Molecular mechanism underlying the regulatory specificity of a \textit{Drosophila} homeodomain protein that specifies myoblast identity

Brian W. Busser$^{1,*}$, Leila Shokri$^{2,*}$, Savina A. Jaeger$^{3,*}$, Stephen S. Gisselbrecht$^2$, Aditi Singhania$^{1,*}$, Michael F. Berger$^{2,5}$, Bo Zhou$^{2,*}$, Martha L. Bulyk$^{2,4,5,6,7,#}$ and Alan M. Michelson$^{1,#}$

**SUMMARY**

A subfamily of \textit{Drosophila} homeodomain (HD) transcription factors (TFs) controls the identities of individual muscle founder cells (FCs). However, the molecular mechanisms by which these TFs generate unique FC genetic programs remain unknown. To investigate this problem, we first applied genome-wide mRNA expression profiling to identify genes that are activated or repressed by the muscle HD TFs Slouch (Slou) and Muscle segment homeobox (Msh). Next, we used protein-binding microarrays to define the sequences that are bound by Slou, Msh and other HD TFs that have mesodermal expression. These studies revealed that a large class of HDs, including Slou and Msh, predominantly recognize TAAT core sequences but that each HD also binds to unique sites that deviate from this canonical motif. To understand better the regulatory specificity of an individual FC identity HD, we evaluated the functions of atypical binding sites that are preferentially bound by Slou relative to other HDs within muscle enhancers that are either activated or repressed by this TF. These studies showed that Slou regulates the activities of particular myoblast enhancers through Slou-preferred sequences, whereas swapping these sequences for sites that are capable of binding to multiple HD family members does not support the normal regulatory functions of Slou. Moreover, atypical Slou-binding sites are overrepresented in putative enhancers associated with additional Slou-responsive FC genes. Collectively, these studies provide new insights into the roles of individual HD TFs in determining cellular identity, and suggest that the diversity of HD binding preferences can confer regulatory specificity.

**KEY WORDS:** Homeodomain, Myoblast, Gene regulation

**INTRODUCTION**

\textit{Drosophila} larval somatic muscles are multinucleated myotubes with individual sizes, shapes, positions, orientations and attachments that are determined by the combinatorial activities of muscle identity genes, each of which has a unique expression pattern (Baylies et al., 1998; Busser et al., 2008). The diversity of myotube identities originates in a population of mononucleated myoblasts termed founder cells (FCs), which fuse with a more homogeneous group of neighboring muscle cells called fusion-competent myoblasts (FCMs) to form muscle precursors (Baylies et al., 1998). A subfamily of muscle identity genes encoding HD TFs (referred to herein as ‘founder cell identity homeodomains’ or FCI-HDs) has been proposed to control the unique gene expression programs of individual FCs (Baylies et al., 1998; Jagla et al., 2001). This hypothesis was investigated for the Ladybird (Lb) HD TFs which showed that Lb target genes include molecules involved in both early specification and later muscle differentiation (Junion et al., 2007). Other FCI-HD TFs include \textit{slouch} (slou) and muscle segment homeobox (\textit{msh}), which display mutually exclusive expression in adjacent FCs (Lord et al., 1995; Nose et al., 1998; Knirr et al., 1999). Both loss-of-function and gain-of-function genetic experiments have demonstrated that the normal activities of Slou, Msh and Lb are required for the proper development of all muscles derived from the FCs that express these TFs (Lord et al., 1995; Nose et al., 1998; Knirr et al., 1999; Jagla et al., 2002). In addition, overexpression of either Slou, Msh or Lb results in muscle fate transformations, consistent with the sufficiency of these TFs to specify cellular identity. However, despite these well-characterized genetic activities, the molecular mechanisms by which FCI-HD TFs interact with and function to control muscle cis-regulatory modules (CRMs) remain poorly understood.

TFs can be classified according to the structural similarity of their DNA-binding domains. For example, the DNA binding and functional specificity of some HD proteins has been shown to reside in the sequence composition of their HDs (Kuziora and McGinnis, 1989; Florence et al., 1991; Schier and Gehring, 1992; McGinnis, 1989; Florence et al., 1991).
Ekker et al., 1994; Mann and Carroll, 2002; Mann et al., 2009). Thus, it is not surprising that for some HD subclasses, such as the NK, Bed, Six and Iroquois groups, the distinct amino acid sequences of their homeodomains create unique binding preferences (Berger et al., 2008; Noyes et al., 2008). In contrast to these HD subclasses, the majority of HD TFs have a restricted range of DNA-binding specificities, which typically are centered on a canonical TAAT core (Mann et al., 2009). The low information content of such DNA-binding sites poses a challenge to understanding how these HD TFs can mediate their precise developmental functions. A further problem in interpreting the functional specificity of HDs is inherent in the widespread binding across the genome that has been documented for this TF class (Biggin, 2011).

