Conservation and Innovation in the DUX4-family Gene Network

Jennifer L. Whiddon

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Stephen J. Tapscott, Chair

Michael Emerman

Edith H. Wang

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Conservation and Innovation in the DUX4-family Gene Network

Jennifer L. Whiddon

Chair of the Supervisory Committee:
Professor Stephen J. Tapscott
Department of Neurology

Facioscapulohumeral dystrophy (FSHD) is caused by the mis-expression of the DUX4 transcription factor in skeletal muscle. Animal models of FSHD have been hampered by incomplete knowledge of the conservation of the DUX4 transcriptional program in other species besides humans. We demonstrate that both mouse Dux and human DUX4 activate repetitive elements and genes associated with cleavage-stage embryos when expressed in muscle cells of their respective species. Specifically, mouse DUX activated the transcription of MERV-L retrotransposons and genes such as Gm4340 (a.k.a. Gm6763), Slc34a2, Tcstv1/3, Tdpoz 3/4, Usp17la-e and Zfp352, all of which are characteristic of mouse two-cell embryos. Human DUX4 activated transcription of orthologs of mouse DUX-induced genes, including orthologs of genes characteristic of mouse two-cell embryos. Despite functional conservation, we found that the binding motifs of mouse DUX and human DUX4 have diverged. To better understand the extent of conservation or divergence between these factors and to assess current mouse models of FSHD, we expressed human DUX4 in mouse muscle cells. In this context, human DUX4 maintained modest activation of early embryo genes driven by conventional promoters, but did not activate MERV-L-promoted genes. These and additional findings indicate that the ancestral DUX4-factor regulated a cleavage-stage embryo program driven by conventional promoters, whereas divergence of the DUX4/Dux homeodomains correlates with their retrotransposon specificity. These results provide insight into how species balance conservation of a core developmental program with innovation at retrotransposon promoters and provide a basis for developing crucial animal models of FSHD.

TABLE OF CONTENTS

List of Figures

Chapter 2 Figures:

Chapter 3 Figures:

Chapter 4 Figures:

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Stephen for being an excellent mentor, equal parts cheerleader and taskmaster. I also need to thank all members of the Tapscott Lab, past and present, for excellent discussions and camaraderie. I also appreciate Harmit and all members of his lab for allowing me to interlope in their lab meetings and offering essential feedback and advice. Finally, I want to thank the remaining members of my entire committee for challenging me to grow as a scientist.

Chapter 1. Introduction

Facioscapulohumeral Muscular Dystrophy

Facioscapulohumeral Muscular Dystrophy (FSHD) is an autosomal dominant genetic disorder that affects an estimated 4-12 per 100,000 people worldwide (1-4). It is a progressive disease that first leads to muscle weakness in the face, shoulders and upper arms of affected individuals. The age at which symptoms begins varies widely between affected individuals from infancy to middle age, but it often begins in a person's twenties(5). This disease greatly affects the quality of life of affected individuals as it impairs facial expressions, lifting of the arms above shoulder level and one of every four affected individuals eventually requires the assistance of a wheelchair(6). Although there are no treatments for this disease currently, progress has been made in recent years towards understanding the mechanism of FSHD.

Although the DUX4 gene was initially discounted as irrelevant to FSHD, it is now widely accepted as the causative agent. FSHD was first described in 1885(7), but it was not until 1990 that Wijmenga *et al.* discovered the genetic basis of the disease(8). Shortly thereafter, Hewitt *et al.* described the genetic lesion that occurs in FSHD-affected individuals, namely a contraction in the number of D4Z4 macrosatellite repeat units on chromosome 4, and that each repeat unit carried the open reading frame (ORF) of the double homeobox 4 (officially called DUX4; called hDUX4 in this manuscript for clarity) gene that was devoid of introns within its coding sequence(9-11). The atypical structure of the hDUX4 gene led investigators to suspect dysregulation of genes near D4Z4 as the causative agent of FSHD, as opposed to the hDUX4 gene within each D4Z4 repeat unit(12). Another reason that hDUX4 was discarded initially as a candidate causative gene is that its mRNA and protein were difficult to detect in patient biopsies and cultured myoblast cell lines derived from affected individuals(13, 14). The first hint that hDUX4 might actually encode a functional protein came from evolutionary studies that revealed that the hDUX4 ORF has been conserved since the last common ancestor of primates (~74 million years) and a similar gene has been conserved in Afrotherians since their last common ancestor (~100 million years; e.g. elephant, hyrax, tenrec)(15). Finally, researchers found individuals with unique recombination events that provided the genetic association between certain 4qA haplotypes (4qA159, 4qA161, 4qA163, 4qA166H, and 4qA168) and FSHD – simply, all affected individuals had one of the 5 haplotypes listed above, all of which encode a polyadenylation signal immediately downstream of the final repeat of the D4Z4, while individuals remained unaffected despite a contracted D4Z4 array if they lacked the polyadenylation signal on the 12 other haplotypes known in the population(16). Subsequent work showed that distal polyadenylation signals were used to stabilize the hDUX4 transcript in adult testes, but for an unknown reason, these distal polyadenylation signals are not used to stabilize hDUX4 transcripts in the skeletal muscles of individuals affected by FSHD(10).

The demonstration that hDUX4 was the causative agent of FSHD was a critical step forward because it unified the FSHD field behind a single mechanism. After this resources were consolidated and researchers focused on modulating hDUX4 levels or activity in skeletal muscle as a key path towards a treatment. Several models of FSHD were developed based on hDUX4 expression, with the ultimate goal of generating a system for screening candidate drug therapies. Current models of FSHD span the gamut from hDUX4 overexpression in cultured myoblasts to transgenic mice carrying long or short D4Z4 arrays (reviewed in: (17)). One motivation for my thesis work was to investigate the extent to which an endogenous mouse gene (officially called *Dux*; called *mDux* in this manuscript for clarity) could serve as a model for FSHD.

Double homeobox gene evolution

In order to understand the rationale for using *mDux* as a model of FSHD, it is essential to understand the evolutionary history of the DUX genes. Double homeobox genes are only found in Eutherians (placental mammals) and a putative single homeodomain ancestor has been identified in species outside of Eutherians (opossum, lizard and chicken) using synteny and sequence similarity, leading to the hypothesis that recombination between two tandemly duplicated single homeodomain genes created the DUX genes (18). Following homeodomain duplication there was an expansion of DUX genes that created three paralogues: DUXA, DUXB and DUXC. Very little is known about most of these genes aside from the disease-associated hDUX4. hDUX4 was created in a retroposition event, during which a processed mRNA was integrated into the genome using enzymes encoded by LINE1 retroelements(19). In contrast to most other retrogenes, hDUX4 actually has an intron in its 3' untranslated region(10), which is thought to have been acquired in an uncommon intron acquisition event following the initial retroposition, which created a gene completely free of introns.

Two lines of evidence support the hypothesis that hDUX4's intron-containing ancestor was a DUXC gene. First, hDUX4 and DUXC from species that retain intron-containing DUXC (e.g.

Laurasiatherians) share a stretch of homology at the extreme C-terminus that includes one or more LxxLL motifs, thought to mediate protein-protein interactions. Neither DUXA nor DUXB genes in any species contain C-terminal LxxLL motifs. Second, a maximum likelihood tree made with the concatenated homeodomains from various species show a distinct DUXA clade, a distinct DUXB clade and a mixed hDUX4/DUXC clade(18), indicating that hDUX4 is more similar in sequence to DUXC genes than to DUXA or DUXB genes (if this tree were great support for hDUX4 and DUXC being orthologs, the mixed clade would have species-tree topology, but it doesn't). Another piece of evidence used to determine orthologous relationships between genes is synteny. If hDUX4 and DUXC were true orthologs, they should exist in syntenic genomic locations. However, anchor genes that define the ancestral DUXC have not been established because DUXC genes are present on different chromosomes with different neighboring genes in various species. One possible explanation for the lack of synteny within DUXC genes could be that the ancestral DUXC locus was present in a dynamic region of the genome that was prone to chromosomal rearrangements. In support of this "dynamic region" explanation is the observation that DUXC in cow, pig and dog are all subtelomeric, which is known to be unstable. Consistently, DUX4 genes in primates are also subtelomeric and a chromosomal breakpoint was identified that perturbed synteny between primates and non-primates(20). A further complication to synteny analyses between hDUX4 and DUXC is that hDUX4 retroposed, which is thought to involve integration into a new locus. Despite equivocal synteny analyses, evidence described above provides reasonable support for the hypothesis that hDUX4 retroposed from an ancestral DUXC mRNA.

In addition to hDUX4, other DUXC retrogenes have been identified in rodent and Afrotherian (e.g. elephant, hyrax, tenrec) lineages. Orthologues of hDUX4 have been found in many primate species and also in an outgroup of primates: treeshrews(15). Mice and rats have DUXC retrogenes, but it is unclear whether the rodent DUXC retrogenes were created in the same retroposition event that created the DUXC retrogenes of treeshrews and primates – it has been argued both ways. One piece of evidence that has been used to support an independent retroposition is the maximum likelihood tree created with concatenated homeodomains in which the mDUX homeodomains are placed outside of the DUXC/DUX4 mixed clade(15). Stronger support of independent retroposition events would be the identification of an intron-containing DUXC genes in a lineage that separates treeshrews and rodents, but no such gene has been identified.

Finally, there is evolutionary evidence to support the hypothesis that all DUXC genes and retrogenes are functional homologs. No species has been identified that contains both a DUXC gene and a DUXC retrogene. Additionally, all DUXC genes and retrogenes surveyed have an uncommon head-to-tail multi-copy array structure. These facts have been used to argue that the retrogenes overwrote the parental intron-containing locus and that DUXC genes and retrogenes could perform the same function(20). A critical barrier to developing non-primate models of FSHD has been the belief that the DUX4 gene is unique to primates. If DUXC genes and retrogenes are indeed functional homologs, their mis-expression in skeletal muscle might cause disease via the same mechanism, which would support using various non-primate organisms (e.g. mice, dogs) and the DUXC gene or DUXC retrogene from those species to model FHSD.

