Discovery, optimization and validation of an optimal DNA-binding sequence for the Six1 homeodomain transcription factor

Yubing Liu1,2, Soumyadeep Nandi1,2, André Martel1,2, Alen Antoun1,2, Ilya Ioshikhes1,2 and Alexandre Blais1,2,*

1Ottawa Institute of Systems Biology and 2Biochemistry, Microbiology and Immunology Department, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

Received March 12, 2012; Revised May 9, 2012; Accepted May 24, 2012

ABSTRACT

The Six1 transcription factor is a homeodomain protein involved in controlling gene expression during embryonic development. Six1 establishes gene expression profiles that enable skeletal myogenesis and nephrogenesis, among others. While several homeodomain factors have been extensively characterized with regards to their DNA-binding properties, relatively little is known of the properties of Six1. We have used the genomic binding profile of Six1 during the myogenic differentiation of myoblasts to obtain a better understanding of its preferences for recognizing certain DNA sequences. DNA sequence analyses on our genomic binding dataset, combined with biochemical characterization using binding assays, reveal that Six1 has a much broader DNA-binding sequence spectrum than had been previously determined. Moreover, using a position weight matrix optimization algorithm, we generated a highly sensitive and specific matrix that can be used to predict novel Six1-binding sites with highest accuracy. Furthermore, our results support the idea of a mode of DNA recognition by this factor where Six1 itself is sufficient for sequence discrimination, and where Six1 domains outside of its homeodomain contribute to binding site selection. Together, our results provide new light on the properties of this important transcription factor, and will enable more accurate modeling of Six1 function in bioinformatic studies.

INTRODUCTION

A defining characteristic of transcription factors (TFs) is their ability to recognize and bind to specific DNA sequences in the genome, within the regulatory region of the target genes they control. Multiple structural classes of TFs exist, among which are homeodomain factors. The primary function of the homeodomain is DNA binding. This group of regulatory factors were first discovered due to their involvement in homeotic conversions in Drosophila, but were later found to exist in essentially all eukaryote species, from yeast to humans. Bioinformatics analyses indicate that 235 homeodomain TFs are encoded in the human genome (not counting splice variants and pseudogenes) (1). This diversity is manifested by various expression profiles, protein domain composition and interaction partners. However, a large number of homeodomains seemingly recognize the same DNA sequence TAAT, or close variants, leading to the question of how redundancy in binding site selection is avoided by these proteins. The prototypical 60 amino acids long homeodomain assumes a three-dimensional structure composed of an unstructured N-terminal arm followed by three alpha-helices. The N-terminal arm and the third helix contribute most of the DNA sequence-binding specificity, and their amino acid composition are thought to contribute to this property (2,3). Two recent large-scale surveys considered the question of binding specificity of homeodomain TFs from mouse (4) or Drosophila (5). It was found that indeed, when homeodomains are considered globally, their amino acid sequence correlates with their DNA sequence-binding preferences within and adjacent to the TAAT core. However, domains residing outside of the homeodomain can also influence DNA binding, either through direct DNA contacts or interaction with dimerization partners (6–9). Therefore, it remains to be established whether homeodomain TF-binding site predictions based solely on in vitro DNA-binding preferences are sufficiently accurate to allow to predict which target genes they regulate.

The Six family of homeodomain TFs is conserved from flies to humans, and in mammalians counts six members, from Six1 to Six6. Like most homeodomain TFs, they are...
involved in controlling the development of various tissue types; for instance, Six1 and Six4 are involved in the development of the eyes, ears, kidneys and skeletal muscle (reviewed in (10,11)). We became interested in this group of TFs when we found that the DNA motif they recognize, the MEF3 sequence element, is enriched within the promoter region of Myogenic Regulatory Factors (MRFs) target genes, in muscle precursor cells (myoblasts) (12). The MEF3 element is a phylogenetically conserved motif (DNA consensus TCAGGTTC) that was originally identified within the regulatory region of only a few muscle-specific genes (13,14). The Six1 and Six4 members of the Six family, two factors essential to myogenesis, were subsequently found to bind specifically to the MEF3 sequence within the myogenin (Myog) gene promoter and activate this gene’s expression during embryogenesis (15–19). MRFs control myogenesis by activating the expression of a large cohort of genes necessary for differentiation and function of muscle cells, and in some instances they accomplish this task by cooperating with transcription factors of the Mef2 and Pbx families (20,21). The connection between the MRF target genes and MEF3 sites led us to postulate that Six family members can also cooperate with the MRFs to regulate myogenesis. This was confirmed by the genome-wide identification of Six1 target genes in mouse myoblasts and myotubes, and by functional assays which showed that Six factors can activate transcription with the MRFs in a synergistic fashion (22).

The wealth of DNA-binding information contained within our genomic Six1-binding profile (22) gave us the opportunity to examine the DNA sequence-binding preferences of this TF in myoblasts and myotubes and to analyze it in the context of the previously identified MEF3 sequence motif. Here, using de novo motif finding and position weight matrix (PWM) optimization, we report that Six1 has a broader than anticipated DNA-binding profile, which extends well beyond the canonical MEF3 consensus sequence. In vitro binding data corroborate these in vivo DNA-binding sequence preferences, suggesting that sequence-specific DNA binding by Six1 does not require interaction with dimerization partners. However, we find a discrepancy between our results and those of an in vitro DNA-binding screen for the Six1 homeodomain, performed by others (4), suggesting that Six1 homeodomain sequence is not the sole determinant of its binding to DNA.

**MATERIALS AND METHODS**

**De novo motif finding**

The Amadeus program was used for de novo DNA sequence motif discovery (23). The ‘bound’ sets corresponded to the sequences bound by Six1 in mouse myoblasts (MB, total of 1022 sites) or myotubes (MT, total of 1853 sites) with a false discovery rate (FDR) less than 10% (22). These two sets of bound sequences were broken down into three (for MB) or five (for MT) randomly assigned subgroups, respectively (i.e. MB-A, MB-B, MB-C and so on). The purpose of this sub-grouping was to obtain one subgroup for motif discovery, and to reserve the other sequences for subsequent validation of the discovered motifs. The Amadeus program was run on each of these eight subgroups, performing ‘large’ searches of 12 base pairs motifs and examining both DNA strands. For these eight searches, the background sequence set corresponded to a randomly selected subset of the genomic regions surveyed in our ChIP-on-chip analyses (20% of surveyed loci, approximately 108 Mb of sequence). Both sets of sequences were repeat-masked using RepeatMasker (24). The top ranking sequence motif (lowest corrected P-value after 25 cycles of boot-strapping where ‘bound’ and ‘background’ sequences are randomly interchanged) was retained. Searches for motifs less than 12 base pairs in length were also run and yielded very similar motifs (not shown). Finally, an averaged PWM (Six1_MB+MT) was calculated by aligning the eight PWMs and averaging the frequencies at each position. PWMs were represented graphically using the Weblogo program (25).