Here, we have undertaken an integrated genomics approach to investigate the mechanisms by which the FCI-HDs Slou and Msh regulate the unique genetic programs of individual muscle FCs. We first identified Slou- and Msh-responsive genes by genome-wide expression profiling. We then used protein-binding microarrays to define the specific sequences that are bound by Msh, Slou and other mesodermal HD TFs. These studies revealed that a large subset of HD TFs, including Slou and Msh, predominantly bind to sites having a TAAT core, but that each HD also recognizes a small number of atypical or non-consensus sequences that we refer to as ‘HD-preferred’ motifs. Site-directed mutageneses revealed that Slou regulates myoblast genes through atypical binding sites that are preferentially bound by Slou relative to other HDs. Furthermore, using a computational algorithm, we found that Slou-preferred binding sequences are enriched within putative enhancers associated with Slou-responsive genes, suggesting that HD binding to atypical preferred sequences may serve as a general mode of regulation by this TF class. These findings provide fresh insights into how FCI-HDs induce the distinct genetic programs and fates of individual myoblasts.

**MATERIALS AND METHODS**

**Fly stocks**

*Drosophila* stocks containing the following transgenes and mutant alleles were used: UAS-slou and slou256 (gifts from M. Frasch, University of Enlargen, Germany), attP2 and nos-phiC31intNLS (Bischof et al., 2007) (gifts from N. Perrimon, Harvard University, USA), UAS-msh (a gift from A. Nose, University of Tokyo, Japan), *lbl-lacZ* and *mih2-lacZ* (Philippakis et al., 2006), and twi-gal4 UAS-2EGFP (Halfon et al., 2002a).

**Cloning, expression and protein binding microarray analysis of Drosophila HD TFs**

The DNA-binding domains of selected *Drosophila* HD TFs were cloned into Gateway-compatible vectors and proteins were produced either by in vitro transcription and translation, or by overexpression in *E. coli* followed by affinity purification. The method for each TF is described in supplementary material Table S2. Protein-binding microarray (PBM) assays were performed as previously described (Berger et al., 2006; Berger et al., 2008). To score 9-mers, 8-mer PBM enrichment scores were generated by a modification of the Seed-and-Wobble algorithm (Berger et al., 2006) using the top 90% of foreground and background features; each 9-mer was then assigned the lesser of its two constituent sub-8-mer scores. This procedure, with a score cutoff value of 0.31, optimally separated bound from unbound sequences in a comparison between PBM and published in vitro footprinting data (Gallo et al., 2011). To score preferred binding sites, any 9-mer with a PBM enrichment score that: (1) scored over 0.31 when a HD was bound, (2) scored less than 0.31 with any of the 10 other HDs examined in this study and (3) scored at least 0.05 less for any of the 10 other HDs examined in this study was considered ‘preferred’.

**Analysis of transgenic reporter constructs and embryo staining**

Enhancer regions were synthesized in vitro (Integrated DNA Technologies, Coralville, IA, USA) and subcloned into the reporter vector pWattB-GFP, which was constructed by blunt-end cloning the 3.3 kb *Afel-BstBI* fragment of *pPelican* (Barolo et al., 2000) (containing a *mini-white* gene) into the *AatI* site of pSP73, and the 285 bp *S. lividans* attB site for phage phiC31 (Groth et al., 2004), along with the 2.6 kb *DraIII-HindIII* fragment of phi-H-Stinger (Barolo et al., 2000) (containing an insulated nuclear-localized GFP-reporter construct) in place of the pSP73 polylinker. All constructs were targeted to attP2 (Markstein et al., 2008) with phiC31-mediated integration, and homozygous viable insertion lines were obtained. Whole-embryo immunohistochemistry, in situ hybridization and fluorescent in situ hybridization with tyramide signal amplification (Invitrogen, Carlsbad, CA, USA) followed standard protocols (Halfon et al., 2000).

**Fluorescence-activated sorting of cells from Drosophila embryos and gene expression profiling experiments**

For gene expression microarray experiments, a single-cell population was prepared and GFP-positive cells were purified by flow cytometry from late stage 11/early stage 12 twi-gal4 UAS-2EGFP UAS-msh, *twi-gal4* UAS-2EGFP UAS-slou and *twi-gal4* UAS-2EGFP embryos, resulting in a 2.5- to 3-fold enrichment of mesodermal cells over whole embryos. Total cellular RNA was isolated and labeled in one round of linear amplification and used for hybridization to *Drosophila* Affymetrix GeneChip 2.0 arrays according to methods recommended by the manufacturer. Experimental details of how flow cytometry and microarray data analysis were performed have previously been described (Estrada et al., 2006).

**Lever analysis**

Motifs and gene sets used in the Lever analysis (Warner et al., 2008) are described in detail in supplementary material Table S5. The background gene set included all genes in the genome not annotated as expressed in FCs. Area under the receiver operating characteristic (ROC) curve (AUC) values of the gene set-motif combination pairs were corrected for length bias. Lever was used with the following options: --R 0 --P 0.001 --LP --W 1500 50. FDR calculations were based on 1000 permutations for calculating the Q-value (false discovery rate) of significance of the enrichment statistics (i.e. AUC values).

