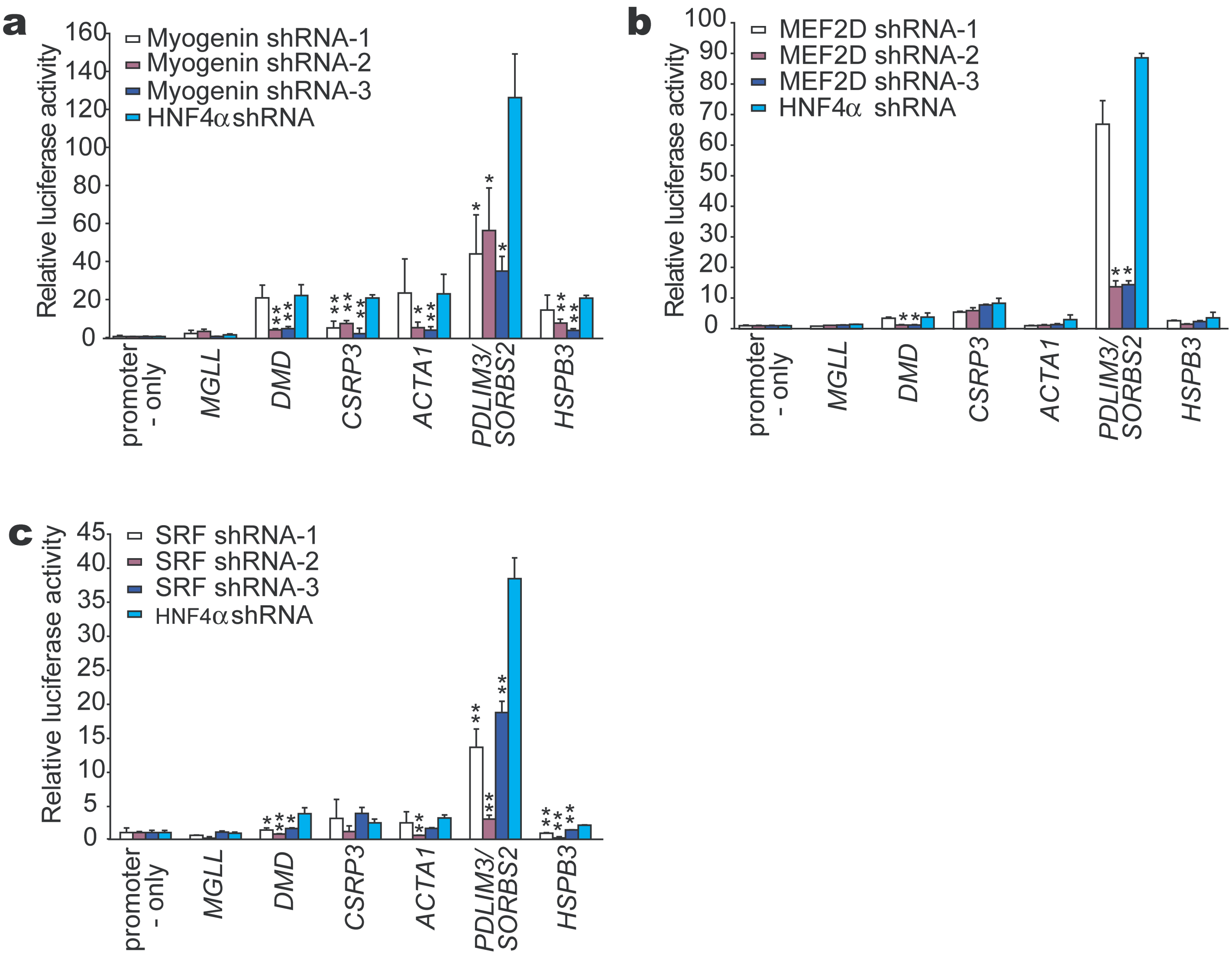
**Supplementary Figure 8 - Verification of transcription upregulation during muscle differentiation.** Total RNA from primary human cells was extracted and processed as described in **Supplementary Methods.** The following sets of transcripts were normalized to *RPS18*: (**a**) muscle transcription factors, (**b**) genes regulated by positive control CRMs, (**c**) genes associated with predicted CRMs.