Mouse Double Homeobox Gene

Although many details of the evolutionary history of the DUX genes remain unknown, hDUX4 is not the only DUXC retrogene, which leads to the intriguing question as to the extent to which various DUXC genes and retrogenes perform the same function. Since mouse is a commonly used model system and it has a DUXC retrogene, I decided to start my investigation there. In addition to shedding light on the evolutionary history of these genes, if mDUX and hDUX4 are functional homologs mDUX could be a useful tool in FSHD studies and drug screens.

Prior to my graduate studies, very little was known about mDUX. One paper discovered mDUX and laid the initial groundwork for future studies(15). They found that mDUX exists in a multi-copy tandem array, but that the array is only found on one chromosome (chr10) as opposed to arrays on multiple chromosomes and dispersed individual repeat units, as is the case with hDUX4 in the human genome. They also found that the number of repeats is polymorphic between outbred mice strains, which is a clear parallel to the polymorphic D4Z4 array lengths observed in humans and of relevance to FSHD as arrays of less than 10 copies leads to disease. Reverse transcription polymerase chain reaction (RT-PCR) and RNA fluorescence *in situ* hybridization (RNA-FISH) showed that mDUX transcripts (sense and anti-sense) are made in several tissues at both early and late developmental times, with the highest expression in the central nervous system (CNS). As this has been the only study of expression patterns and mDUX lacks introns, it seems prudent to interpret the RT-PCR data conservatively; however, the RNA-FISH data certainly bolsters the support for mDUX expression in the CNS. An interesting side note is that this group could not amplify full-length transcripts with primers at the 5'- and 3'-termini, which suggests cryptic splicing may create a truncated version of mDUX in some tissues as has been shown for hDUX4 (i.e. hDUX-fullLength and hDUX4-short).

A second paper focused on disease-relevant similarities between mDUX and hDUX4(21). They found that forced overexpression of mDUX lead to cytotoxicity in cultured mouse skeletal muscle cells, mouse fibroblasts and mouse embryonic stem cells. At lower levels of forced mDUX expression in cultured mouse skeletal muscle cells, they observed downregulation of several myogenic regulators and inhibition of differentiation into fused myotubes.

As both mDUX and hDUX4 are transcriptional activators, one key unknown in the DUX/FSHD field is the genome-wide transcriptome of mDUX. These data would facilitate broad comparisons between mDUX and hDUX4 because the hDUX4 transcriptome is well established in cultured human skeletal muscle cells, a common model of FSHD. The genes that hDUX4 activates fall into several interesting categories: germline and stem cell-expressed, RNA processing, ubiquitin pathway, immunity and innate defense, and cancer-testis antigens(22). Another key feature of the hDUX4 transcriptome is that hDUX4 directly binds and activates transcription of many repetitive elements, such as mammalian apparent LTR-retrotransposons (MaLRs), endogenous retrovirus (ERVL and ERVK) elements, and pericentromeric satellite HSATII sequences(23). Furthermore, some of these repetitive elements are used as promoters for various protein-coding genes, non-coding RNAs and antisense transcripts. As many of these repetitive elements are specific to primates, transcriptome studies of mDUX also would enable evaluation of a connection between mDUX and rodent-specific repetitive elements, with implications for both the evolutionary history of DUX genes and modeling of FSHD.

Physiological function of DUX genes

Given the disease relevance of hDUX4, extensive studies have investigated hDUX4 in skeletal muscle – the disease context. However, DUX4 orthologs have been identified across primates, arguing that the DUX4 ORF has been conserved for at least ~74 million years of evolution and thus that it serves some physiological purpose (ignoring for a moment DUXC genes and retrogenes outside of primates as their similarity to hDUX4 is not yet established). Two pieces of evidence provided hints as to the physiological function of hDUX4 prior to my graduate studies. hDUX4 is expressed in the testes, possibly the germline, of individuals not-affected by FSHD and many genes typical of germ and stem cells are upregulated following hDUX4 expression(22). My graduate work on mDUX and a fortuitous collaboration with a group studying early human and mouse development (Dr. Bradley Cairns' Group of the Huntsman Cancer Institute in Salt Lake City, Utah) further strengthened the connection between DUXC genes/retrogenes and a physiological function in early preimplantation embryo development.

Therefore, I will offer a brief summary of what is currently known about early development in mice and humans. Given that samples from early human development are limited, we know much more about early mouse development than in humans, but observational studies of early human development have been performed a handful of times at this point and their quality continues to improve along with improvements in single-cell RNA-sequencing technologies and analysis methods(24-26).

One of the earliest key findings was that of Peaston *et al.* in 2004(27). They found transcripts of transposable elements in both oocytes and early embryos and even observed transposable elements serving as alternative promoters and alternative first exons in these cell types. This was the first demonstration that transposable element regulation could lead to the coordinated expression of a broad transcriptional program. Since then, others have further developed their initial observation using Next-Generation sequencing technologies(28).

In addition to the availability of early mice embryos, mouse embryonic stem cells (mESC) have been studied extensively as a model of early development. mESC are generated through culturing the inner cell mass and thus it was thought that mESC are pluripotent like the inner cell mass. Since there is some confusion around the term "pluripotent", I am referring to cells that are expressing Oct4 and Nanog and retain the potential to differentiate into many cell types, but not trophoblasts (precursors of the placenta). However, it is now well-established that cultured mESC are a heterogeneous population with about 5% of the population at any one time possessing a developmental potential greater than pluripotency (i.e. totipotency)(29-31). Again, there is confusion and controversy around the use of the term "totipotency," but herein I will use the term totipotent to refer to cells that are not expressing Oct4 nor Nanog and can differentiate into all cell types, including trophoblasts. Two markers have been used to identify the totipotent subpopulation in mESC: expression of Zscan4 and expression of mouse endogenous retroviruses with leucine tRNA primer (MERV-L) (28, 31). Studies of mESC that are positive for these markers have a distinct transcriptome from cells that are not expressing these markers, and thus, we have a reasonable idea of the transcriptome of totipotent embryos. Although the field lacks a strong consensus, most agree that two-cell mice embryos are totipotent, but that totipotency is lost shortly thereafter. Two-cell embryos also are the developmental stage at which the embryonic genome is transcribed for the first time (termed embryonic genome activation, EGA). Although chromatin modifiers have been identified and characterized in the early embryo and models thereof (e.g. LSD1/KDM1A, CAF-1)(32-34), a transcriptional activator has thus far been conspicuously absent.

Implications for Development

A key implication of the work described herein is that DUXC genes and retrogenes in each species across placental mammals may be the missing transcriptional activator responsible for activating the embryonic genome for the first time. If subsequent studies corroborate these findings, DUXC genes and retrogenes are far more functionally conserved than their sequences imply. In this vein, targets shared between all DUXC-factors likely represent the core ancestral program necessary during totipotency. Conversely, species-specific targets (such as, but not limited to, retroelements) might contribute species-specific differences that may be layered on top of the core developmental program. Understanding both the conserved and diverged features of early development and totipotency will be essential when applying these findings to assistive reproductive technologies and to the possible creation of induced totipotent stem cells from patient-derived fibroblasts, similar to our current ability to create induced pluripotent stem cells using the Yamanaka Factors(35).

Implications for Disease

Another key implication of the findings described herein is that hDUX4 may be less unique to primates than was previously thought, which opens up new avenues for the development of FSHD disease models. Various models have been proposed and implemented; however, mDUX has not been widely utilized, likely due to its underappreciated similarity to hDUX4. This work establishes direct transcriptional activation of repetitive elements by mDUX in the mouse genome. Interestingly, this relationship is not shared between hDUX4 and the mouse genome – raising the possibility that animal models of FSHD might require forced expression of the DUXC-factor from the particular animal being used in the model. Interesting species to consider first would be mouse/mDUX and possibly, canine/canine DUXC as they are tractable model organisms with many tools available. Another disease-relevant implication of this work is that shared genes regulated by hDUX4 in human cells and mDUX in mouse cells provide a relatively short list of candidate genes responsible for the cytotoxicity of both of these DUX-factors.

Chapter 2. Conservation and Innovation in the DUX4-Family Gene Network

This chapter is in the process of publication as:

Whiddon, J.L., Langford, A.T., Wong, C.J., Zhong, J.W., Tapscott, S.J. (2016) Conservation and innovation in the DUX4-family gene network. Nature Genetics. *In review.*

Introduction

Facioscapulohumeral dystrophy (FSHD) is caused by the expression of the human DUX4 double homeodomain transcription factor in skeletal muscle(36). The DUX4 retrogene is embedded in the D4Z4 macrosatellite arrays in the subtelomeric regions of chromosomes 4 and 10. DUX4 is normally expressed in cells in the lumen of the testis(37), likely the early germline cells, and possibly in the thymus(38), but not skeletal muscle where the D4Z4 array that contains DUX4 is epigenetically repressed in a repeat-dependent manner. The most common form of FSHD (FSHD1; OMIM #158900) is caused by arrays of ten or fewer D4Z4 units(39), whereas FSHD2 is caused by a mutation in SMCHD1, a member of the condensin/cohesion family necessary to maintain D4Z4 epigenetic repression (OMIM #158901)(40). The inefficient epigenetic repression of DUX4 in FSHD results in occasional bursts of DUX4 expression in FSHD muscle cells with the consequent transcriptional activation of several hundred genes regulated by this transcription factor(37). DUX4 induces expression of genes associated with stem cells, alters RNA processing, modulates the innate immune response, and activates the ERVL-MaLR and HERVL families of LTRs (22, 23). Although the continued expression of DUX4 leads to apoptosis, the mechanism of disease pathology *in vivo* remains largely unknown and might be complex based on the large number of genes regulated by DUX4. Expression of human DUX4 in mouse muscle also causes apoptosis (17, 41-43); however, incomplete knowledge of how accurately the expression of human DUX4 in mice recapitulates the FSHD transcriptional program limits the utility of mice as a model for FSHD.