**Gene ontology (GO) analysis**

Upregulated probe sets were defined as having a Q-value of less than 0.001 which totaled 1058 for *Twi>*msh and 591 for *Twi>*S59. Over-represented GO categories were defined with FuncAssociate2.0 using standard parameters (1000 simulations, significance cutoff=0.05) (Berriz et al., 2009).

**Chromatin immunoprecipitation coupled to quantitative real-time PCR**

A single-cell suspension was prepared from late stage 11 twi-gal4 UAS-2EGFP embryos and fixed in 1.8% formaldehyde. GFP-positive cells were isolated using flow cytometry. Chromatin was prepared, fragmented (200 to 500 bp), and immunoprecipitated with an antibody to Slou (Baylies et al., 1995) according to previously published procedures (Zeitlinger et al., 2007). Duplicate immunoprecipitations were analyzed. Quantitative real-time PCR (qPCR) using SYBR Green (Applied Biosystems) was used to assess the enrichment of genomic fragments which include the Slou-preferred binding sites in the *lbl* and *mih2* enhancers from immunoprecipitated DNA versus non-immunoprecipitated DNA. A genomic region associated with the rp49 gene was included as a control.

**Data access**

Mouse embryonic stem cell and embryonic day 11.5 data were available from the UniProbe database (Robasky and Bulyk, 2011) and from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) under accession number GSE11239. *Drosophila* PBM data are available from GEO under accession number GSE35380, and gene expression microarray data can be obtained from GEO under the accession number GSE27163.
RESULTS

Individual FC genes are differentially responsive to the HD TFs Slou and Msh

To define candidate transcriptional targets of Slou and Msh, we determined the genome-wide mRNA expression profiles of primary mesodermal cells purified from embryos in which Slou or Msh was overexpressed at a developmental time when FCs are specified. We previously used a similar approach to predict hundreds of novel FC expression patterns, a large number of which were independently verified in vivo (Estrada et al., 2006). These studies revealed that there were 1051 and 327 genes that exhibited statistically significant (with $Q<0.1$) 1.5-fold and 4-fold differences in expression in the Slou gain-of-function experiment, respectively. Similarly, for the Msh gain-of-function experiment, there were 1525 (1.5-fold differences) and 380 (4-fold differences) genes that exhibited statistically significant differences in expression. Next, all genes in the genome were ranked based on their responses to ectopic Slou or Msh, and known FC genes were mapped onto these distributions such that their responsiveness to both FCI-HD TFs could be compared (Fig. 1A; supplementary material Fig. S1). Different FC genes were activated, repressed or unaffected by one or both of these TFs, findings that were validated by whole-embryo in situ hybridization (Fig. 1B-D; supplementary material Table S1). Seventeen out of 22 (77.3%) and 16 of 26 (61.5%) of the tested FC genes that were found by microarray-based expression profiling to be Slou- or Msh-responsive, respectively, were verified by in situ hybridization to have these predicted patterns (see supplementary material Table S1). Furthermore, analysis of over-represented Gene Ontology annotation terms among the differentially expressed genes revealed that these FCI-HD TFs regulate both upstream (e.g. signaling molecules and transcription factors) and downstream (terminal differentiation gene products such as muscle structural and extracellular matrix proteins) components of the myogenic regulatory network (see supplementary material Table S1). Taken together, these results establish that individual FC genes are differentially responsive to FCI-HD TFs, with Slou and Msh targeting both upstream and downstream components of the myogenic regulatory network.

Slou and Msh predominantly recognize DNA sequences containing a TAAT core but also exhibit preferences for variant binding sites that are unique to each HD

The differential responsiveness of individual FC genes to overexpression of Slou or Msh suggests that these FCI-HD TFs exhibit functional specificity in regulating FC enhancers. To better understand the molecular mechanisms underlying this specificity, we determined the in vitro DNA-binding preferences of Slou, Msh and eight other mesodermally expressed Drosophila HD TFs using high-resolution universal protein-binding microarrays (PBMs) (see supplementary material Table S2 for details of clones used) (Berger et al., 2006). Previously, all possible 8-mer binding sites for mouse HD TFs were used to construct position weight matrix (PWM)-based motif representations using the PRIORITY algorithm and corresponding graphical sequence logos (Narlikar et al., 2006) (Fig. 2A; see supplementary material Table S4).

Fig. 1. Differential responsiveness of individual FC genes to Slou or Msh overexpression. (A) mRNA expression profiles of mesodermal cells overexpressing Slou or Msh under control of Twist (Twi)-Gal4 (Twi>Slou and Twi>Msh, respectively) compared with wild type. On each axis, genes are ranked by Bayesian t-statistic (Choe et al., 2005) from the most likely upregulated relative to wild type (lower left corner) to the most likely downregulated. Responses of previously documented FC (red) and other (gray) genes are shown, ranked from the most likely upregulated (lower left) to the most likely downregulated (upper right). ‘Other genes’ include genes known to be not expressed in FCs and genes not tested for expression in FCs. (B-D) Expression of Nidogen (Ndg) (B-D) mRNA in wild-type (B), Slou overexpressing (C) and Msh-overexpressing (D) stage 12 embryos. Arrows indicate groups of cells that have increased expression of Ndg, which occurs in different somatic mesodermal cells in the Twi>Slou and Twi>Msh embryos. See also supplementary material Table S1.