Motif abundance within sets of sequences was estimated using the CisGenome program (26). The number of PWM ‘hits’ within ‘bound’ and ‘background’ sequences was determined with a likelihood ratio setting of 500. The bound set of sequences for this purpose was composed of those excluded from the initial motif discovery (e.g. for the matrix identified using subgroup MB-A, the sequences in MB-B and MB-C were combined and used to test its performance). Here the background sequences were a distinct set of 20% of all surveyed loci. The relative enrichment score (the ratio of frequency of PWM hits among bound regions over that in background regions) was calculated. We also calculated the cumulative hypergeometric probability, the chance of finding at least a certain number of hits to a PWM among the bound regions (sample) given that a certain number of hits exist among the background (population) sequences (sample and population sizes expressed as total base pair length). Finally, searches were also repeated on the top 5% most phylogenetically conserved portion of the sequences.

**PWM optimization**

The methods described by Staden (27) and Bucher (28) were utilized to calculate the PWM. We adopted the base frequencies reported by Amadeus in the de novo motif searches (above). These base frequencies were converted into odds scores by dividing the frequencies by expected frequency which is calculated from the Database of Transcription Start Sites (DBTSS) for mouse using the formula described in (27):

\[
e_{bi} = n_{bi} / \sum_{i=1}^{L} n_{bi}
\]

where \( e_{bi} \) is one of 4 nucleotides (A,C,G or T) at position \( b \) of the motif and \( L \) is the length of the sequence.

The sequences used from DBTSS were of the length 1201 base pairs (−1000 to +201 from the TSS) and were aligned with respect to the TSS.
The weight for each position of the matrix is derived using the formula described in (29):

\[ w_{bi} = \ln \left( \frac{n_{bi}}{e_{bi} + s_i} \right) + c_i \]

where \( b \) is one of the 4 nucleotides, \( n_{bi} \) is the number of times base \( b \) occurs at the \( i \)-th position of the motif, \( c_i \) is a constant providing column maximum value to be zero, \( s_i \) is a smoothing parameter preventing the logarithm of zero (or too small a value). We adopted the criteria as: \( s_i = 0 \) if the first term under logarithm in Formula is larger than 0.01 \( \times \frac{n}{N_{total}} \) and \( s_i = 0.01 \times \frac{n}{N_{total}} \) otherwise, where \( n = \sum_{i=1}^{L} n_{bi} \).

To calculate the similarity score for a specific sequence within the PWM we used the formula as:

\[ S = \sum_{i=1}^{L} w_{bi} \]

where \( L \) is the length of PWM, \( w_{bi} \) is the log-odds weight of nucleotide \( b \) at position \( i \) in the PWM. To optimize the derived PWM, we have used correlation coefficient (CC) also known as Simple Matching Coefficient (SMC) (30) as the objective function. This function takes into account sensitivity and specificity of the predicted TFBS. The process of optimization started with evaluating the performance of the PWM by calculating CC at each cutoff. The CC is calculated as:

\[ CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}} \]

CC is calculated for each cutoff starting from a very stringent threshold and relaxing the threshold until we get the maximal CC. To calculate the CC we have divided the sequences into two different datasets depending on the binding preference of Six1 in myoblasts and myotubes. The sequences where Six1 is found experimentally to bind are regarded as positive and the sequences where Six1 did not bind are regarded as negative dataset. These two datasets were utilized to find out the four parameters to calculate CC: true positives, true negative, false positives and false negatives (TP, TN, FP and FN, respectively). From the above mentioned datasets, we designated TP as the number of sites from the experimental-positive dataset positively identified by the PWM with a given cutoff, and we regarded any sites computationally identified from the experimental-negative dataset where Six1 do not bind as FP. TN is calculated as the difference between the total number of sequences in the negative dataset and FP, while FN is calculated as the difference between the total number of sequences in the positive dataset and TP. The FN was calculated assuming each sequence in the positive datasets should have at least one binding site. The above step is repeated for each cutoff with the increment of 0.1, and maximal CC with respective cutoff was identified. We selected the corresponding cutoff and we further refined the performance of the PWM as follows. The PWM with the previously optimized cutoff shows sensitivity of 54% and specificity of 76%. This matrix becomes our initial PWM for the next step of optimization. Again we start with the stringent cutoff and refined the motif list used to build the PWM at each 0.1 increment in the cutoff. At each cutoff, the matrix was used to find motifs from the positive dataset. The list of motifs thus obtained from positive dataset was utilized to build the new PWM. With this PWM, we searched motifs in the negative dataset and compared the search results with those obtained from the positive dataset. We subtracted the motifs from the positive search list and rebuilt the PWM with the remaining list. The new PWM was evaluated with the function CC. We repeated the latter step for a large range of cutoffs, from stringent to relaxed, and selected the cutoff where the CC attained the maximum. The resulting matrix provides better discrimination between the positive (bound) and negative (surveyed) datasets. The respective algorithm of PWM optimization is hence named ‘Bound/Surveyed sequence Discrimination’ (BSD) algorithm. The PWM was optimized using binding data for Six1 in C2C12 myoblasts and in fully differentiated myotubes. For further validation on an independent dataset, we used genomic regions bound by Six1 only at 24 hours of differentiation, but not bound in myoblasts nor in myotubes. This corresponds to a total of 187 DNA sequences with an average length of 1061 bp, which have no overlap with the myoblasts and myotubes datasets.