While the human DUX4 (hereinafter hDUX4) transcriptome is known(22, 23), the mouse DUX (hereinafter mDUX) transcriptome remains largely unknown(21) and there is not yet consensus on whether hDUX4 and mDUX are true orthologs(15, 18, 20). mDux and hDUX4 have been called "retro-orthologs" of DUXC (Leidenroth 2012). Orthologs of the hDUX4 retrogene have been identified throughout the primate lineage and primates have lost DUXC, raising the possibility that the retroposition of the DUXC cDNA directly into the DUXC gene converted it to a retrogene. Similarly, mDux is a retrogene in an array of direct repeats, and mice have lost DUXC, although it is unknown whether mDUX was created in a shared retroposition event before the primate and murine lineages diverged or an independent retroposition event after these lineages diverged. Like hDUX4, mDux is expressed in the testes, suggesting a conserved role in some tissues. However, the homeodomain regions of hDUX4 and mDux have only 35% and 58% amino acid identity of the first and second homeodomains, respectively, compared to 100% mouse/human for the majority (63%) of homeodomains(44), suggesting a possible functional divergence between hDUX4 and mDux.

In this study we addressed the extent to which mDUX and hDUX4 are functional homologs by performing expression studies of mDux and hDUX4 in mouse myoblasts and comparing these data to prior studies expressing hDUX4 in human myoblasts. We found that mDUX and hDUX4, when expressed in mouse and human muscle cells, respectively, regulated genes and retroelements characteristic of mouse two-cell embryos, suggesting a conserved role in the establishment of a transcriptional program associated with the totipotent early cleavage embryo. In contrast, their binding motifs diverged resulting in much weaker activation of the two-cell embryo gene signature and almost no activation of LTRs by hDUX4 in mouse muscle cells, which was similar to results with the canine DUXC gene. Together, these data indicate that the ancestral DUX4-family factor regulated a core program related to early embryogenesis, but binding domains and sites have diverged among species, possibly driven by binding and activation of retrotransposons, findings that will inform future non-primate models of FSHD.

Results

mDUX activates genes characteristic of two-cell embryos in mouse myoblasts

The set of genes regulated by human DUX4 has been determined in skeletal muscle cells because its mis-expression in skeletal muscle causes FSHD(22, 23, 45, 46). Therefore, in order to determine a comparable mDUX transcriptome, we transduced mouse C2C12 myoblast cells with a lentivirus containing a codon-altered mDUX transgene regulated by a doxycycline-inducible promoter and used antibiotic selection to create a clonal population (Clone15B). We codon-altered mDUX to decrease the overall CpG content because this was shown to enhance transgene expression of a similar inducible hDUX4 vector(46). We induced mDUX expression in biological triplicates for 36 hours in growth medium and performed RNA-seq. We observed increased expression of 962 genes and decreased expression of 204 genes (absolute log2-foldchange>=2 with adjusted p-value <=0.05), compared to un-induced cells (Fig. 1a, Supplementary Tables 1-2).

To confirm mDUX expression was specific to doxycycline induction, we also performed RNA-seq on a clonal C2C12 cell line where doxycycline treatment induced expression of the firefly luciferase gene and used luciferase-expressing cells as a negative control to determine the mDUX transcriptome. Both methods of determining the mDUX transcriptome were highly concordant (Supplementary Fig. 1a). To check for non-specific gene regulation, we compared luciferase-expressing cells to un-induced cells and observed increased expression of 4 genes and decreased expression of 65 genes using the same cut-off criteria as for the mDUX analysis (Supplementary Fig. 1b). Only one gene affected by mDUX was also affected by firefly luciferase (Serpinb9g) and it was removed from further analyses.

Gene ontology analysis of the mDUX up-regulated genes showed enrichment of 25 gene ontology (GO) categories (P-value<0.05; Supplementary Table 3). Nine of these GO terms were related to development, including the GO term: embryo development. Notably, of the 20 genes that contributed to the enrichment of the embryo development term, there were 12 homeobox genes, including most of the family of oocyte specific homeoboxes (Obox1/2/3/5/6). However, embryo development remains a poorly annotated GO term. For example, some of the genes robustly activated by mDUX, such as Zscan4a-e and Tcstv1/3, are expressed at the 2-cell stage of mouse embryo development but were not present in the embryo development GO category.

Because mDUX activated expression of 2-cell embryo genes Zscan4a-e and Tcstv1/3)(28, 47, 48), we sought to determine whether mDUX regulates a broader set of genes characteristic of the 2-cell embryo (2C). Mouse embryonic stem cells (mESCs) are known to be a heterogeneous population and much work has been done to sort and characterize subpopulations with distinct characteristics. For example, the subpopulation of mESCs expressing Zscan4c are thought to closely resemble the transcriptional landscape of the 2-cell embryo (2C-like) based on RNA-seq profiling (31). Gene Set Enrichment Analysis (GSEA) using this dataset revealed that the mDUX transcriptome is significantly enriched for the Zscan4c+ 2C-like gene signature (NES = 12.56, p-value < 0.001; Fig. 1b). As a negative control, we used GSEA to assess enrichment of the 2C-like state gene signature in a transcriptome where one does not expect to find enrichment (Supplementary Fig. 2a). In addition to Zscan4, many other established 2-cell specific genes contributed to this enrichment, for example: Gm4340 (aka Gm6763), Slc34a2, Tcstv1/3, Tdpoz 3/4, Usp17la-e and Zfp352. Importantly, the previously published 2C-like transcriptome included mDUX itself and mDUX RNA is expressed in mESC

(J. Whiddon, unpublished data). Therefore, expressing mDux in C2C12 mouse myoblasts activated transcription of genes characteristic of the mouse 2-cell embryo.

To determine which genes in this 2C-like gene signature were direct targets of mDUX, we used two complementary strategies to determine mDUX binding locations with chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). First, we used two commercially available mDUX antibodies on the mDUX-inducible C2C12 cell line (Clone 1D). Second, we created a polyclonal population of cells with the doxycycline inducible vector expressing a chimeric protein that fuses the mDUX-CA homeodomains with the hDUX4-CA carboxyterminus (MMH). The MMH-chimera maintains the DNA binding domain of mDUX and the carboxyterminal epitopes of hDUX4, permitting us to use the same DUX4 antisera to IP the chimera and hDUX4 (Supplementary Fig. 3a). We confirmed that the MMH-chimera retained the mDUX DNA-binding specificity by comparing the ChIP-seq peaks of the chimera to those of mDUX. Although the mDUX antibodies had a lower signal-to-noise ratio, and thus identified fewer peaks, the vast majority of the peaks identified by the mDUX-antibody were a subset of the chimera-identified peaks (Supplementary Fig. 3b). ChIP-seq with one mDUX antibody, A-19, found 2,400 peaks, 95% of these peaks overlap a peak in the MMH-chimera dataset. Similarly, ChIP-seq with a second mDUX antibody, S-20, found 628 peaks, 99% of these peaks overlap with a peak in the MMH-chimera dataset. Furthermore, the MEME motif predication algorithm predicted nearly identical motifs for A-19 peaks and MMH peaks (Supplementary Figure 3c). We therefore used the ChIP-seq data set from the MMH-chimera that contained the mDUX binding region with the human DUX4 carboxyterminal epitopes immunoprecipitated with antisera to the human epitopes because of superior signal-to-noise compared to the commercially available antisera to mDUX.

Using the MMH-chimera dataset, we identified 8,187 peaks. Of these peaks, 3% were within 1 kilobase (+/-) of an annotated transcriptional start site (TSS). We defined direct targets as genes that were regulated by mDUX according to RNAseq (as described above) and had a ChIP-seq peak within 1 kilobase of its annotated TSS. By these criteria, we identified 67 genes as direct targets of mDUX. These direct targets comprised 7% of up-regulated genes and zero down-regulated genes. Of the 67 upregulated direct targets of mDUX, 30 of these were in the 2C-like gene signature (Fig. 1c). This is 20-fold more genes than the 1.47 genes expected by chance based on hypergeometric testing (p=7.8E-31). Interestingly, many of the hallmark 2C-like genes, such as Zscan4a-f, Tcstv1/3, Usp17lb/d and Zfp352 were identified as direct targets of mDUX (Supplementary Fig. 4a). We cloned a 450 base-pair region upstream of the Zscan4c TSS that encompassed a mDUX chIP-seq peak and carried two predicted mDUX binding sites into a luciferase reporter and observed strong induction in a mDUX-dependent manner (Supplementary Fig. 4b). Together, these results demonstrate that mDUX expressed in myoblasts directly regulates a large portion of the 2C-like gene signature.

mDUX and hDUX4 activate orthologous genes in myoblasts of their respective species, including genes in the mouse 2C-like gene signature

Despite considerable sequence divergence in their two DNA-binding homeodomain regions (Fig. 1d), we found that mDUX and hDUX4 activated orthologous genes in myoblasts of their respective species, including genes in the mouse 2C-like gene signature. To do this analysis, we compared our mDUX dataset to a previously published RNA-seq dataset from human myoblasts expressing hDUX4(46) that we re-analyzed so that both datasets were processed with the same bioinformatics pipeline. Using the same filtering criteria as for mDUX, we observed increased expression of 1,634 genes and decreased expression of 151 genes with hDUX4 expression. mDUX and hDUX4 strongly induced a similar set of genes in complex, repetitive gene families. For example, the top 30 genes induced by mDUX consisted of seven ZSCAN family members, eleven PRAME family members, three USP17 genes, two THOC4/ALYREF family members, and two EIF1A-like genes. Each of these gene families was also induced by hDUX4 in human myoblasts.

To broaden this analysis, we considered all upregulated genes with 1:1 mouse-to-human orthology according to HomoloGene(49) that were detected in both datasets. There were 885 hDUX4-upregulated and 454 mDUX-upregulated genes that met these criteria. The mDUX-upregulated gene list shares 143 genes with the hDUX4-upregulated gene list, which is significantly more than the 34 expected by chance (p=1.21E-53 by hypergeometric testing). As a second method of comparison, we used GSEA and it revealed that the 500 genes most upregulated by hDUX4 were significantly enriched in the genes most upregulated by mDUX (NES=8.16, p-value<0.001; Fig. 1e) and vice versa (NES=6.01, p-value<0.001; Supplementary Fig. 5a).