The PBM data indicate that a large class of Drosophila HDs – including members of the Hox subclass (Ubx, AbdB), Slou, Msh, Eve and Lbl – primarily recognize sequences with the canonical TAAT core sequence, in general agreement with prior studies of Drosophila and mouse HD DNA-binding specificities (Berger et al., 2008; Noyes et al., 2008) (supplementary material Fig. S2). In addition, some HD subclasses – including Six, Paired HD and certain members of the NK subclass, Tin and Bap – exhibit DNA-binding profiles that are distinct from this canonical sequence (Fig. 2A). Furthermore, the present PBM results show that many of the HD TFs that bind predominantly to TAAT-containing sequences also recognize atypical binding sites that are unique to each TF. For example, Slou and Msh each recognize a small set of sequences (AbdB), paired HD [Paired-type homeobox 1 (Ptx1)] and Six (Six4) families of HD TFs, as well as Even skipped (Eve) and Msh. Two-dimensional hierarchical clustering analysis of the PBM enrichment scores (E-scores) of all 9816 ungapped 9-mers (see supplementary material Table S3) that were bound by at least one HD TF with E-score$$>0.31$$ is shown in Fig. 2A (see Materials and methods for details of how binding thresholds were determined). In order to represent DNA-binding specificities, we constructed position weight matrix (PWM)-based motif representations using the PRIORITY algorithm and corresponding graphical sequence logos (Narlikar et al., 2006) (Fig. 2A; see supplementary material Table S4).
that are not bound by any other of the examined *Drosophila* HDs (Fig. 2A,B; supplementary material Table S3). These Slou- and Msh-preferred sequences are also preferentially bound by the orthologous mouse HDs (Fig. 2A). To visualize these distinctive DNA-binding specificities, we constructed motifs from the sequences preferentially bound by Slou and Msh (Fig. 2D,E; supplementary material Tables S1, S2). As many of the FCI-HDs (Slou, Msh, Lhx2, Eve and Lbl) bind similar sequences and exert related regulatory roles in each of the FCS in which they are expressed, we also created motifs from these shared, or ‘common’, binding sequences (Fig. 2C; supplementary material Tables S4, S5). These data show that *Drosophila* FCI-HDs have both shared (HD-common) and individual sequence preferences (HD-preferred) that differ markedly from each other.

**The cell-specific effects of Slou are mediated by single Slou-preferred DNA-binding sequences**

To understand the molecular basis for the specificity of FCI-HD TFs, we asked whether Slou-preferred binding sites are responsible for cell type-specific gene regulation by this HD TF. To test this hypothesis, we first identified conserved Slou-preferred DNA-binding sequences in previously characterized enhancers from Slou-responsive FC genes (Fig. 1; supplementary material see Fig. S3) (Halfon et al., 2000; Capovilla et al., 2001; Halfon et al., 2002b; Philippakis et al., 2006). We focused our functional studies of Slou-preferred binding sites by choosing FC enhancers associated with *lbl* and *mib2* (Philippakis et al., 2006), which represent upstream myogenic regulatory and downstream muscle differentiation genes, respectively (Busser et al., 2008). These genes were shown to be responsive to ectopic Slou using whole-embryo in situ hybridization in spite of both being scored as non-responsive in the Slou gain-of-function microarray experiment (see supplementary material Table S1). These discrepancies probably reflect the limited sensitivity of microarray-based expression profiling of minority members of heterogeneous cell populations, and underscore the importance of independently validating microarray results at single-cell resolution in intact embryos.

Slou and Lbl are expressed in mutually exclusive patterns in adjacent FCs and adult muscle precursors in the lateral embryonic mesoderm (Fig. 3C,J) (Jagla et al., 1998; Knirr et al., 1999). Slou activity is required in the two Slou-expressing FCs that form the muscles lateral oblique 1 (LO1) and ventral transverse 1 (VT1). In *slou* mutant embryos, the loss of these two muscles is associated with Lbl derepression and a duplication of the segment border muscle, which derives from the normal Lbl-expressing FC and the Lbl-expressing adult muscle precursors (Fig. 3E,K) (Knirr et al., 1999). Such cross-repressive interactions among FCI-HD TFs are thought to maintain the individual localized expression of these genes (Jagla et al., 2002; Lacin et al., 2009).

We asked whether Slou repression of *lbl* in the segment border muscle is mediated by Slou-preferred binding sites in the *lbl* FC enhancer. To investigate this issue, we first showed with loss-of-function (Fig. 3E) and gain-of-function (Fig. 3F) genetic experiments that the effects of Slou on endogenous *lbl* expression are mirrored at the level of the isolated enhancer in transgenic reporter assays. Both the *lbl* gene and *lbl-lacZ* reporter are normally expressed in three mesodermal cells, the segment border muscle and two adult muscle precursors (Fig. 3D); *slou* mutants, however, show an increase in both *lbl* gene and enhancer-regulated reporter expression in five cells (Fig. 3E). *slou* gain of function elicits the reciprocal effect of extinguishing both *lbl* gene and *lbl* enhancer-driven reporter activity within the mesoderm (Fig. 3F,L). Taken together, these results confirm that the isolated *lbl* enhancer is repressed by Slou.