**Recombinant Six1 purification**

The full-length coding sequence of mouse Six1 was amplified from C2C12 cells and cloned in frame in the pHis2 plasmid, which codes for an N-terminal hexa-histidine tag followed by a linker region. The protein was produced in the *E. coli* STAR strain. Cells were grown to an optical density at 600 nm (OD\( _{600} \)) of 0.6 and induced to produce the protein with 0.1 mM isopropyl \( \beta \)-D-1-thiogalactopyranoside for 1 h 30 min at 37°C. Cells were pelleted, resuspended in binding buffer (50 mM Tris pH 8.0, 50 mM NaCl, 0.5 mM DTT, 10 mM imidazole and 2 mM phenylmethylsulfonyl fluoride) and sonicated using a microtip sonifier. The lysate was spun at 17000g for 15 min at 4°C, and the supernatant was applied to Nickel-Sepharose beads (GE Healthcare). Beads were washed in binding buffer containing 40 mM imidazole, and elution was carried out using binding buffer containing 500 mM imidazole. The eluate was immediately bound to heparin-Sepharose beads (GE Healthcare). The beads were washed with wash buffer (50 mM Tris pH 8.0, 650 mM NaCl, 0.5 mM DTT and 2 mM phenylmethylsulfonyl fluoride) and eluted in a similar buffer containing 750 mM NaCl. The eluate was dialyzed for 18 hours against a similar buffer reduced to 150 mM NaCl and containing no imidazole, then concentrated using Amicon Ultra (30 kDa cut-off, Millipore), aliquoted, frozen in liquid nitrogen and stored at \(-80°C\). Coomasie blue staining of the purified Six1 protein indicated an estimated purity of 90% (Figure 3A). For work with the homeodomain of Six1, amino acids 110 to 201 of mouse Six1 were cloned using the same strategy as for the full-length protein. This
Six1-HD protein therefore contains the homeodomain and 15 amino acids of flanking sequence on each side, which conforms to what Berger et al. have used. Purification of the histidine-tagged Six1-HD protein from E. coli was performed in the same way as for the full-length protein, except that dialysis and concentration were performed using devices with smaller pore sizes (3 kD cut-off).

Electrophoretic mobility shift assays and calculation of dissociation constant ($K_{\text{app}}$)

EMSA experiments were performed using His-Six1 and fluorescently-labelled double-stranded DNA probes, which were prepared by end labelling of double-stranded DNA containing a G nucleotide overhang at each end of the molecule, with the Klenow enzyme Exo- (NEB) and Cy5-labelled deoxycytidine triphosphate (dCTP) (GE Healthcare). In all cases, the probe sequence context was gTTAGAGGGGGCTCAGGTTTC as the top strand. The initial small script ‘g’ represents the additional nucleotide overhang used for labelling, while the underlined nucleotides constitute the MEF3 site. In order to test the influence of various MEF3 nucleotide substitutions on Six1 binding, and to disregard the putative influence of surrounding nucleotides, we changed the sequence of the MEF3 site while retaining the same surrounding sequence context. EMSA-binding reactions contained varying amounts of His-Six1 protein in a fixed volume (4 μl in 50 mM Tris, 150 mM NaCl and 1 mM DTT) and 20 fmoles of probe in a final volume of 10 μl. The binding buffer was composed of Hepes 25 mM pH 7.6, KCl 8 mM, dDc 1 μg, MgCl2 5 mM, Glycerol 10% v/v. The reactions were set up on ice, then incubated at 37°C for 5 min, and loaded on a 5% w/v acrylamide gel (29:1 ratio acrylamide to bis-acrylamide) containing 2% glycerol, with TGE 0.5× (12.5 mM Tris, 95 mM Glycine, 0.5 mM EDTA) as the running buffer. After separation, the gels were rinsed in water, and the fluorescent signal was quantitated using a Typhoon Phosphorimager (GE Healthcare), adjusting the photomultiplier tube voltage so that none of the signal is saturated. The ratio of the volume of shifted probes over that of the total (shifted and free probes) was calculated using ImageQuant software (GE Healthcare) for each concentration of Six1. To determine the $K_{\text{app}}$ of protein-DNA binding, we determined the concentration of Six1 protein (in nanomolar) required to reach half maximal binding, using the function of one site saturation in SigmaPlot and following recommendations outlined in (31).

RESULTS

Identification of a novel MEF3-like motif

We have previously obtained ChIP-on-chip data for Six1 binding in the C2C12 cell line of mouse myoblasts (22). The experiment was performed in proliferating myoblasts and in differentiated myotubes and led to the identification of 1022 and 1853 high-confidence bound genomic loci in these two cell types, respectively. The average length of Six1-bound loci is 1230 bp; we used Amadeus, a de novo motif finding program, to precisely identify the DNA sequence motif most likely recognized by Six1 in these genomic regions. Our assumption is that the most abundant motif in these bound regions should be the one directly bound by Six1. For the purpose of motif discovery and subsequent testing, the search was run multiple times on the Six1 ChIP-on-chip target sequences and partitioned in eight subsets of approximately 330–370 sequence (see materials and methods for details). As shown in Figure 1, we obtained eight similar PWMs with little if any difference between the PWMs of Six1-bound genomic sequences detected in myoblasts and myotubes. Additionally, to summarize these results we also combined these eight PWMs to obtain an average matrix, called Six1_MB+MT (Figure 1 and Supplementary Table S1).

Figure 1 clearly shows that our novel matrices share a strong resemblance with the previously identified MEF3 sequence motif represented in the TRANSFAC database. However, the new PWMs are clearly more degenerate, since at multiple positions more than one nucleotide is allowed; this is most obvious near the center of the motif, at positions 6–8, where the preference for the AGG nucleotides is weaker than in the TRANSFAC motif. In contrast, positions 4 (T) and 12 (C) display the lowest variability. These differences have important implications for the prediction of target gene binding by Six1: using the inflexible TRANSFAC MEF3 element would possibly overlook a large number of true targets.

To determine if these differences are significant and if the increased degeneracy in our novel PWMs has an impact of their predictive value, we evaluated the sensitivity and specificity of the PWMs. We compared the number of matches to each PWM that can be found in bound sequences and in control sequences using the Cisgenome motif mapping program: a larger number of matches indicate higher sensitivity, while the specificity is given by the enrichment ratio (frequency of matches in bound sequences divided by that in control sequences). We first verified that hits to the Six1_MB+MT PWM are substantially enriched among all strata of the Six1-bound loci found by ChIP-on-chip, not solely among the top (highest confidence) loci (Figure 2). Table 1 gives the results of the comparison of our de novo-identified motifs and reveals that all nine motifs we generated from our ChIP-on-chip data are present in large numbers among the bound genomic regions (i.e. the PWMs are sensitive) and are characterized by substantial enrichment levels (i.e. they are also specific). Importantly, the results presented in Table 1 clearly show that the novel matrices outperform the TRANSFAC motif in both specificity (higher enrichment levels) and sensitivity (larger number of sites). These analyses were performed giving equal consideration to all genomic sequences. When only the phylogenetically conserved regions of the bound loci were studied, the enrichment level of the PWMs increased, as can be expected for the binding sites of a developmentally important TF (32,33). Here again, our de novo PWMs outperformed the TRANSFAC matrix with their...
Figure 1. Position weight matrices for Six1 and other homeodomain TFs. (A) PWMs generated in this study. (B) Previously reported PWMs for Six1. (C) The PWMs of other well-characterized homeodomain TFs. The PWMs were represented graphically using sequence logos.
higher enrichment levels (Table 1, enrichment—conserved sites only).