To determine whether activation of the early embryo network was conserved between mDux and hDUX4, we used GSEA to assess the extent to which hDUX4 activated the human orthologs of the mouse 2C-like gene signature. Of the 469 genes that comprise the mouse 2C-like gene signature, HomoloGene predicts 297 have simple 1:1 orthologs in humans. GSEA shows that these genes were enriched at the top of the hDUX4 ranked transcriptome of genes with simple 1:1 orthologs in mouse (NES=2.24, p-value = 0.002, Fig. 1f).

It should be noted that these analyses of similarity using the HomoloGene method was very conservative. Complex gene families, such as the ZSCAN4, PRAME, THOC4/ALYREF, and USP17 families were excluded from the HomoloGene dataset because 1:1 orthology cannot be established, but members of each of these complex gene families were upregulated in both species. Together, these data demonstrate a strong functional conservation for mDUX and hDUX4 in regulation of this 2C-like network in their respective species.

mDUX and hDUX4 have partially conserved binding motifs

To better understand the functional conservation between mDUX and hDUX4, we determined the mDUX binding motif and compared it to the hDUX4 binding motif. To do this, we used the de novo motif-finding algorithm, MEME(50), on our mDUX ChIP-seq dataset and re-analyzed our previously published hDUX4 ChIP-seq dataset(22) to avoid analysis differences between the two motifs. Despite their functional conservation, we identified a mDUX binding motif that diverged from the hDUX4 binding motif in the first half of the motif but the second half of the motif was almost completely conserved (Fig. 2a). It is possible that the modularity of the motifs corresponds to separate contributions of each homeodomain, which makes it interesting to note that the four residues predicted to determine DNA-binding-specificity are identical between hDUX4 and mDUX in the second homeodomain but not the first(51) (Fig. 1e).

hDUX4 modestly activates the 2C-like gene signature when expressed in mouse myoblasts

Because of the apparent paradox of the functional conservation of their transcriptomes and the partial divergence of their binding motifs, we next generated RNA-seq and ChIP-seq datasets for hDUX4 in mouse muscle cells to better understand their conservation and divergence (Supplementary Tables 4-5). Using the same methodology as with mDUX, we transduced mouse C2C12 myoblast cells with a lentivirus containing a codon-altered hDUX4 transgene regulated by a doxycycline-inducible promoter and used antibiotic selection to create a clonal population (Clone7). We induced hDUX4 expression in biological triplicates for 36 hours in growth medium and performed RNA-seq. hDUX4 induction increased expression of 582 genes and decreased expression of 428 genes (absolute log2-Foldchange >=2 with adjusted p-value <=0.05), compared to un-induced cells (Supplementary Fig. 4a). Comparing hDUX4-expressing cells to luciferase-expressing cells yielded a similar transcriptome, showing that hDUX4 expression is specific to doxycycline treatment (Supplementary Fig. 4b). One of the hDUX4-upregulated genes (Adgrg1) and 55 of the hDUX4-downregulated genes also changed expression in luciferase-expressing cells and were removed from further analyses.

Overall, hDUX4 regulated many genes that were not orthologous to mDUX-regulated genes and generally showed little similarity to the mDUX transcriptome (Supplementary Fig. 4c). However, GSEA showed significant enrichment of the 2C-like gene signature in the genes upregulated by hDUX4 in mouse cells (NES = 4.25, p-value<0.001; Fig. 2b). The activation of this signature, however, was not as robust by log2 fold-change as the activation when mDUX was expressed in mouse cells. For example, Tcstv3 and Zscan4d had log2 fold-changes of only 0.92 and 0.66, respectively, compared to 10.1 and 12.4 by mDUX, indicating that hDUX4 activates the 2C-like gene signature through moderate induction of many members, perhaps reflecting the partial motif divergence. Importantly, chIP-seq data revealed that hDUX4 bound the same motif in mouse cells as in human cells (Supplementary Fig. 4d, indicating that cofactors do not contribute to the differences in binding specificity between mDUX and hDUX4.

To determine which genes in mouse myoblasts were direct targets of hDUX4, we analyzed the chIP-seq dataset with the same computational pipeline as with mDUX and identified 46,136 hDUX4 peaks, 1.4% of which were within 1 kilobase (+/-) of an annotated TSS. We identified 48 direct targets of hDUX4, which comprised 7.4% of upregulated genes and 1.2% of downregulated genes. Of the 43 upregulated direct targets of hDUX4, Gm13119 and Pdlim3 are in the mouse 2C-like gene signature. Although little is known about Gm13119, it has homology to known PRAME-family genes and several PRAME-family members were upregulated upon hDUX4 expression in human myoblasts, whereas Pdlim3 has no known ortholog in the human genome. As noted above, hDUX4 expressed in mouse cells shows specific but relatively modest transcriptional activation of 2C-like genes, and therefore the identification of only a few direct targets of hDUX4 in the mouse 2C-like gene signature may be the result of weak binding to the mDUX motifs near the 2C-like genes that fails to meet the peak height threshold cutoff.

mDUX, but not hDUX4, binds and activates transcription of repetitive elements in mouse myoblasts

In contrast to the moderate conservation of the 2C-like program in mouse cells, activation of retrotransposons by hDUX4 in mouse cells has completely diverged. Transcription of repetitive elements has been reported in 2C-like mouse ES cells(27, 28). We found that mDUX, but not hDUX4, induced expression of MERV-L elements by 100-fold and pericentromeric satellite DNA by 50-fold (Fig. 3a-c, Supplementary Fig. 7a-c, Supplementary Tables 6-7), both of which are characteristic of 2C embryos and 2C-like ES cells. ChIP-seq data indicated that MERV-L elements were a direct target of mDUX, but not hDUX4 (Supplementary Fig. 8a-b). This is consistent with the finding that mDUX, but not hDUX4, activated a reporter driven by a MERV-L element (Fig. 3d), which carries a good match to the mDUX binding motif (P-value < 0.0001). MERV-L elements have been reported to function as alternative promoters in 2C-embryos(27, 28), which we observed in mDUX-expressing, but not hDUX4-expressing, mouse cells (Fig. 3e, Supplementary Tables 8-9). These results indicate that hDUX4 activated a portion of the 2C-like gene signature in mouse cells, but it did not activate repetitive elements characteristic of the 2C mouse embryo.

Notably, although hDUX4 did not bind MERV-L elements, hDUX4 bound ERVL-MaLR elements in mouse cells (Supplementary Fig. 8b) and in at least 30 cases used them as alternative promoters (Fig. 4a). In some cases, hDUX4 binding to an ERVL-MaLR retroelement caused robust expression of the adjacent gene (Fig. 4b), consistent with our previous finding that hDUX4 binds ERVL-MaLRs when expressed in human cells and uses them as alternative promoters(23). Because activation of retroelements has been tied to creation of the placenta at the base of placental mammals(52), we looked for hDUX4 expression in the placenta. By RT-qPCR and Western blotting, hDUX4 RNA and protein was detected in purchased human placenta samples (Supplementary Fig. 9a-b).

Ancestral function of DUX4-family was to regulate early embryo genes

The above results indicate that mDUX and hDUX4 have maintained the ability to regulate a set of 2C-like genes in mouse cells despite considerable divergence of their homeodomains; however, this functional conservation did not extend to the retrotransposons activated by each. We used chimeric proteins to identify the regions of mDUX and hDUX4 responsible for this partial conservation of function (Fig. 5a). An initial chimera with the mDUX homeodomains and the hDUX4 carboxy-terminus (MMH) matched the transcriptional activity of mDUX (Fig. 5a-c), even on genes where hDUX4 was not active. Thus, the transcriptional divergence between mDUX and hDUX4 mapped to the region containing the two homeodomains.

To determine the relative contribution of each homeodomain, we introduced each human homeodomain individually into mDUX to create the MHM and HMM chimeras (Fig. 5a). Neither MHM nor HMM activated transcription of MERV-L-promoted 2C-like genes (Fig. 5b). However, for 2C-like genes with conventional promoters, the individual hDUX4 homeodomains showed different capacities to substitute for the corresponding mDUX homeodomain. The second hDUX4 homeodomain (MHM) consistently showed stronger activation of the target genes compared to the first hDUX4 homeodomain (HMM; Fig. 5c-d), consistent with the higher similarity of the second homeodomains of mDUX and hDUX4 (Fig. 1d). (We confirmed MHM and HMM expression and stability using a reporter assay (Supplementary Fig. 10a). Reciprocal experiments in human cells also demonstrated that the second homeodomains were more equivalent than the first homeodomains (Figure 5e-f). Therefore, the similarity of the second homeodomain was important to maintain the functional conservation of the 2C-like gene signature at conventional promoters.

To further explore the evolutionary conservation of the DUX4-family to activate an early embryo gene signature, we expanded our analysis to include the canine DUXC gene (hereinafter cDUXC) as a proxy for the ancestral DUX4-family gene. Both mDUX and hDUX4 are retroposed copies of ancestral DUXC mRNA and neither mice nor humans have retained DUXC(15, 18, 20). cDUXC has two canonical homeodomains like mDUX and hDUX4 (percent amino acid identify HD1/HD2 = 42/55 compared to mouse and 60/73 compared to human; Fig. 1d). When expressed in mouse muscle cells, cDUXC did not activate MERV-L-promoted genes (Fig. 5b). However, cDUXC did activate transcription of 2C-like genes with conventional promoters (Fig. 5c-d). Although based on a small number of promoters, this result is consistent with our genome-wide data comparing mDUX and hDUX4, and suggests that the ancestral DUX4-like gene activated an early embryonic developmental program that was independent of retrotransposon-promoted genes.