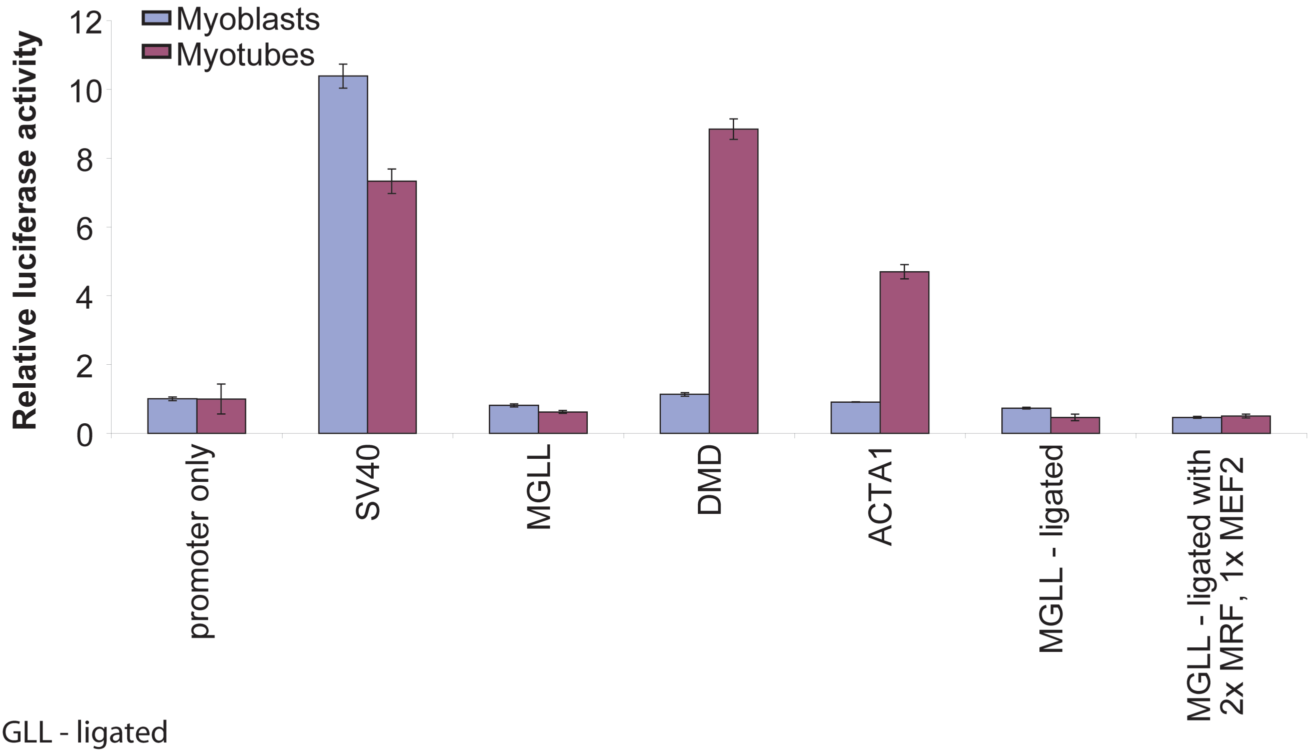
**Supplementary Figure 9: Western blots to detect levels of muscle transcription factors.** (**a**) Western blots were performed as described in **Supplementary Methods** to detect known muscle transcription factors. A lamin B1 antibody was used as normalization control. (**b**) Quantitation of bands in panel **a** was performed using lamin B1 for normalization relative to 0 hours.



**Supplementary Figure 10: Western blot analyses after RNAi knockdown.** An antibody against Lamin B1 was used to control for gel loading.



**Supplementary Figure 11: Luciferase reporter assays of predicted CRMs after shRNA knockdown. (a-c)** C2C12 myoblasts were infected with lentivirus encoding shRNAs directed against known myogenic TFs. In all experiments, lentivirus encoding shRNA against HNF4α, a liver-specific TF, was used as a negative control. Experimental knockdowns were directed against **(a)** Myogenin, **(b)** MEF2D, and **(c)** SRF. In **(a-c)**, * indicates $P < 0.05$, while vertically stacked double asterisks indicate $P < 0.005$, comparing luciferase activity in the experimental knockdown versus the HNF4α knockdown.



```
>MGLL - ligated
catgatgcattcacctcccaccaggccccaccttcaacattggggattacagttcaaaatgaggtttggtggggacacagatccaaaccatatca
ACTTGTAGGGGCAGAAAGACGTCACCTTTACTTGAATTGCAACCCTTACCTTTTCATCGCAGGCTGTAGGAG

>MGLL - ligated with MRF/Mef2/MRF sites
catgatgcattcacctcccaccaggccccaccttcaacattggggCAGCTGgttcaaaatgaggtttggtggggacacagatccaaaccatatca
ACTTGTAGGGGCAGAACTAAAAATAGTTTACTTGAATTGCAACCCTTACCTTTTCATCGCAGGCTGCAGCTG
```

**Supplementary Figure 12: Luciferase reporter assays for a synthetic CRM containing binding sites for MRF AND Mef2.** Putative and control CRMs were cloned either upstream (BglII) or downstream (BamHI) of the luciferase reporter gene of the pGL3-Promoter vector (Promega) in order to reflect the genomic location of the CRM. 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, Fig 6). As a negative control we used a human noncoding genomic region (MGLL) not enriched for matches to the four known myogenic motifs. As described in Supplementary Methods, we created variants of a shorter 167-bp MGLL negative control region by ligating segments of the original MGLL region (MGLL - ligated) or by ligating segments of the MGLL region that have two consensus MRF sites (shown in blue) and one consensus Mef2 site (shown in red). 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 cotransfected 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.