Secondly, we also compared our novel matrices to that identified by Berger et al. (4) by probing protein-binding microarrays with the bacterially expressed Six1 homeodomain in isolation, excluding the N-terminal Six domain as well as the C-terminal region (11). Again, our de novo matrices outperform this motif, both when all sequences or only the conserved subset were considered (Table 1). We note that the similarity between our and Berger’s matrices is limited to positions 4–6 of our motif (10–12 of their motif, consensus TCA).

Finally, we also verified whether the binding sites of other homeodomain transcription factors, including some that are involved in controlling myogenesis (Nkx2.5, Msx1, Pbx1, Pax3), are enriched among the genomic sites bound by Six1. None of these were enriched to a significant level within the Six1-bound genomic regions (Table 1). The canonical ‘ATTA’ (reverse-complement of TAAT) DNA sequence motif recognized by homeodomain transcription factors (e.g. Nkx2.5, in Figure 1) is observed in the Six1_MB+MT matrix at positions 8 to 11 (consensus (A/G)TT(T/A)). However, among the matches to the PWM that we have identified, only 222 out of 1873 conform to the canonical TAAT sequence at these positions; this sub-motif ranks fourth in frequency, behind GTTT, GTTA and ATTT (536, 374 and 268 hits respectively, Supplementary Table S2). Together, the results of this analysis suggest that Six1-bound DNA elements are not limited to the canonical ‘ATTA’ DNA sequence motif shared by several homeodomain TFs and provide a PWM that characterizes Six1 DNA-binding preferences with improved accuracy over all other existing matrices.

Broad sequence specificity of DNA binding by Six1

The Six1 PWM that we established is substantially different from other homeodomain TFs as well as from the PWMs previously reported for Six1 (TRANSFAC and Berger et al.). We were especially intrigued by the rather degenerate nature of the central portion of the matrix. Consequently, we used electrophoretic mobility shift assays (EMSA) to probe Six1’s ability to bind a range of DNA sequences that is wider than previously expected, avoiding the contribution of other confounding factors.

First, we set out to determine the apparent equilibrium dissociation constant (K_{dpp}^{pp}) of purified Six1 for the MEF3 site present in the Myog proximal promoter (15,22,34,35), a site that is identical to the consensus DNA motif established by the TRANSFAC PWM, and which is to date the best characterized Six1-binding site. EMSA reactions were performed in the presence of fixed amounts of Cy5-labelled probe and increasing amounts of the Six1 protein (Figure 3A). We found that Six1 binds to the mouse Myog MEF3 site with a K_{dpp}^{pp} of 35 nM (Figure 3B). Next, we aimed to verify whether the sequence preferences given by our PWM reflect the

Figure 2. Enrichment of the Six1_MB+MT matrix hits within the loci bound by Six1 in myotubes, as a function of their score in the ChIP-on-chip experiment. The 1853 genomic loci bound by Six1 in myotubes were ranked in decreasing order of ChIP enrichment and subdivided in bins of 50 regions. Matches to the Six1_MB+MT matrix were identified, and the numbers of PWM hit per bin (left-hand y axis) or per base pair (right-hand y axis) were calculated. As controls, five sets of randomly selected genomic regions were also scanned for PWM hits. Additionally, two sets of sequences totaling 108 Mb and originally surveyed in the ChIP-on-chip experiments were also scanned here.
Table 1. Results of the sensitivity and specificity searches for new and existing Six1 PWMs

| Name                      | Target list | No. of sites (all)b | Enrichment (all sites) | P-value (all sites) | No. of sites (conserved only)b | Enrichment (conserved sites only) | P-value (conserved sites) |
|---------------------------|-------------|---------------------|------------------------|---------------------|-------------------------------|----------------------------------|--------------------------|
| Six1_MB-A_m01.mat         | MB_BC       | 485                 | 4.73                   | <1E-16              | 139                           | 6.49                             | <1E-16                   |
| Six1_MB-B_m01.mat         | MB_AC       | 449                 | 4.45                   | <1E-16              | 141                           | 7.21                             | <1E-16                   |
| Six1_MB-C_m01.mat         | MB_AB       | 471                 | 5.23                   | <1E-16              | 144                           | 7.31                             | <1E-16                   |
| Six1_MT-A_m01.mat         | MT_BCDE     | 977                 | 2.58                   | <1E-16              | 246                           | 4.60                             | <1E-16                   |
| Six1_MT-B_m01.mat         | MT_ACDE     | 1043                | 2.96                   | <1E-16              | 260                           | 4.52                             | <1E-16                   |
| Six1_MT-C_m01.mat         | MT_ABDE     | 888                 | 2.73                   | 2.2E-16             | 240                           | 4.70                             | <1E-16                   |
| Six1_MT-D_m01.mat         | MT_ABCE     | 891                 | 3.08                   | <1E-16              | 248                           | 4.92                             | 1.3E-15                  |
| Six1_MT-E_m01.mat         | MT_AB CDC    | 895                 | 2.36                   | <1E-16              | 242                           | 3.93                             | <1E-16                   |
| Six1_MB+C+MT.mat          | MB ABC      | 1144                | 3.50                   | <1E-16              | 321                           | 4.99                             | <1E-16                   |
| Berger_Six1_0935.mat      | MB ABC      | 308                 | 1.38                   | 2.8E-08             | 51                            | 1.52                             | 2.6E-03                  |
| M00319-V$MEF3_B02.mat     | MB ABC      | 350                 | 0.67                   | 1.0E+00             | 116                           | 0.67                             | 1.0E+00                  |
| M00510-V$MSX1_01-Msx-1.mat| MT AB CDC    | 782                 | 0.77                   | 1.0E+00             | 203                           | 0.72                             | 1.0E+00                  |
| M00640-V$HOXA4_Q2-HOXA4.mat| MB ABC     | 697                 | 0.89                   | 1.0E+00             | 185                           | 0.99                             | 5.9E-01                  |
| M00241-V$NKX25_02-Nkx2-5.mat| MB ABC    | 472                 | 0.77                   | 1.0E+00             | 116                           | 0.69                             | 1.0E+00                  |
| M00360-V$PAX3_01-Pax-3.mat| MT ABC      | 866                 | 0.72                   | 1.0E+00             | 188                           | 0.68                             | 1.0E+00                  |
| M00394-V$LHX3_01-Lhx3a.mat| MT ABC      | 63                  | 1.13                   | 2.9E-01             | 2                             | 0.39                             | 9.7E-01                  |
| M00394-V$MEF3_B02.mat     | MB ABC      | 241                 | 1.00                   | 5.2E-01             | 52                            | 0.80                             | 9.5E-01                  |
| M00394-V$MSX1_01-Msx-1.mat| MT ABC      | 528                 | 1.12                   | 6.4E-03             | 105                           | 1.00                             | 5.2E-01                  |
| M00096-V$SPBX1_01-Pbx1a.mat| MB ABC      | 541                 | 0.80                   | 1.0E+00             | 147                           | 0.94                             | 7.7E-01                  |
| M00096-V$MEF3_B02.mat     | MB ABC      | 977                 | 0.76                   | 1.0E+00             | 211                           | 0.83                             | 1.0E+00                  |