Discussion

Together our data indicate that an ancestral DUXC gene evolved to regulate a core transcriptional program characteristic of early embryos, and that the retroposed mDux and hDUX4 have maintained this function in rodents and primates, respectively. This functional conservation was maintained despite significant divergence of their homeodomains and their corresponding DNA binding motifs, particularly the first homeodomain and the first half of the motif. As a consequence, although hDUX4 is able to activate many genes within the 2C-like gene signature when expressed in mouse cells, the transcriptional activation was much less robust and ChIP-seq identified fewer binding events at these genes, both indicative of weaker binding of hDUX4 to the mDux binding sites in these genes. The second homeodomain is more highly conserved between mDux, hDUX4 and cDUXC, and one half of the mDUX and hDUX4 binding motifs are highly conserved. These data also support the suggestion that a single homeodomain factor might have driven this pluripotency program prior to the generation of the DUX family in placental mammals, and that activation of endogenous retroelements necessary for placental development was the original selective advantage conferred by homeodomain duplication. This hypothesis is consistent with our detection of hDUX4 in human placenta (Geng, L. N., unpublished data) and reports of hDUX4 target gene expression in trophectoderm(53).

In contrast to the conservation of the second homeodomain and the activation of the 2C transcriptional program, retrotransposons and retrotransposon-promoted genes likely reflect species-specific additions to the early embryo transcriptome. One hypothesis is that mDUX and hDUX4 have diverged because of a dynamic evolutionary relationship between these factors and distinct families of retrotransposons. MERV-L elements were robustly activated by mDUX, but not hDUX4. However, hDUX4 robustly activated ERVL-MaLR elements in human cells. This near dichotomy of activity suggests that MERV-L activation might have driven the divergence of mDUX from hDUX4. This is particularly interesting because hDUX4 did activate transcription from ERVL-MaLRs in mouse cells that were not activated by mDux, suggesting that mDUX might have diverged away from binding these sequences. Several of the hDUX4-activated ERVL-MaLRs in mouse cells acted as alternative promoters for nearby genes, indicating how quickly and profoundly the transcriptome can be rewired based on the divergence of the DUX homeodomain sequences.

Such comparisons are particularly relevant to FSHD where it remains unclear how to model this disease in non-primate animals. The fact that both hDUX4 and mDUX expression leads to apoptosis in mouse muscle cells supported the use of hDUX4 in mice as a model of FSHD(21, 54). However, although hDUX4 modestly activated the early embryonic 2C-like gene signature when expressed in mouse muscle cells, the activation was both weak and incomplete when compared to its activation of the orthologous genes in human cells, and the LTR-driven program was completely different. This suggests that expression of hDUX4 in mice will, at best, recapitulate only a small portion of the program it activates in human cells and might not be an effective model of the human disease. Or, at a minimum, experiments using such a model need to be carefully focused on the limited cross-species activities. Our study further suggests that using mDUX-expression in mice might be an alternative approach to model FSHD and should be explored in parallel with other models.

A very recent study reported comparisons between hDUX4 and mDUX (55) and although our studies were similarly motivated, our approaches differ considerably such that our study addresses key questions raised in their study. First, we created clonal cell lines with stably integrated doxycycline-inducible transgenes in mouse myoblasts, while they performed transient transfections of human myoblasts. They asserted that the partial functional homology they observed in human myoblasts had positive implications for modeling FSHD with hDUX4 expression in mice, which is the strategy of several current disease models; our data in mouse myoblasts addressed their assertion directly and we found hDUX4 targets differed distinctly between human and mouse myoblasts. Second, our analyses included investigation of transcriptional activation of retroelements, which is completely lacking in their analyses, but is likely critical to understanding both FSHD and the physiological functions of these factors. Third, because we expressed mDUX in its relevant mouse genomic context, we observed a robust gene signature and identified this gene signature as characteristic of the cleavage-stage embryo. This is the strongest indication yet of a physiological role of the DUX4-family genes, which we extended not only to hDUX4 in its relevant human genomic context, but also to canine DUXC. Therefore, our study provides a clear extension to their study by suggesting an alternative method for modeling FSHD and by revealing an evolutionarily conserved physiological function of the DUX4-family transcription factors that is, however, largely restricted to their relevant genomic contexts, likely due to dynamic relationships between these factors and retroelements.

In conclusion, we found that mDUX and hDUX4 are likely functional homologs with a conserved role in establishing the transcriptome in the totipotent cleavage embryo when expressed in cells from their respective species, but largely activate transcription of disparate genes and few retroelements when expressed in a cross-species context (namely, hDUX4 in mouse cells in this study). This is consistent with the divergence of their binding motifs in the first half of the motif, while conservation in the second half of the motifs may explain why hDUX4 retained a modest ability to activate the two-cell embryo gene signature. Neither hDUX4 nor canine DUXC activated mDUX-activated retroelements, which raised the interesting possibility that interactions with retroelements drove the sequence and binding site divergence between hDUX4 and mDUX. These findings inform future non-primate models of FSHD and provide a model for studying genome evolution especially in regards to the critical balance between conservation of a key developmental program with the innovation driven by binding to mobile retrotransposon promoters.

Materials and Methods

General Statistical Methods

Standard statistical tests were used and described for each individual application. Biological triplicates were used for RNA-seq and RT-PCR as indicated. The ChIP-seq studies were multiple singleton experiments with several antibodies that would IP the same binding domain, as described. No statistical methods were used to predetermine sample size.

Whole genome RNA-sequencing (RNA-seq)

C2C12, mouse myoblasts (ATCC® CRL-1772™), were grown in DMEM (Gibco/Life Technologies) supplemented with 10% fetal bovine serum (Thermo Scientific) and 1% penicillin/streptomycin (Life Technologies). These cells were obtained from ATCC and passaged without losing the ability to differentiate into myotubes but have not routinely been checked for mycoplasma. We cloned mDUX transgene into the pCW57.1 lentiviral vector, a gift from David Root (Addgene plasmid #41393), which has a doxycycline-inducible promoter. mDUX and hDUX4 transgenes were codon-altered to decrease overall CpG content because this was shown to enhance transgene expression of the inducible hDUX4 vector(46). To create monoclonal cell lines, we first transduced pCW57.1-mDUX into 293T cells (ATCC® CRL-3216™), along with the packaging and envelope plasmids pMD2.G and psPAX2 using lipofectamine 2000 reagent (ThermoFisher). Viral-like-particles containing pCW57.1-hDUX4 was a gift from Sean Shadle and was prepared in a similar manner. C2C12 were plated at low density and transduced with lentivirus at a low multiplicity of infection (MOI < 1) in the presence of polybrene. Cells were selected and maintained in 2.6ug/ml puromycin. Individual clones were isolated using cloning cylinders about 7 days after transfection and chosen for analysis based on robust transgene expression following 2ug/ml doxycycline treatment for 36 hours.

Biological triplicates were prepared and total RNA was extracted from whole cells using NucleoSpin RNA kit (Macherey-Nagel) following the manufacturer's instructions. Total RNA integrity was checked using an Agilent 2200 TapeStation (Agilent Technologies, Inc., Santa Clara, CA) and quantified using a Trinean DropSense96 spectrophotometer (Caliper Life Sciences, Hopkinton, MA). RNA-seq libraries were prepared from total RNA using the TruSeq

RNA Sample Prep v2 Kit (Illumina, Inc., San Diego, CA, USA) and a Sciclone NGSx Workstation (PerkinElmer, Waltham, MA, USA). Library size distributions were validated using an Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). Additional library QC, blending of pooled indexed libraries, and cluster optimization were performed using Life Technologies' Invitrogen Qubit® 2.0 Fluorometer (Life Technologies-Invitrogen, Carlsbad, CA, USA). RNA-seq libraries were pooled (14-plex) and clustered onto two flow cell lanes. Sequencing was performed using an Illumina HiSeq 2500 in "rapid run" mode employing a single-read, 100 base read length (SR100) sequencing strategy. Image analysis and base calling was performed using Illumina's Real Time Analysis v1.18 software, followed by 'demultiplexing' of indexed reads and generation of FASTQ files, using Illumina's bcl2fastq Conversion Software v1.8.4 (http://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html).

RNA-seq Data Analysis

Reads of low quality were filtered prior to alignment to the reference genome (mm10 assembly) using R (development version 3.4.0) and Bioconductor (3.3.0) to call TopHat v2.1.0(56), Bowtie and GenomicAlignments. Reads were allowed to map up to 20 locations. Reads overlapping UCSC known genes were counted using summerizeOverlaps and differential gene expression was determined using DESeq2, which calculated P-values using the Wald test and adjusted P-values for multiple testing using the procedure of Benjamini and Hochberg. DESeq2 estimates variance for each gene using the average expression level across all samples(57). Gene Set Enrichment Analysis (GSEA) was performed using the GSEApreranked module of the Broad Institute's GenePattern(58) algorithm. Specifically, we used 1,000 gene list permutations to determine P-value and the classic scoring scheme(59). As we only compared to one gene set (from Akiyama et al.(31)), we did not need to correct for multiple tests. For GSEA plot interpretation, see Figure 1b legend. For negative control, see Supplementary Fig. 1. Gene Ontology analysis (GO) analysis was done using Gene List Analysis tool of the PANTHER Classification System(60) (version: 10.0), which calculated P-values using the binomial statistic as described in the PANTHER User Manual (http://pantherdb.org/help/PANTHER_user_manual.pdf). Repeat element analysis was accomplished using repStats (version: 0.99.0). Briefly, repStats uses summarizeOverlaps to count reads that overlap RepeatMasker-annotated repeat elements. Note, reads counts based on reads that mapped to multiple locations were divided by the number of mapped locations. Reads that support repeats used as alternative promoters or alternative first exons were identified and activation scores were calculated as described previously(23), with the one exception that we retained reads that linked chIPseq peaks to annotated exons regardless of whether they spliced across an intron or not.

