Supplementary Methods

MyoD binding matrix training

A pilot prediction of MyoD using M00184 (V$MYOD_Q6) from TRANSFAC reported 138 potential binding sites. However, comparing to the ChIP-seq results of E-box binding TFs from the ENCODE Project (12,13) indicated that 17.98% of these sites were for Hey1 binding, 7.19% for USF1 and 33.81% for c-Myc, respectively. In contrast, only 21.58% were genuine MyoD binding sites (50), suggesting that the PSSM does not well match the preference of MyoD binding.

To train a more specific MyoD binding matrix, we obtained top 50 MyoD ChIP-seq binding peaks from C2C12 myoblasts (95% confluency) and C2C12 myotubes from the summary datasets described in Cao et al. (50). Corresponding Human/Mouse/Rat alignments within ChIP-seq peaks were obtained using Galaxy (42). Sequences matched consensus NNNNCANNTGNNNN and conserved in human, mouse and rat were considered as MyoD binding sites. A total of 132 E-box alignments were obtained, corresponding to 396 MyoD binding sites in human, mouse or rat. A MyoD binding matrix was built based on these sites and the resultant matrix was shown as following:

A: 88 57 94 99 0 396 30 29 0 0 45 57 88 67
C: 105 174 71 114 396 0 52 279 0 0 159 128 111 125
G: 152 96 172 141 0 0 272 55 0 396 105 107 128 133
T: 51 69 59 42 0 0 42 33 396 0 87 104 69 71

Identification of MyoD ChIP-seq peaks

MyoD binding sites have been profiled from several skeletal muscle cell types using ChIP-seq technology (50). In Cao et al., a MyoD ChIP-seq peak was defined as a maximal contiguous region, which was composed by regions of coverage at least 4 and must contain a region of coverage at least 12, and its height was defined as the maximum coverage found in that peak region. Currently, only ChIP-seq peaks of C2C12 myoblasts (95% confluent) and myotubes were provided as summary datasets in the companion website (http://www.cs.washington.edu/homes/ruzzo/papers/DevCell/2010a/). The ChIP-seq results of C2C12 myoblasts (50% confluence), mouse embryonic fibroblasts transfected with MyoD, and primary mouse myotubes were provided as full datasets, which gave the coverage of all reads in MyoD ChIP-seq experiments mapped to the mouse genome. To define peaks
for these ChIP-seq experiments, we employed a simple strategy similar to the definition in summary
datasets. A stringent threshold of coverage at least 12 (corresponding to a false discovery rate ~10^-7, as
described in Cao et al.) was used to define peaks. The mapping regions of reads with height greater
than 12 were extracted from full datasets. Regions less than 200 bp were extended to 200 bp and the
overlapping regions were merged. The resultant regions were defined as ChIP-seq peaks and the height
of ChIP peak was defined as the maximum coverage found in that peak region.