Enrichment of binding sites predicted by PWMs discovered for Six1, for existing Six1 PWMs and for other homeodomain transcription factors has been illustrated.

aList of target genomic regions scanned with a given PWM. MB indicates Six1-bound targets in myoblasts, and MT those bound in myotubes.

bNumber of sites corresponding to ‘hits’ to the PWM, irrespective of their phylogenetic conservation.

cThe enrichment is given as the ratio of hits found in the indicated target set over those found in a fraction of the ChIP-surveyed sequence space, prorated by the length of each group of sequences in base pairs.

dxThe P-value represents the cumulative hypergeometric probability subtracted from 1.

Regions outside the homeodomain contribute to DNA-binding sequence specificity

In their large-scale study of mouse homeodomains, Berger et al. reported a DNA sequence motif preferred by the Six1 homeodomain (Six1-HD, Figure 1B) that is rather different from the one we report here for the full-length Six1 protein (Figure 1A). This is consistent with the fact that the Berger et al. PWM gives a substantially weaker match compared to the GTA prefix, which is present in mut02 but absent in the wild-type probe. In contrast, the full-length Six1 protein binds both sequences with comparable affinities (Table 2, 34.7 and 28.7 nM for WT and mut02, respectively).
mut02, respectively), in accordance with the fact that our Six1_MB+MT PWM attributes considerably less importance to the prefix sequence (Figure 1).

Secondly, we designed new Myog mutant probes (mut32, mut33 and mut34), which conform to the Berger et al. PWM at the prefix (GTA instead of GGC), but they deviate from the wild-type sequence near the 3' end (Table 2). Using these new probes, we again compared Six1-HD and Six1 by EMSA. We found that Six1-HD is mostly unaffected by these mutations, including replacement of the suffix C by a G (mut34, Figure 4B, top row). This is consistent with the reported PWM, which does not dictate any sequence preference at these positions. On the other hand, two of these mutations greatly affect full-length Six1 binding (Figure 4B, bottom row). Based on these substantial differences in DNA sequence preferences between Six1-HD and Six1, we conclude that binding site selection by Six1 involves not only its homeodomain but also regions outside of it.

**Computational optimization of the Six1 PWM**

As outlined above the novel Six1_MB+MT PWM discovered de novo outperforms other existing matrices. However, we postulated that it could still be improved using bioinformatic approaches, considering a potential ‘dilution effect’ due to the length of peaks from the ChIP-on-Chip technique (a stretch of 12 bp motif in peaks of 1230 bp on average). We employed for this purpose a novel BSD approach (Bound/Surveyed Discrimination) developed in this study (see Methods for details). Using our ChIP-on-chip data, we sought to determine if changes to the nucleotide weight values within the matrix could be introduced and further increases its specificity and sensitivity to discriminate between the ‘bound’ and ‘surveyed’ sets of sequences. In doing this, we made the assumption that matches to the PWM in ‘bound’ sequences represent biologically true binding sites, while matches to the ‘surveyed’ regions represent mostly ‘not-bound’ sites. We note, however, that the ‘bound’ set of sequences is a subset of the ‘surveyed’ sequences and so surveyed sequences do contain truly bound sites. This led to the generation of the Six1-opti matrix (Figure 1A, bottom). A receiver-operator characteristics curve analysis of the two PWMs reveals that indeed, the optimized PWM performs better than the original Six1_MB+MT matrix (or the TRANSFAC and Berger PWMs, Figure 5A), since at any given level of specificity, the optimized PWM has enhanced sensitivity. The benefits of using our optimization procedure are unlikely to
Multiple changes

Suffix changes

| Name | Sequence | Rationale | $K_d^{app}$ (nM)$^d$ |
|------|----------|-----------|---------------------|
| WT   | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Wild-type sequence | 34.7 ± 7.9 |
| Myog_WT | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Most frequent prefix | 28.7 ± 4.3 |
| Myog_mut01 | gTTAGAGGGGATCAGTTTCTGTCGGTGGC | Prefix conforms to Berger et al. | 16.8 ± 2.4 |
| Myog_mut03 | gTTAGAGGGGATTCTGTCGGTGGC | Least frequent prefix | >350 |
| Myog_mut04 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | G is second most frequent after A | 40.9 ± 4.1 |
| Myog_mut05 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | C is least frequent nucleotide | 63.8 ± 7.7 |
| Myog_mut06 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | T is third most frequent after A | >350 |
| Myog_mut07 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 52.3 ± 5.6 |
| Myog_mut08 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 34.2 ± 1.8 |
| Myog_mut09 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 29.4 ± 2 |
| Myog_mut10 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 32.4 ± 2.1 |
| Myog_mut11 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | A is second most frequent after T | 24.2 ± 1.3 |
| Myog_mut12 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Rare nucleotide | 81.6 ± 10 |
| Myog_mut13 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Rare nucleotide | 72.2 ± 3.1 |

Core changes

Prefix changes

| Name | Sequence | Rationale | $K_d^{app}$ (nM)$^d$ |
|------|----------|-----------|---------------------|
| Myog_mut01 | gTTAGAGGGGATCAGTTTCTGTCGGTGGC | Most frequent prefix | 28.7 ± 4.3 |
| Myog_mut02 | gTTAGAGGGGATCAGTTTCTGTCGGTGGC | Prefix conforms to Berger et al. | 16.8 ± 2.4 |
| Myog_mut03 | gTTAGAGGGGATTCTGTCGGTGGC | Least frequent prefix | >350 |

Suffix changes

| Name | Sequence | Rationale | $K_d^{app}$ (nM)$^d$ |
|------|----------|-----------|---------------------|
| Myog_mut04 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | G is second most frequent after A | 40.9 ± 4.1 |
| Myog_mut05 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | C is least frequent nucleotide | 63.8 ± 7.7 |
| Myog_mut06 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | T is third most frequent after A | >350 |
| Myog_mut07 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 52.3 ± 5.6 |
| Myog_mut08 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 34.2 ± 1.8 |
| Myog_mut09 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 29.4 ± 2 |
| Myog_mut10 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 32.4 ± 2.1 |
| Myog_mut11 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | A is second most frequent after T | 24.2 ± 1.3 |
| Myog_mut12 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Rare nucleotide | 81.6 ± 10 |
| Myog_mut13 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Rare nucleotide | 72.2 ± 3.1 |