The *lbl* FC enhancer (Philippakis et al., 2006) contains over 20 separate sites capable of binding Slou, including eight sequences that can bind all FCI-HD TFs (see supplementary material Fig. S3A). In addition, there is a single, evolutionarily conserved sequence that is preferentially bound by Slou (Fig. 3A,B). To investigate the potential role of Slou in regulating this *lbl* enhancer, we used chromatin immunoprecipitation followed by quantitative real-time polymerase chain reaction (ChIP-qPCR) to show that a genomic sequence that includes this Slou-preferred binding site is bound by Slou in purified primary mesodermal cells (see supplementary material Fig. S4). This result establishes that Slou binds to the *lbl* FC enhancer in vivo, and is consistent with the possibility that Slou directly regulates this element.

To test whether the conserved Slou-preferred motif in the *lbl* FC enhancer mediates the previously described repressive activity of Slou on *lbl* expression, we mutated this sequence such that Slou binding is significantly reduced, as judged by the PBM E-score of the mutant site (Fig. 3A), and a crucial nearby T-box-binding site is unaffected (Y. Kim, B.W.B. and A.M.M., unpublished). A GFP
reporter driven by the wild-type *lbl* enhancer is expressed in three Lbl-positive cells (Fig. 3G) and is not co-expressed with Slou (Fig. 3H). However, mutagenesis of the Slou-preferred binding site in the *lbl* enhancer results in derepression of the reporter in two nearby Slou-expressing FCs (Fig. 3I,M), the same cells in which endogenous *lbl* is derepressed in *slou* mutant embryos (Fig. 3E). These results suggest that a single Slou-preferred binding site is capable of mediating the cell-specific effects of Slou in individual embryonic cells, consistent with the known activity of this FCI-HD TF.

To assess the role of a second Slou-preferred binding site, we tested the function of an independent motif of this class in the FC enhancer associated with *mib2*, a gene that encodes a putative E3 ubiquitin ligase involved in maintaining myotube integrity (Nguyen...
et al., 2007; Carrasco-Rando and Ruiz-Gomez, 2008). This experiment also provided the opportunity to assess the function of Slou-preferred sites in regulating downstream targets of muscle differentiation. We previously characterized an enhancer from the mib2 gene that is active in all mib2-expressing FCs (Fig. 4C) (Philippakis et al., 2006), a subset of which also expresses Slou (Fig. 4F,J). The latter cells correspond to the same Slou-expressing FCs that exhibit reporter derepression when the Slou-preferred site in the lbl FC enhancer is inactivated (muscles LO1 and VT1; Fig. 3M). slou mutant embryos show a loss of both endogenous mib2 and mib2 enhancer-driven reporter expression in muscle LO1 and VT1 FCs (Fig. 4D,K), whereas slou gain of function (Fig. 4E,L) induces ectopic activity of both the endogenous mib2 gene and mib2 enhancer in adjacent mesodermal cells that normally do not express mib2. These results support the model that Slou directly activates the mib2 enhancer in a specific subset of FCs.
Similar to the \( \textit{ lbl} \) FC enhancer, Slou also binds in vivo to the \( \textit{ mib2} \) FC enhancer, as determined by ChIP-qPCR (see supplementary material Fig. S4). Although the \( \textit{ mib2} \) enhancer contains multiple sequences that can bind both Slou and other HD TFs, it – like the \( \textit{ lbl} \) enhancer – possesses one evolutionarily conserved Slou-preferred binding site (Fig. 4A,B; supplementary material Fig. S3B). To test the potential function of this Slou-preferred site, we mutated it in an otherwise wild-type \( \textit{ mib2} \) enhancer such that Slou can no longer bind (Fig. 4A). This mutation caused an attenuation of \( \textit{ mib2} \) reporter activity in FCs LO1 and VT1 that normally express both \( \textit{ slou} \) and \( \textit{ mib2} \) at stage 11 (compare Fig. 4F with 4G), an effect that is markedly increased as the FCs fuse with FCMs to form muscle precursors at a later developmental stage (compare Fig. 4H with 4I). Of note, the Slou-preferred binding site mutation did not alter \( \textit{ mib2} \) enhancer activity in any other FCs, as expected for a site that mediates the effects of this particular FCI-HD TF. Moreover, these cell-specific findings for the cis mutation of the Slou-preferred binding site in the \( \textit{ mib2} \) enhancer precisely correlate with the \textit{ trans} effect of \( \textit{ slou} \) loss-of-function on both endogenous \( \textit{ mib2} \) expression and \( \textit{ mib2} \) enhancer activity (Fig. 4D). These results are summarized schematically in Fig. 4J-M. Collectively, these studies show that the HD-binding preferences of an FCI-HD TF can mediate distinct biological effects in individual embryonic cells, establishing a previously uncharacterized mechanism underlying HD-specific functions.