DNA sequencing after chromatin immunoprecipitation (ChIP-seq)

hDUX4 ChIP-seq datasets were based on monoclonal cell lines described above and were straight-forward given the availability of polyclonal antibodies to hDUX4; MO488 and MO489 were used in this study (previously described in Geng et al.(22)). We performed ChIP-seq for mDUX using two complementary approaches. First, we used two commercially available mDUX antibodies on a mDUX-inducible C2C12 clonal cell line prepared as described for RNA-seq (A-19, catalogue number: sc-385089 and S-20, catalogue number: sc-385090, Santa Cruz Biotechnology). Second, we created a polyclonal population of cells with the doxycycline inducible vector expressing a chimeric protein that fuses the codon-altered mDUX homeodomains with the codon-altered hDUX4 carboxyterminus (MMH). The MMH-chimera maintains the DNA binding domain of mDUX and the carboxy-terminal epitopes of hDUX4, permitting us to use the same hDUX4 antisera to IP the MMH-chimera and hDUX4 (Supplementary Fig. 8a). We confirmed that the MMH-chimera retained the mDUX DNA-binding specificity by comparing the ChIP-seq peaks of the chimera to those of mDUX. Although the mDUX antibodies had a lower signal-to-noise ratio, and thus identified fewer peaks, the vast majority of the peaks identified by the mDUX-antibody were a subset of the chimera-identified peaks (Supplementary Fig. 8b). ChIP-seq with one mDUX antibody, A-19, found 2,400 peaks, 90% of these peaks overlap a peak in the MMH-chimera dataset (8,187 peaks). Similarly, ChIP-seq with a second mDUX antibody, S-20, found 628 peaks, 97% of these peaks overlap with a peak in the MMH-chimera dataset. Furthermore, the MEME motif predication algorithm predicted nearly identical motifs for A-19 peaks and MMH peaks (Supplementary Fig. 8c). We therefore used the ChIP-seq data set from the MMH-chimera for all the analyses described in the main text because of the superior signal-to-noise compared to the commercially available antisera to mDUX.

Cross-linked ChIP was performed similar to previous reports for other transcription factors(61, 62). Briefly, ~6x10^7 cells were fixed in 1% formaldehyde for 11 minutes, quenched with glycine, lysed, and then sonicated to generate final DNA fragments of 150–600 bp. The soluble chromatin was diluted 1:10 and pre-cleared with protein A:G beads for 2 hours. Remaining chromatin was incubated with primary antibody overnight, then protein A:G beads were added for an additional 2 hours. Beads were washed and then de-crosslinked overnight. ChIP samples were validated by RT-qPCR and then prepared for sequencing per the Nugen Ovation Ultralow library system protocol with direct read barcodes. ChIP-seq libraries were prepared from IP samples using an Ovation Ultralow Library System kit (NuGEN Technologies., San Carlos, CA, USA). Library size distributions were validated using an Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). Additional library QC, blending of pooled indexed libraries, and cluster optimization were performed using Life Technologies' Invitrogen Qubit® 2.0 Fluorometer (Life Technologies-Invitrogen, Carlsbad, CA, USA). ChIP-seq libraries were pooled (12-plex) and clustered onto two flow cell lanes. Sequencing was performed using an Illumina HiSeq 2500 in Rapid Mode employing a single-read, 100-base read length (SR100) sequencing strategy. hDUX4 ChIP-seq was performed separately from mDUX and MMH.

ChIP-seq Data Analysis

Image analysis and base calling were performed using Illumina's Real Time Analysis v1.18 software, followed by 'demultiplexing' of indexed reads and generation of FASTQ files, using Illumina's bcl2fastq Conversion Software v1.8.4 (http://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html). Reads of low quality were filtered out prior to alignment to mm10, using BWA 0.7.10(63). Further ChIPseq computational analyses were performed using R (development version 3.4.0) and Bioconductor (3.3.0). Raw reads were aligned to mm10 using Rsamtools, ShortRead, and Rsubread. Peak calling was done with MACS2 (macs2 2.1.0.20151222), only peaks with q-value < 0.01 were considered. MACS2 calculated q-values p-values using the Benjamini-Hochberg procedure. Motif prediction was done with MEME-ChIP 4.11.2(50), which includes FIMO analysis.

Transient transfection and RT-qPCR

Transient DNA transfections of C2C12 cells were performed using SuperFect (QIAGEN) according to manufacturer specifications. Briefly, 80,000 cells were seeded per well of a 6-well plate the day prior to transfection, 2ug DNA/well and 10ul SuperFect/well. 24hrs post-transfection, total RNA was extracted from whole cells using NucleoSpin RNA kit (Macherey-Nagel) following the manufacturer's instructions. One microgram of total RNA was digested with DNAseI (Invitrogen) and then reverse transcribed into first strand cDNA in a 20 uL reaction using SuperScript III (Invitrogen) and oligo(dT) (Invitrogen). cDNA was diluted and used for RT-qPCR with iTaq Universal SYBR Green Supermix (Bio-Rad). Primer efficiency was determined by standard curve and all primer sets used were >90% efficient. Relative expression levels were normalized to the endogenous control locus Timm17b and empty vector by DeltaDeltaCT. The primers used in this study are listed in Supplementary Table 10.

Transient transfection and dual luciferase assay

Transient DNA transfections of C2C12 cells were performed using SuperFect (QIAGEN) according to manufacturer specifications. Briefly, 16,000 cells were seeded per well of a 24-well plate the day prior to transfection, 1μg total DNA/well and 5μl SuperFect/well. Cells to be analyzed via RT-qPCR were transfected with the expression plasmid indicated and RNA was harvested 24 hours post-transfection, then RT-qPCR proceeded as described above. Cells to be analyzed via dual luciferase assay were co-transfected with a pCS2 expression vector carrying the affector construct indicated (500ng/well), a pCS2 expression vector carrying renilla luciferase (20ng/well) and a pGL3-basic reporter vector (500ng/well) carrying test promoter fragment upstream of the firefly luciferase gene. Cells were lysed 24 hours post-transfection in Passive Lysis Buffer (Promega). Luciferase activities were quantified using reagents from the Dual-Luciferase Reporter Assay System (Promega) following manufacturer’s instructions. Light emission was measured using BioTek Synergy2 luminometer. Luciferase data are given as the averages ± SEM of at least triplicates.

RT-qPCR and Western Blotting for hDUX4 in human tissues

RT-qPCR for full length hDUX4 and western blotting for hDUX4 in human tissues was performed as described in Snider et al.(37). RNA and protein lysates from human tissues were purchased from BioChain (Hayward, CA) and Origene (Rockville, MD).

Code Availability

The code that supports the findings of this study are available from GitHub at the following link: https://github.com/TapscottLab.

Data Availability

The data generated in this publication have been deposited in NCBI's Gene Expression Omnibus(64) and are accessible through GEO Series accession number GSE87282 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87282).The RNA-seq data of hDUX4-expressing human myoblasts from Jagannathan et al. (46) has GEO Series accession number GSE85461.

Acknowledgements

Thank you to Sean Shadle for making and sharing the pCW57.1-hDUX4 virus-like particles used to generate the hDUX4-inducible C2C12 clonal cell lines. Thank you to Peter Hendrickson for 2C::TdTomato Reporter (Addgene Plasmid #40281), from which we cloned the MERV-L used to create the MERV-L reporter in this study. Thank you to the Fred Hutchinson Cancer Research Center Sequencing Core Facility. The authors would also like to thank members of the research groups of Stephen Tapscott and Harmit Malik for helpful discussions and critical reading of the manuscript. This work was supported by: NIH NIAMS R01AR045203, National Science Foundation Graduate Research Fellowship Program: DGE-1256082, Interdisciplinary Training in Genome Sciences: T32 HG00035 from the National Human Genome Research Institute.

Conflict of Interest Statement

The authors declare no conflict of interest.

Legends to Figures

Figure 1. mDUX and hDUX4 activated an early embryo gene signature in muscle cells of their respective species

(a) mDUX transcriptome in C2C12 mouse muscle cells: red dots are genes affected more than absolute(log2FoldChange)>=2 and adjusted p-value<=0.05.

(b) GSEA: gene set is 2C-like gene signature; x-axis is log2FoldChange-ranked mDUX transcriptome. Green line is running enrichment score(ES); ES increases when a gene in the mDUX transcriptome is also in 2C-like gene set; ES decreases when a gene isn't in 2C-like gene set. Increases are also indicated by vertical black bars. Enrichment score at the peak normalized by gene set size is NES. P-value was empirically determined based on permutations of ranked gene lists(59).

(c) Direct targets are defined by RNA-seq (absolute(log2FoldChange)>=2 and adjusted p-value<=0.05) and ChIP-seq (peak within one kilobase +/- of transcriptional start site, TSS). Shown are the 30 genes in the 2C-like state gene signature out of 67 total mDUX direct targets.

(d) Homeodomain alignments (%=amino acid identity, *=four predicted DNA-contacting residues, cDUXC=canine DUXC).

(e) GSEA: gene set is the top 500 most upregulated genes in hDUX4-expressing human cells, x-axis is log2FoldChange-ranked mDUX transcriptome in mouse cells. This cross-species comparison required limiting both gene set and transcriptome to 1:1 mouse-to-human orthologs. The opposite comparison is in Supplementary Fig. 2b.

(f) GSEA: gene set is the human orthologs of the mouse 2C-like gene signature, x-axis is log2FoldChange-ranked hDUX4 transcriptome in human muscle cells. Both gene set and transcriptome are limited to 1:1 mouse-to-human orthologs. Note: mouse 2C-like gene signature has 469 genes total, 297 gene have simple 1:1 mouse-to-human orthology.

Figure 2. Despite binding motif divergence and general transcriptome divergence, hDUX4 transcriptome in mouse muscle cells is enriched for the 2C-like gene signature

(a) Comparison of mDUX and hDUX4 binding motifs as determined by MEME. Note the divergence in the first half of the motif and the conservation of the second half of the motif.

(b) GSEA: gene set is the mouse 2C-like gene signature, x-axis is the log2FoldChange-ranked hDUX4 transcriptome in mouse cells. Since the mouse 2C-like gene signature and this hDUX4 transcriptome were both identified in mouse cells, neither gene set nor transcriptome was limited to genes with 1:1 mouse-to-human orthology.

Figure 3. mDUX, but not hDUX4, activates transcription of repetitive elements characteristic of the early embryo in mouse muscle cells

(a) Expression levels of repeats during mDUX expression in mouse cells. Each dot is a repeatName as defined by RepeatMasker. Red color indicates differential expression at absolute log2-Foldchange>=1 and adjusted p-value<=0.05. Number in parentheses is log2-FoldChange.

(b) Same as (a) for hDUX4-expressing mouse muscle cells.