Table 2. Summary of EMSA experiments

| Name$^a$ | Sequence$^b$ | Rationale$^c$ | $K_d^{app}$ (nM)$^d$ |
|----------|--------------|--------------|---------------------|
| Myog_WT  | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Wild-type sequence | 34.7 ± 7.9 |
| Myog_mut01 | gTTAGAGGGGATCAGTTTCTGTCGGTGGC | Most frequent prefix | 28.7 ± 4.3 |
| Myog_mut02 | gTTAGAGGGGATCAGTTTCTGTCGGTGGC | Prefix conforms to Berger et al. | 16.8 ± 2.4 |
| Myog_mut03 | gTTAGAGGGGATTCTGTCGGTGGC | Least frequent prefix | >350 |
| Myog_mut04 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | G is second most frequent after A | 40.9 ± 4.1 |
| Myog_mut05 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | C is least frequent nucleotide | 63.8 ± 7.7 |
| Myog_mut06 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | T is third most frequent after A | >350 |
| Myog_mut07 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 52.3 ± 5.6 |
| Myog_mut08 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 34.2 ± 1.8 |
| Myog_mut09 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 29.4 ± 2 |
| Myog_mut10 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Very frequent dinucleotide | 32.4 ± 2.1 |
| Myog_mut11 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | A is second most frequent after T | 24.2 ± 1.3 |
| Myog_mut12 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Rare nucleotide | 81.6 ± 10 |
| Myog_mut13 | gTTAGAGGGGACCTCAGTTTCTGTCGGTGGC | Rare nucleotide | 72.2 ± 3.1 |

$^a$Myog_WT is the Myog probe with wild-type MEF3 consensus in the center. Myog_mut01 to Myog_mut30 are probes with various mutations in the MEF3 consensus. Myog_mut31 is the same probe as Myog_mut03 cited as a different rationale.

$^b$Mutated nucleotides in the MEF3 consensus are highlighted in black. The lower cap ‘g’ nucleotide was added for fluorescent labelling purposes.

$^c$Mutated nucleotides in the MEF3 consensus are highlighted in black. The lower cap ‘g’ nucleotide was added for fluorescent labelling purposes.

$^d$The improved Six1-opti PWM allowed us to identify new potential binding sites that may have been missed using the original Six1 MB + MT matrix. At similar sensitivity and specificity (~60% and ~73%, respectively), the Six1-opti matrix identified 322 novel putative binding sequences (occurring a total of 505 times among our Six1-bound genomic loci) that were missed with the starting matrix (Figure 5B). On the other hand, the starting matrix identified only 11 sequences (17 occurrences) not found by the optimized PWM. It is also noteworthy to consider these results in terms of putative target gene identification, since this is a common use of PWM scanning programs. Using the Six1-opti PWM would allow to identify 1051 target genes (i.e. sequences with at least one hit to the PWM), while the original matrix would only recognize 747 of them. This represents a 40.7% increase in sensitivity.
To determine if Six1 binds the new sequences identified by the optimized matrix, we used EMSA to assess the binding affinity of Six1 to them. We selected 10 novel sites unique to Six1-opti and 5 sites uniquely identified with Six1_MB+MT for validation (Table 2). Interestingly, 9 out of 10 sites unique to Six1-opti have comparable affinity to the Myog probe, whereas only 2 out of 5 sites unique to Six1_MB+MT are bound by Six1 with a measurable $K_d$, suggesting that our optimization approach improved the discriminatory power of our Six1 PWM. Finally, as an ultimate test of the relevance of the binding site predictions made by the optimized PWM, we repeated the search on an independent set of Six1 target loci that are bound by Six1 24 hours after the onset of myoblast differentiation, but not in myoblasts or myotubes (see materials and methods). As expected, with 49 common binding sequences out of 187 Six1-bound regions, 23 binding sites were identified solely with the Six1-opti PWM, while only 2 binding sequences were uniquely defined with Six1 MB+MT (Figure 5C). This confirms the superiority of the optimized PWM over the original matrix.

**DISCUSSION**

Using bioinformatic analysis of the genomic binding profile of Six1 in muscle cells, we have found that this transcription factor has the ability to bind to a much broader range of DNA sequences than anticipated. While the previously reported MEF3 sequence motif is clearly enriched among genomic regions bound by Six1, other sequences that diverge substantially from this DNA element are also found preferentially at Six1-bound loci. We used an *in vitro* binding assay with recombinant Six1 to confirm that the protein indeed binds to these divergent sequences with high affinity, thereby ruling out an artificial effect of the de novo sequence discovery algorithm we used. The novel PWM should prove to be useful to studies that employ TF target site prediction to elucidate the structure of regulatory networks (41). For Six1, a TF involved in the genesis of multiple tissue types, a more precise DNA binding motif may contribute to discovering novel direct targets, elucidating composite regulatory networks and rationalizing its implication in diseases such as breast cancer or Branchio-oto-renal syndrome (7,42,43).

---

**Figure 4.** Substantial differences in DNA sequence selectivity between Six1-HD and Six1. (A) EMSA gels were performed with increasing amounts of Six1 homeodomain (Six1-HD) using the Myogenin WT probe (left), or the mut02 probe (right), which conforms to the consensus reported by Berger et al. using protein-binding microarrays. (B) EMSA experiments performed with the Six1-HD (top row) or Six1 (bottom row) proteins, on three derivatives of the mut02 probe (mutated positions are underlined, compared with the WT probe). Note that because the Six1-HD has a relatively low affinity for DNA in these assays, the amounts of protein used are higher than those used for the full-length Six1 protein. The $K_d$ values (all in nanomolars) for each protein on each probe are given underneath the respective gel images.
Figure 5. Performance comparison of the Six1_MB+MT and Six1-opti PWMs. (A) ROC curves. The y-axis is the sensitivity and the x-axis is the 1-specificity value. Performances of the original TRANSFAC MEF3 and Berger et al. PWMs for Six1 are also given for comparison. (B) Venn diagram indicating the number of hits to each PWM, and their overlap, at their respective optimal thresholds, among loci targeted by Six1 in myotubes. Numbers in parentheses are the number of unique sequences (one sequence can occur more than once). (C) Results for the comparison between predictions made with the Six1-opti and Six1_MB+MT PWMs, on sequences bound by Six1 only at 24 hours post-differentiation.