**FCI-HD-preferred binding sequences are over-represented within putative CRMs of FCI-HD-responsive genes**

Having demonstrated the functional significance of Slou-preferred binding sites in two FC enhancers, we next asked whether FCI-HD-preferred binding sequences are more generally involved in the regulation of FC gene expression. We reasoned that if FCI-HD-preferred sites confer transcriptional specificity to FC enhancers, then these sequences should be over-represented in the noncoding regulatory regions of the correspondingly responsive FC target genes.

To examine this possibility, we used a computational algorithm called Lever (Warner et al., 2008) to evaluate the enrichment of Slou- or Msh-preferred binding sequences in combination with DNA-binding motifs for Pointed (Pnt), Twist (Twi) and Tin – TFs with known FC regulatory functions (Halfon et al., 2000; Halfon et al., 2002b; Philippakis et al., 2006) – within putative CRMs identified in the noncoding regions of Slou- or Msh-responsive genes. The gene sets used in these analyses were composed of 44 Slou-responsive, 31 Msh-responsive, 12 Slou-non-responsive and 14 Msh-non-responsive genes (supplementary material Table S5).

This analysis revealed that predicted CRMs associated with Slou-responsive FC genes are equally enriched for Slou-preferred sites, together with Pnt and Twi, as the previously delineated combination of FC regulators, Pnt, Twi and Tin (Fig. 5A; supplementary material Table S5) (Philippakis et al., 2006). Importantly, no such enrichment of Slou-preferred sites was observed for FC genes that are not responsive to Slou (Fig. 5B). In addition, Msh-preferred sites are also enriched along with Pnt and Twi sites within putative CRMs of Msh-responsive FC genes (Fig. 5C; supplementary material Table S5), but no enrichment is seen among FC genes that are not responsive to Msh (Fig. 5D; supplementary material Table S5). Moreover, when the HD-preferred motifs are exchanged for Pnt or Twi sites – as opposed to Tin sites – these combinations are also discriminatory for appropriately HD-responsive FC genes (supplementary material Fig. S5A,B), further supporting transcriptional co-regulation through HD-preferred motifs. Interestingly, HD-common sites are also enriched along with Pnt and Twi sites within putative CRMs associated with Slou- or Msh-responsive FC genes (supplementary material Fig. S6A,B, Table S5). This latter finding is consistent with HD-common motifs that mediate the activities of a broad spectrum of HDs, including members of the Hox family (Capovilla et al., 2001; Enriquez et al., 2010). Nevertheless, both our experimental and computational results demonstrate that HD-preferred motifs contribute significantly to the transcriptional specificity of FCI-HDs.

**Fig. 5. Enrichment of Slou-preferred and Msh-preferred binding sites located within putative CRMs in the noncoding regions of Slou- and Msh-responsive FC genes.** (A-D) Receiver operating characteristic (ROC) curves showing the discrimination of Slou-responsive (A), Slou-nonresponsive (B), Msh-responsive (C) and Msh-non-responsive (D) FC genes by the indicated AND combinations of Pnt, Twi, Tin, Slou-preferred and Msh-preferred binding motifs (see supplementary material Table S5 for the entire set of Lever results). The area under the ROC curve (AUC) for each gene set and motif combination is shown. Foreground gene sets are listed in supplementary material Table S5 and the background was generated as described in the Materials and methods. Slou-preferred and Msh-preferred sites are over-represented together with the known FC regulators Pnt and Twi, in the noncoding regions of Slou-responsive or Msh-responsive FC genes, respectively. This effect does not occur with FC genes that are known not to be Slou or Msh responsive.
Molecular specificity of a homeodomain binding preferences, that determines the activity of an enhancer. To possible that it is the location of the sites, rather than their particular sequence, order and spacing of TF-binding sites are known to be crucial for enhancer function (Ludwig et al., 2000; Senger et al., 2010). Thus, it remains be a widespread mechanism underlying the regulatory specificity of other FC co-regulatory TFs among the correspondingly HD-preferred and Msh-preferred motifs are enriched along with two S5A,B). Collectively, these results led us to conclude that Slou-preferred site to a sequence bound by all HDs (Fig. 6C). However, this new Slou-preferred site was also incapable of mediating the normal function activity by substituting the wild-type Slou-preferred site for an alternative sequence that is also preferred by Slou when compared with other HDs (Slou-pref-alt; Fig. 6A). Nevertheless, the Lever analysis demonstrated that each of the over-represented motif combinations is only partially able to discriminate among the members of the included gene sets, a finding that most probably reflects the heterogeneity of TF combinations that regulate individual members of these co-expressed genes. Consistent with this idea, no combination of TF-binding sites that included Slou- and Msh-preferred motifs was able to as effectively distinguish among a much larger collection of FC genes that is not biased towards being responsive to Pnt, Slou or Msh (supplementary material Fig. S7). Similarly, the heterogeneity of gene expression and combinatorial regulation amongst the individual genes that make up these gene sets probably explains the inability to see greater enrichment of Pnt+Twi+HD-preferred motifs when compared with Pnt+Twi motifs alone (supplementary material Fig. S5C,D). In this context, it is also important to note that the more constrained three-way ‘AND’ combination is as applicable to the gene set in question as the combination that contains only two known FC co-regulatory TFs. Because of the statistical constraint associated with increasing the combinatorial specificity through the addition of a third motif, we focused on comparing the previously delineated three-way ‘AND’ combination of Pnt+Twi+Tin with three-way ‘AND’ combinations involving two known FC co-regulatory motifs together with HD-preferred motifs (Fig. 5A,C; supplementary material Fig. S5A,B). Collectively, these results led us to conclude that Slou-preferred and Msh-preferred motifs are enriched along with two other FC co-regulatory TFs among the correspondingly HD-responsive FC gene sets. In summary, both our computational and experimental results suggest that binding to HD-preferred sites may be a widespread mechanism underlying the regulatory specificity of FCI-HD TFs (Capovilla et al., 2001; Enriquez et al., 2010).