(c) Same as (a) for hDUX4-expressing human muscle cells, data previously published(46).

(d) Luciferase assay showing mDUX induction of luciferase driven by a 2C-active MERV-L element, which contains a match to the mDUX motif. Data shown are mean of 3 biological replicates with s.e.m. error bars. This experiment, with biological replicates, was also repeated on three separate occasions with consistent results.

(e) Black bars are counts of genes in the 2C-like gene signature that are MERV-L promoted and activated by the indicated factor. White bars are genes detected by RNAseq, but are not upregulated. Gray bars are genes with no reads by RNAseq.

Figure 4. hDUX4 bound repetitive elements that also have RNAseq reads that connect the ChIP-seq peak to an annotated exon in mouse muscle cells

(a) LTR-family distribution of bound elements with RNAseq reads that connect the element to an annotated exon.

(b) Two examples of hDUX4 binding an LTR to induce novel transcription. Repeat = black box.

Figure 5. Transcriptional divergence between hDUX4 and mDUX maps to the two DNA-binding homeodomains

(a) Cartoons of chimeric proteins; MMH is the two mDUX homeodomains and the hDUX4 C-terminus; MHM is mDUX with HD2 from hDUX4; HMM is mDUX with HD1 from hDUX4.

(b-d) RT-qPCR data for 2C-like genes in mouse muscle cells of various classes. Data shown are mean of 3 biological replicates with s.e.m. error bars. The experiments in (b) and (d) were also repeated on three separate days with biological triplicates and showed consistent results. The experiments in (c) have not yet been repeated on another day.

(b) 2C-like genes with MERV-L promoters

(c) 2C-like genes with conventional promoters that are induced by hDUX4 and mDUX

(d) 2C-like genes with conventional promoters that are induced only by mDUX

(e) Cartoons of reciprocal set of chimeric proteins; HHM is the two hDUX4 homeodomains and the mDUX C-terminus; HMH is hDUX4 with HD2 from mDUX; MHH is hDUX4 with HD1 from mDUX.

(f) RT-qPCR data for hDUX4-target genes in human rhabdomyosarcoma cells. Data shown are mean of 3 biological replicates with s.e.m. error bars. These experiments have not yet been repeated on another day.

Legends to Supplementary Figures

Supplementary Figure 1. Using an alternative negative control to determine the mDUX and hDUX4 transcriptomes revealed consistent transcriptomes.

(a) Comparing the log2FoldChange of genes using two methods of determining mDUX transcriptome: x-axis log2FoldChange is calculated between mDUX +doxy and -doxy; y-axis log2FoldChange is calculated between mDUX+doxy and Luciferase+doxy.

(b) RNA-seq of biological triplicates of a clonal mouse muscle cell line induced to express Luciferase by treatment with doxycycline (y-axis) compared to un-induced cells (x-axis).

Supplementary Figure 2. Negative control for Gene Set Enrichment Analyses (GSEA)

(a) As a negative control, we used GSEA to assess enrichment of the 2C-like gene signature in a transcriptome where one does not expect to find enrichment. The transcriptome we used was a published dataset representing the MyoD transcriptome when expressed lentivirally in mouse embryonic fibroblasts(61). MyoD has no known role in the 2C mouse embryo, rather it is the master regulator of muscle lineage specification(65-67). That this graph peaks near the center of the x-axis indicates that the majority of the 2C-like state genes are unaffected by MyoD (vertical hash mark). This contrasts distinctly with the taller, left-shifted peak seen in Fig. 1b, for example.

GSEA determined p-values by permuting the transcriptome 1,000 times, hence our report of "p-value<0.001". It seems likely that with more permutations there would be more distinction between the p-value reported for this transcriptome and the p-values reported elsewhere in this study.

Supplementary Figure 3. mDUX binding sites were identified using two complementary ChIP-seq approaches

(a) Cartoons of antibodies and chimera combinations used in ChIP-seq.

(b) Amount of overlapping peaks by genomic coordinates.

(c) De novo motif prediction for peaks called from mDUX_A-19 and MMH_MO488/489.

Supplementary Figure 4. Zscan4c is a direct target of mDUX

(a) ChIP-seq and RNA-seq coverage near the Zscan4c locus. Black rectangle shows location of 450bp sequence (chr7:11,005,309-11,005,758) that was synthesized and cloned upstream of luciferase to create the Zscan4c reporter. Find Individual Motif Occurrences (FIMO) identified two mDUX binding motifs that overlap the Zscan4c reporter region. Figure prepared with Integrative Genomics Viewer(68, 69).

(b) Luciferase assay data using reporter that includes 450bp DNA under the mDUX ChIP-seq peak near the TSS of Zscan4c and either mDUX or an empty vector. Data shown are mean of 3 biological replicates with s.e.m. error bars. This experiment, with biological replicates, was also repeated on three separate occasions with consistent results.

Supplementary Figure 5. Reciprocal GSEA showing mDUX and hDUX4 activate orthologous genes in their respective species

(a) Making the opposite comparison as the graph in main text Fig. 1e, this GSEA shows that the 500 genes most upregulated by mDUX were significantly enriched in the genes most upregulated by hDUX4. The x-axis is the log2FoldChange-ranked hDUX4 transcriptome. This analysis compared mDUX-expressing mouse cells to hDUX4-expressing human cells. Since this comparison is between species, we limited both gene set and transcriptome to genes with simple 1:1 mouse-to-human orthologs.

Supplementary Figure 6. RNA-seq and ChIP-seq data for hDUX4 expressed in mouse muscle cells

(a) hDUX4 transcriptome in mouse muscle cells. Red dots are genes affected more than absolute(log2FoldChange)>=2 and adjusted p-value<=0.05 are shown in red.

(b) Comparing the log2FoldChange of genes using two methods of determining hDUX4 transcriptome: x-axis log2FoldChange is calculated between hDUX4 +doxy and -doxy; y-axis log2FoldChange is calculated between hDUX4+doxy and Luciferase+doxy.

(c) Comparison of transcriptome induced by hDUX4 and mDUX in mouse muscle cells. Only genes for which we had reads in both data sets are included: 13,515 genes total. Spearman's rank correlation coefficient is 0.1812.

(d) Comparison of hDUX4 binding motifs in mouse and human muscle cells as determined by MEME.

Supplementary Figure 7. Distribution of transcribed repeats broken down by repFamily

(a) mDUX-expressing mouse muscle cells

(b) hDUX4-expressing mouse muscle cells

(c) re-analyzed data from hDUX4-expressing human muscle cells

Supplementary Figure 8. ChIP-seq supports mDUX, but not hDUX4, binding to MERV-L in mouse muscle cells

(a) mDUX and hDUX4 ChIP-seq coverage in mouse muscle cells at a MERV-L LTR.

(b) 26% of the 8187 total mDUX binding sites we identified fall within LTR elements, which is 2-fold more than expected if these binding sites were evenly distributed across the genome. Both ERVK and ERVL elements contributed to the enrichment. Although hDUX4 binding sites are not overrepresented in LTR elements in mouse cells (compare third bar to second bar), hDUX4 has 1.7-fold more binding sites in ERVL-MaLRs than expected by genomic distribution. Previously published hDUX4 binding site distribution in human muscle cells shown for comparison(22, 23).

(c) The MERV-L LTR consensus sequence carries a match to the mDUX binding motif (q-value = 0.0132).

Supplementary Figure 10. Luciferase assay with (HUMAN)ZSCAN4 promoter

(a) To confirm that the chimeric proteins were expressed and stable, we tested the chimeras by luciferase assay on a reporter that responds to both hDUX4 and mDUX (J. Whiddon, unpublished data). Such a reporter is the published (HUMAN)ZSCAN4 promoter driving luciferase(22), which has four good matches to the hDUX4 binding motif and two good matches to the mDUX binding motif. Data shown are mean of 3 biological replicates with s.e.m. error bars. This experiment has not yet been repeated on a separate occasion.

NOTE: Supplementary Tables can be found in the publication:

Whiddon, J.L., Langford, A.T., Wong, C.J., Zhong, J.W., Tapscott, S.J. (2016) Conservation and innovation in the DUX4-family gene network. Nature Genetics. *In review.*

Supplementary Table Titles

Supplementary_Table1_mDUX_RNAseq_genes.xlsx Genes up- or down-regulated in mDUX-expressing mouse muscle cells

Supplementary_Table2_mDUX_ChIPseq.xlsx ChIP-seq peaks for mDUX- and MMH-expressing mouse muscle cells

Supplementary_Table3_mDUX_GO.xlsx GO analysis of mDUX transcriptome

Supplementary_Table4_hDUX4_RNAseq_genes.xlsx Genes up- or down-regulated in hDUX4-expressing mouse muscle cells

Supplementary_Table5_ hDUX4_ChIPseq.xlsx ChIP-seq peaks for hDUX4-expressing mouse muscle cells

Supplementary_Table6_mDUX_RNAseq_repeats.xlsx Repetitive element differential expression analysis in mDUX-expressing mouse muscle cells

Supplementary_Table7_hDUX4_RNAseq_repeats.xlsx Repetitive element differential expression analysis in hDUX4-expressing mouse muscle cells

Supplementary_Table8_mDUX_peakAssociatedGenes.xlsx Peak associated genes in mDUX-expressing mouse muscle cells