The results of our analyses provide useful information that may guide structure-function studies of Six1-DNA interactions. The Six family homeodomains, including that of Six1, belong to the K50 class of homeodomains, as they differ from the majority of other homeodomains classes by the change of a key DNA-binding residue, asparagine at position 50 of the HD, to a lysine. The importance of that residue in encoding DNA-binding specificity has been highlighted by a number of biochemical and structural studies (44–48). Other residues implicated in DNA-binding specificity or stability, for example those within the N-terminal arm, differ between Six and other homeodomains or within Six family members [discussed in (49–51)]. Yet, it remains to be determined precisely how DNA-binding specificity is established by Six family TFs, and by Six1 in particular. The various family members have been shown to bind to different sequences: Six3 and Six6 can bind to the canonical TAAT sequence that is bound by most homeodomain TFs, while Six1/2/4/5 have all been shown to be able to bind to sequences resembling the MEF3 element [reviewed in (11)]. Some Six proteins can also bind DNA as heterodimers with Eya family proteins, and these interactions are thought to enhance their affinity for DNA (7,51). Most TFs, including several homeodomain factors, bind to DNA in vivo as homo- or heterodimers, and stabilization of such oligomeric states has been put forth as one possible mechanism to explain the influence of Eya proteins in regulating Six proteins (51). In our experiments, one predominant oligomeric state of Six1 was detected in EMSA; however, the oligomeric state of Six1 in solution remains unclear (i.e. whether Six1 binds DNA as a monomer or as a multimer in our assays), although at very high protein concentrations slower migrating complexes became visible on EMSA gels (data not shown). We cannot tell at this point if these species represent aggregates or functionally and physiologically relevant oligomers. Further analysis of the precise mode of DNA recognition by Six1, the influence of interaction partners and possible involvement of oligomeric states, should help us understand the function of this protein and rationalize its implication in diseased states.

ChIP-on-chip analysis captures a snapshot of protein-DNA interactions as they occur in live cells, and although chromatin is immunoprecipitated with an antibody against Six1, putative DNA-binding partners of Six1 were possibly involved in the interactions we have discovered. It would be interesting to determine the genomic binding profiles of Eya proteins in myoblasts to see if indeed these proteins tend to bind DNA along with Six1, and if co-binding with Eya proteins alters DNA-binding preferences in any way. However, these experiments may prove excessively difficult to perform in vivo: while it is possible to ChIP a specific Eya-Six chromatin-bound complex (using sequential ChIP assays), it would be much more challenging to devise a way of pulling-down only Six1 chromatin complexes that do not contain Eya proteins. Interestingly, however, the results of our in vitro experiments corroborate those of our ChIP-on-chip experiments: the variety of sequence motifs enriched among Six1-bound loci is reflected in EMSA using DNA probes and Six1 alone. This leads us to postulate that protein domain(s) within Six1 itself are the main determinant of DNA sequence selection by Six1 in vivo (at least in muscle cells, where our analysis was done). As shown by others and noted above, Eya binding could influence predominantly the binding affinity rather than sequence selectivity (7,51).

The PWM we have generated is fairly close to the initial MEF3 PWM reported more than 15 years ago. The stringency (and therefore low sensitivity) of the TRANSFAC MEF3 PWM comes from the fact that it was derived from the DNA sequences of only five sites within muscle gene promoters. One can easily imagine that with larger sampling, the PWM would perform better at predicting Six1-binding sites.

The significant discrepancy between our data and those reported by Berger et al. was more puzzling. The authors used protein-binding microarrays to determine the sequence specificity of the mouse Six1 homeodomain and reported a Six1 PWM that has only limited resemblance to our PWM or to the TRANSFAC MEF3 element (4). We reason that these differences originate from the fact that only the homeodomain region of Six1 was studied, whereas our ChIP-on-chip and binding studies were...
performed with full-length Six1. Indeed, we confirmed with EMSA experiments that Six1-HD exhibits a DNA sequence preference that is in line with what Berger et al. reported, but that is substantially different from that exhibited by the full-length protein. We therefore conclude that regions outside of the Six1 homeodomain participate in DNA binding, either through direct DNA contacts or indirectly, perhaps by enabling structural stabilization.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online: Supplementary Tables 1–2.

ACKNOWLEDGEMENTS
Supplementary Tables 1–2.

Conflicts of interest statement. None declared.