**The particular nucleotide sequence of a Slou-preferred binding site is crucial for the regulatory activity of Slou**

The sequence, order and spacing of TF-binding sites are known to be crucial for enhancer function (Ludwig et al., 2000; Senger et al., 2004; Panne et al., 2007; Swanson et al., 2010). Thus, it remains possible that it is the location of the sites, rather than their particular binding preferences, that determines the activity of an enhancer. To address this issue, we performed site specificity swaps in an otherwise wild-type mib2 enhancer. We first changed the specificity of the previously identified functional Slou-preferred site for one that can bind all FCI-HD TFs (HD-common, Fig. 6A). We reasoned that if only the location of the Slou-binding site is crucial, then exchanging it for another site that can also bind Slou should have no effect on transcriptional activity. However, substituting the Slou-preferred site for a HD-common sequence caused an attenuation of the enhancer in Slou-expressing muscle precursors LO1 and VT1 (Fig. 6C) when compared with the wild-type enhancer (Fig. 6B). This result is equivalent to that occurring with mutation of the same site such that it cannot bind Slou at all (Fig. 4I). Thus, simply the ability to bind Slou at a particular location in an enhancer is insufficient to mediate the regulatory activity of this FCI-HD TF. Rather, the actual sequence of the HD-binding site appears to contribute to TF function in this context.

We have extended these analyses by asking whether a different Slou-preferred binding site would be sufficient for mib2 enhancer activity by substituting the wild-type Slou-preferred site for an alternative sequence that is also preferred by Slou when compared with other HDs (Slou-pref-alt; Fig. 6A). However, this new Slou-preferred site was also incapable of mediating the normal function of the mib2 enhancer in Slou-expressing muscle precursors LO1 and VT1 (Fig. 6D), the same effect as produced by either completely inactivating Slou binding (Fig. 4I) or changing the Slou-preferred site to a sequence bound by all HDs (Fig. 6C). Collectively, these results indicate that the precise nucleotide sequence of a Slou-preferred site is crucial for the function of this HD TF, a conclusion that is further supported by the high degree of evolutionary conservation of the two Slou-preferred binding sites whose functions we have validated (Fig. 3B, Fig. 4B).

**DISCUSSION**

Here, we used an integrated genomics approach to interrogate the molecular mechanisms of action of a subset of identity HD TFs that have been proposed to control the unique gene expression programs of muscle FCs (Baylies et al., 1998; Tixier et al., 2010). We first showed that FC genes are differentially responsive to Slou and Msh, which suggests functional specificity in the regulation of FC genes by these FCI-HD TFs, and is consistent with the known effects of these TFs on muscle cell fates (Lord et al., 1995; Nose et
al., 1998; Knirr et al., 1999; Tixier et al., 2010). PBM assays defined the specific sequences that are bound by these HDs, revealing that the majority of binding sites contain TAAT core sequences that are shared by all FCI-HD TFs, but that each HD also binds to a small number of unique, atypical sequences. In each of the two Slou-responsive FC enhancers, we found that the transcriptional specificity of Slou is mediated by its binding to a single motif that is preferred by Slou and that is not bound by other mesodernally expressed HDs that were examined. Genome-wide computational studies provide further evidence for the potential importance of HD-preferred binding sites within the myogenic network of FC genes. Nevertheless, mesodermal HD proteins do not exclusively act through these atypical motifs as Hox TFs have been documented to regulate other muscle enhancers through HD-common binding sites (Capovilla et al., 2001; Enriquez et al., 2010).

Our data show that the diversity of HD-binding preferences may confer the cell-specific effects of HDs by controlling which member of a related TF family is able to bind to and function at a particular site in a given CRM. This feature of enhancers may be especially important in developmental contexts where multiple family members that have different activities are co-expressed, resulting in potential competition for TF binding to shared sites. Such would be the case for FCI-HD and Hox TFs, both of which participate in the myogenic program but with distinct regulatory functions (Michelson, 1994; Baylies et al., 1998). Given the high level of conservation of these individual binding sites, there appears to be strong evolutionary selection for a particular HD-preferred sequence, a process that may be driven by the requirement for maintaining essential interactions with other TFs in a given regulatory context. For example, the DNA specificity of Hox HDs is known to be modified by interactions with co-factors such as the PBC and MEIS subclasses of TALE HD proteins (Moens and Selleri, 2006; Mann et al., 2009). Although there is currently no evidence that these co-factors interact with Drosophila FCI-HD TFs, PBC proteins are thought to interact with similar classes of vertebrate TFs (In der Rieden et al., 2004). Other forms of collaboration with FCI-HD TFs may also occur, including TF heterodimerization (Landschulz et al., 1988; Grove et al., 2009), cooperative interactions with other co-factors (Mahlafey, 2005), or formation of multi-protein complexes of signal-activated and tissue-restricted TFs that have convergent effects on mesodermal gene expression (Busser et al., 2008; Mann et al., 2009).

The existence of functional HD-preferred binding sites raises the issue of how such sequences mediate their regulatory effects, especially as our site specificity swap experiments revealed that the particular nucleotide sequence of a Slou-preferred site appears to be crucial for its function. It is possible that the specific sequences of HD-preferred DNA-binding sites form unique structures that are recognized by some HDs and not by others in certain contexts (Joshi et al., 2007). Alternatively, binding to such sequences may induce a distinct protein conformation that is essential for enabling the HD to activate or repress the corresponding CRM, for example, by facilitating interactions with co-factors or other regulatory proteins (Leung et al., 2004).

Although our results support a central role for sequences preferred by one particular HD TF, the complexity of FC gene expression makes it likely that additional HD input occurs through sequences preferred by other co-expressed HDs. As many FCI-HD TFs have mutually exclusive expression patterns (Tixier et al., 2010), a DNA binding site specific to, for example, Slou, Msh and Lb will be used by each TF in the cells in which they are differentially expressed.

Thus, the HD-binding profile of enhancers should be re-examined as a collection of sequences with the ability to bind one or many HDs and where the functions of those sites in individual cells are dependent on the expression of the corresponding TF. The cumulative effects of these cell-specific binding events will then direct the discrete regulatory responses of the target genes.

In conclusion, we present a previously uncharacterized mechanism by which different members of the FCI-HD class of TFs determines the unique genetic programs of single myoblasts in a developing embryo. This regulatory process involves the selective recognition of particular DNA sequences by individual HDs. The ability of distinct DNA-binding sequences to generate an additional level of regulatory complexity may be of general importance in the architecture of transcriptional networks and in the evolution of TF families and CRMs. Finally, the approach used here provides a general strategy for investigating similar issues about the specialized roles played by individual members of other TF families, and how those functions may be precisely encoded in the cis-regulatory language of the genome.

Acknowledgements
We thank N. Perrimon, M. Frasch, A. Nose, K. Jagla and M. Baylies for providing fly strains and antibodies; A. Vedenko, E. Lane and C. Sonnenbrot for technical assistance; D. Hill and K. Salehi-Ashtiani (Center for Cancer Systems Biology, Dana Farber Cancer Institute) for assistance with Gateway cloning; R. Gordän for advice in using the PRIORITY algorithm; and M. Knepper, J. Zhu, B. Oliver, A. J. M. Walhout, R. Adelstein and R. Maas for comments on the manuscript. N. Raghavachari (Genomics Core Facility, NHLBI Division of Intramural Research) and R. Steen (Biopolymers Facility, Harvard Medical School) provided help with microarray experiments, P. McCoy (Flow Cytometry Core Facility, NHLBI Division of Intramural Research) was instrumental in performing cell purifications, and T. Ni and J. Zhu (DNA Sequencing Core Facility, NHLBI Division of Intramural Research) offered invaluable advice on the ChIP-qPCR experiments.

Funding
This work was funded by National Institutes of Health/National Institutes of General Medical Sciences (NH/NIGMS) [U01 GM076603 to M.L.B.], by NH/National Human Genome Research Institute (NHGRI) [R01 HG005287 to M.L.B.], by the National Heart, Blood and Lung Institute (NHLBI) Division of Intramural Research (A.M.M.), by a NIH Training Grant [5 T32 GM007748-31 to L.S.], and by a NIH NRSA [1 F32 GM090645-01A1 to L.S.]. Deposited in PMC for immediate release.

Author contributions
A.M.M., B.W.B. and M.L.B. designed the overall research project and wrote the manuscript. S.A.J. performed the Lever computational analyses. B.W.B., B.Z. and L.S. cloned TFs and purified protein for PBM assays. L.S. and M.F.B. performed the ChIP-qPCR experiments. A.M.M., B.W.B. and M.L.B. wrote the manuscript. B.W.B., A.M.M., B.W.B. and M.L.B. provided help with microarray experiments, P. McCoy (Flow Cytometry Core Facility, NHLBI Division of Intramural Research) was instrumental in performing cell purifications, and T. Ni and J. Zhu (DNA Sequencing Core Facility, NHLBI Division of Intramural Research) offered invaluable advice on the ChIP-qPCR experiments.

Competing interests statement
The authors declare no competing financial interests.

Supplementary material
Supplementary material available online at http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.077362/-/DC1

References