REFERENCES
1. Holland,P.W., Booth,H.A. and Bruford,E.A. (2007) Classification and nomenclature of all human homeobox genes. BMC Biol., 5, 47.
2. Gehring,W.J., Qian,Y.Q., Billeter,M., Furukubo-Tokunaga,K., Schier,A.F., Resendez-Perez,D., Affolter,M., Otting,G. and Wuthrich,K. (1994) Homeodomain-DNA recognition. Cell, 78, 211–223.
3. Svingen,T. and Tonissen,K.F. (2006) Hox transcription factors and their elusive mammalian gene targets. Heredity, 97, 88–96.
4. Berger,M.F., Badis,G., Gehwe,S.A., Talmud,K., Phhillipakis,A.A., Pena-Castillo,L., Alleyn,T.M., Mnaininh,S., Botvinikn,O.B., Chan,E.T. et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell, 133, 1266–1276.
5. Noyes,M.B., Christensen,R.G., Wakabayashi,A., Stormo,G.D., Brodsky,M.H. and Wolfe,S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell, 133, 1277–1289.
6. Li,T., Stark,M.R., Johnson,A.D. and Wolberger,C. (1995) Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. Science, 270, 262–269.
7. Patrick,A.N., Schiennann,B.J., Yang,K., Zhao,R. and Ford,H.L. (2009) Biochemical and functional characterization of six branchio-oto-renal syndrome mutations. J. Biol. Chem., 284, 20781–20790.
8. Chang,C.P., Broccchier,L., Shen,W.F., Largman,C. and Cleary,M.L. (1996) Pbx modulation of hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the hox locus. Mol. Cell. Biol., 16, 1734–1745.
9. Shen,W.F., Montana,J.C., Rozenfeld,S., Moscow,J.L., Lawrence,H.J., Bachberg,A.M. and Largman,C. (1997) AbdB-like hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. Mol. Cell. Biol., 17, 6448–6458.
10. Kawakami,K., Sato,S., Ozaki,H. and Ikeda,K. (2000) Six family genes—structure and function as transcription factors and their roles in development. Bioessays, 22, 616–626.
11. Kumar,J.P. (2009) The sine oculi homeobox (SIX) family of transcription factors as regulators of development and disease. Cell Mol. Life Sci., 66, 565–583.
12. Blais,A., Tskitis,M., Acosta-Alvear,D., Sharan,R., Kluger,Y. and Dyluthf,B.D. (2005) An initial blueprint for myogenic differentiation. Genes Dev., 19, 553–569.
13. Parmace,M.S., Ip,H.S., Jung,F., Shen,T., Martin,J.F., Vora,A.J., Olson,E.N. and Leiden,J.M. (1994) A novel myogenic regulatory circuit controls slow/cardiac troponin C gene transcription in skeletal muscle. Mol. Cell Biol., 14, 1870–1885.
14. Spitz,F., Salmenen,M., Demignon,J., Kahn,A., Daegelen,D. and Maires,P. (1997) A combination of MEF3 and NFI proteins activates transcription in a subset of fast-twitch muscles. Mol. Cell. Biol., 17, 656–666.
15. Spitz,F., Demignon,J., Porteu,A., Kahn,A., Conderet,J.P., Daegelen,D. and Maires,P. (1998) Expression of myogenin during embryogenesis is controlled by Six/sine oculis homeoproteins through a conserved MEF3 binding site. Proc. Natl Acad. Sci. USA, 95, 14220–14225.
16. Laclef,C., Hamard,G., Demignon,J., Souil,E., Houbronn,C. and Maires,P. (2003) Altered myogenesis in Six1-deficient mice. Development, 130, 2239–2252.
17. Grifone,J., Demignon,J., Houbcron,C., Souil,E., Niro,C., Seller,M.J., Hamard,G. and Maires,P. (2005) Six1 and Six4 homeoproteins are required for Pax3 and mrf expression during myogenesis in the mouse embryo. Development, 132, 2235–2249.
18. Li,X., Oghi,K.A., Zhang,J., Krones,A., Bush,K.T., Glass,C.K., Nigam,S.K., Aggarwa,A.K., Maas,R., Rose,D.W. et al. (2003) Eya protein phosphatase activity regulates Six1-dach-eya transcriptional effects in mammary organogenesis. Nature, 426, 247–254.
19. Bessarab,D.A., Chong,S.W., Srinivas,P.B. and Korzh,V. (2008) Six1a is required for the onset of fast muscle differentiation in zebrafl. Dev. Biol., 323, 216–228.
20. Molken, J.D. and Olson,E.N. (1996) Combinatorial control of muscle development by basic helix-loop-helix and MADS-box transcription factors. Proc. Natl Acad. Sci. USA, 93, 9366–9373.
21. Berkes,C.A., Bergstrom,D.A., Penn,B.H., Seaver,K.J., Knoepler,P.S. and Tappq,S.J. (2004) Pbx marks genes for activation by MyoD indicating a role for a homeodomain protein in establishing myogenic potential. Mol. Cell, 14, 465–477.
22. In,Y., Chua,A., Chakroun,J., Islam,U. and Blais,A. (2010) Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation. Nucleic Acids Res., 38, 6857–6871.
23. Linharn,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: The amadeus platform and a compendium of metazoan target sets. Genome Res., 18, 1180–1189.
24. Tarailo-Graovac,M. and Chen,N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics, Chapter 4, Unit 4.10.
25. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. Genome Res., 14, 1189–1190.
26. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Olson,E.N. (1994) A novel myogenic regulatory factor of development and disease. Curr. Protoc. Mol. Biol., 18, 27081–27090.
27. Shiga,F., Takanabe,R., Raya,M., Maeda,T. and Nishimura,H. (2003) Altered myogenesis in Six1-deficient mice. Development, 130, 2239–2252.
28. Shiga,F., Takanabe,R., Raya,M., Maeda,T. and Nishimura,H. (2003) Altered myogenesis in Six1-deficient mice. Development, 130, 2239–2252.
29. Gershenzon,N.I., Stormo,G.D. and Ioshikhes,I.P. (2005) DNA binding recognition sites. Mol. Cell. Biol., 25, 88–96.
30. Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. Nucleic Acids Res., 23, 6448–6458.
41. Vavouri, T. and Elgar, G. (2005) Prediction of cis-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. *Curr. Opin. Genet. Dev.*, 15, 395–402.

42. Coletta, R.D., Christensen, K., Reichenberger, K.J., Lamb, J., Micomonaco, D., Huang, I., Wolf, D.M., Muller-Tidow, C., Golub, T.R., Kawakami, K. et al. (2004) The Six1 homeoprotein stimulates tumorigenesis by reactivation of cyclin A1. *Proc. Natl Acad. Sci. USA.*, 101, 6478–6483.

43. Ruf, R.G., Xu, P.X., Silvius, D., Otto, E.A., Beekmann, F., Muerb, U.T., Kumar, S., Neuhauer, T.J., Kemper, M.J., Raymond, R.M. Jr et al. (2004) SIX1 mutations cause branchio-oto-renal syndrome by disruption of EYA1-SIX1-DNA complexes. *Proc. Natl. Acad. Sci. USA.*, 101, 8090–8095.

44. Wilson, D.S., Sheng, G., Jun, S. and Desplan, C. (1996) Conservation and diversification in homeodomain-DNA interactions: a comparative genetic analysis. *Proc. Natl Acad. Sci. USA.*, 93, 6886–6891.

45. Tucker-Kellogg, L., Rould, M.A., Chambers, K.A., Ades, S.E., Sauer, R.T. and Pabo, C.O. (1997) Engrailed (Gln50–>Lys) homeodomain-DNA complex at 1.9A resolution: structural basis for enhanced affinity and altered specificity. *Structure*, 5, 1047–1054.

46. Chaney, B.A., Clark-Baldwin, K., Duve, V., Ma, J. and Rance, M. (2005) Solution structure of the K50 class homeodomain PITX2 bound to DNA and implications for mutations that cause rieger syndrome. *Biochemistry*, 44, 7497–7511.

47. Ades, S.E. and Sauer, R.T. (1994) Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. *Biochemistry*, 33, 9187–9194.

48. Grant, R.A., Rould, M.A., Klemm, J.D. and Pabo, C.O. (2000) Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50–>ala) complex at 2.0 A. *Biochemistry*, 39, 8187–8192.

49. Seo, H.C., Curtiss, J., Mlodzik, M. and Fjose, A. (1999) Six class homeobox genes in drosophila belong to three distinct families and are involved in head development. *Mech. Dev.*, 83, 127–139.

50. Suh, C.S., Ellingsen, S., Austbo, L., Zhao, X.F., Seo, H.C. and Fjose, A. (2010) Autoregulatory binding sites in the zebrafish six3a promoter region define a new recognition sequence for Six3 proteins. *FEBS J.*, 277, 1761–1775.

51. Hu, S., Mamedova, A. and Hegde, R.S. (2008) DNA-binding and regulation mechanisms of the SIX family of retinal determination proteins. *Biochemistry*, 47, 3586–3594.