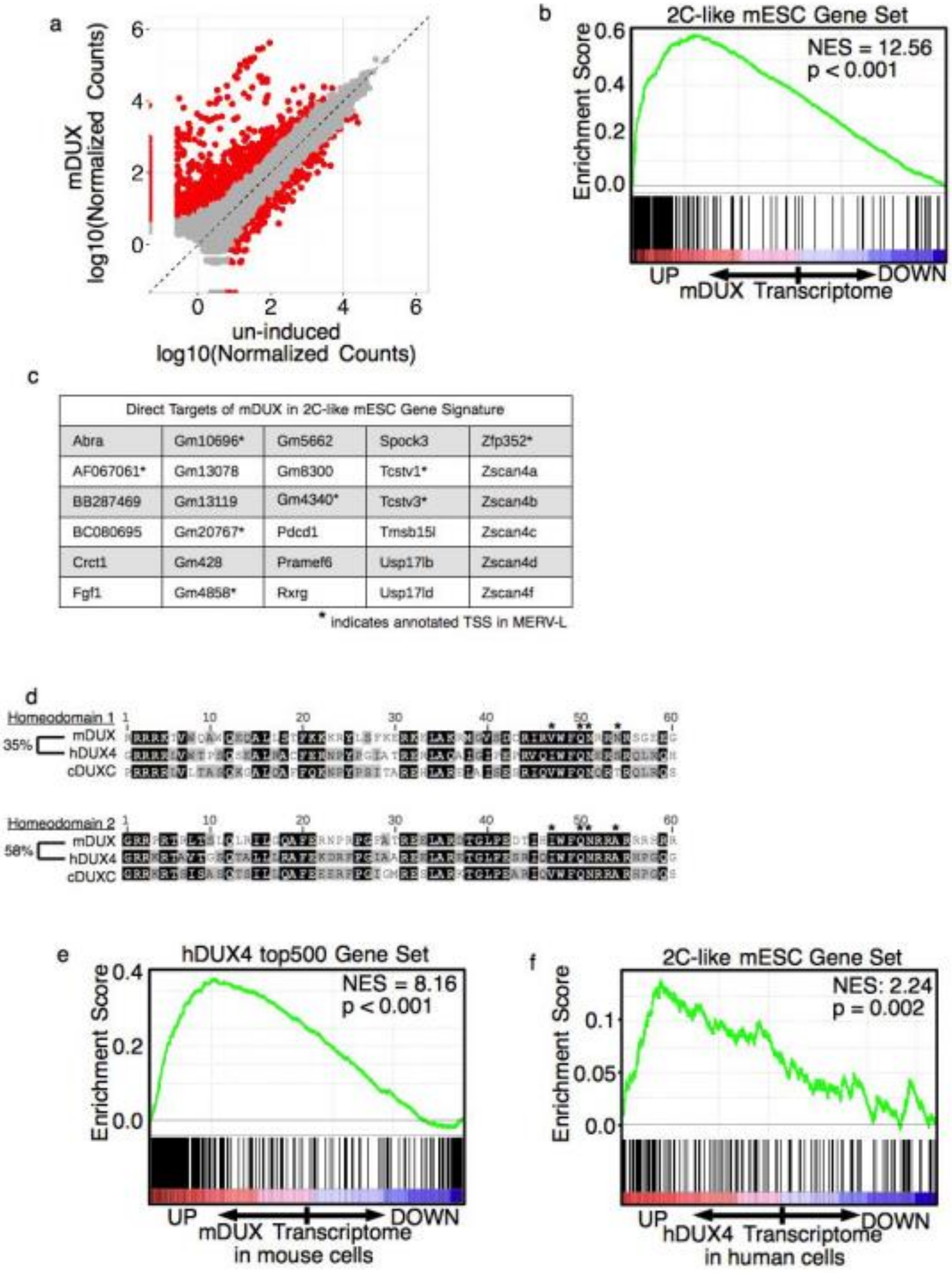
Supplementary_Table9_hDUX4_peakAssociatedGenes.xlsx Peak associated genes in hDUX4-expressing mouse muscle cells

Supplementary_Table10_RT-qPCR_primers.xlsx RT-qPCR primers

Abbreviations

FSHD	Facioscapulohumeral muscular dystrophy
mDUX	mouse DUX
hDUX4	human DUX4
cDUXC	canine DUXC

Figure 1



Direct Targets of mDUX in 2C-like mESC Gene Signature				
Abra	Gm10696*	Gm5662	Spock3	Zfp352*
AF067061*	Gm13078	Gm8300	Tcstv1*	Zscan4a
BB287469	Gm13119	Gm4340*	Tcstv3*	Zscan4b
BC080695	Gm20767*	Pdcd1	Tmsb15l	Zscan4c
Crct1	Gm428	Pramef6	Usp17lb	Zscan4d
Fgf1	Gm4858*	Rxrg	Usp17ld	Zscan4f

* indicates annotated TSS in MERV-L

Figure 2

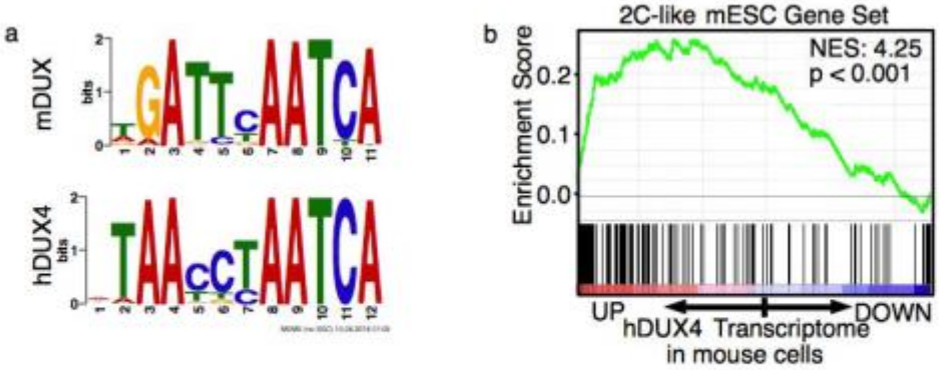

a
mDUX
bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11
hDUX4
bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12
b
2C-like mESC Gene Set
Enrichment Score
0.2
0.1
0.0
NES: 4.25
p < 0.001
UP
DOWN
hDUX4 Transcriptome
in mouse cells

Figure 3

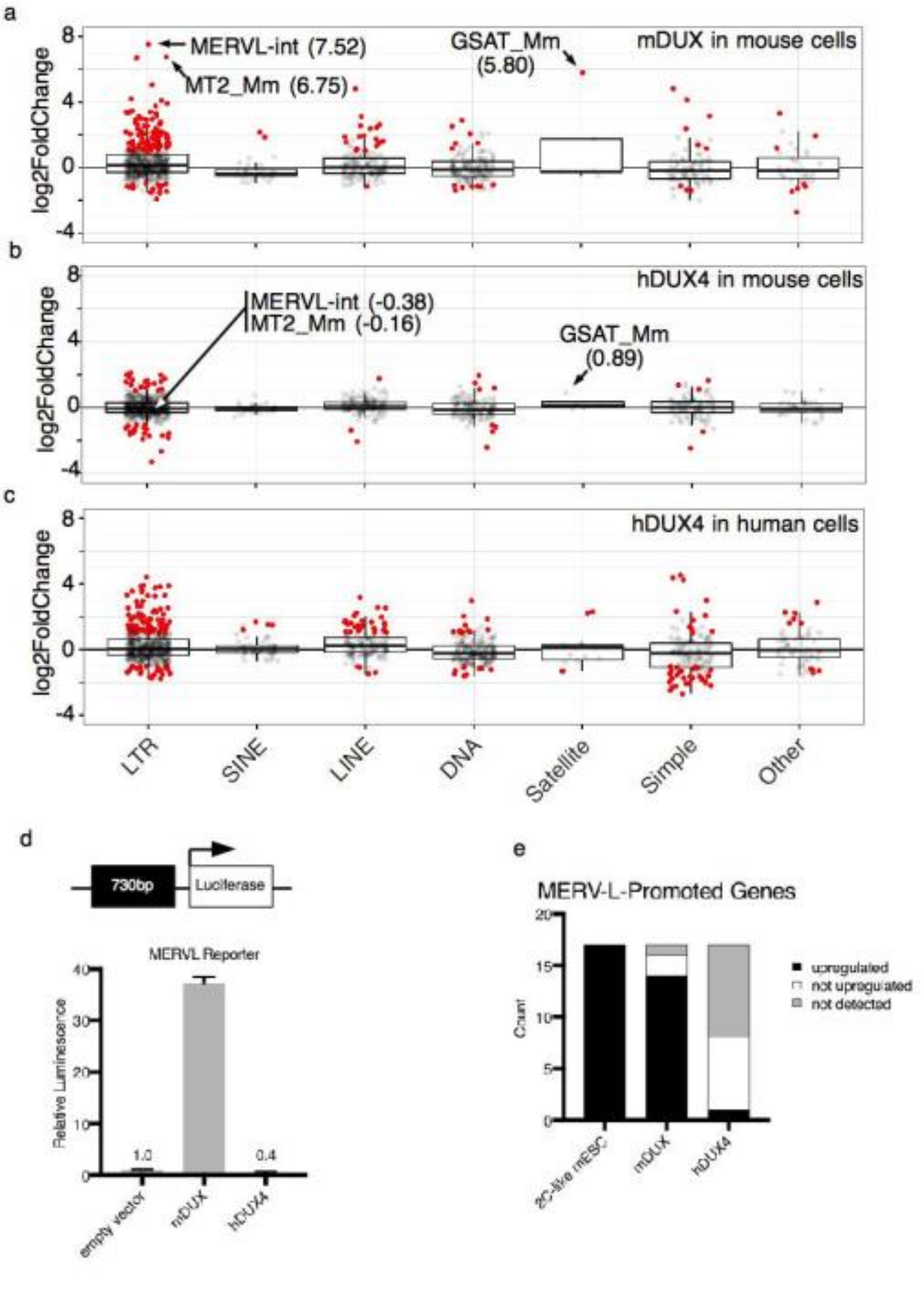

a
mDUX in mouse cells
MERVL-int (7.52)
MT2_Mm (6.75)
GSAT_Mm
(5.80)
b
hDUX4 in mouse cells
MERVL-int (-0.38)
MT2_Mm (-0.16)
GSAT_Mm
(0.89)
c
hDUX4 in human cells
log2FoldChange
8
4
0
-4
LTR
SINE
LINE
DNA
Satellite
Simple
Other
d
730bp
Luciferase
MERVL Reporter
Relative Luminescence
40
30
20
10
0
1.0
0.4
empty vector
mDUX
hDUX4
e
MERV-L-Promoted Genes
Count
20
15
10
5
0
upregulated
not upregulated
not detected
2C-like mESC
mDUX
hDUX4

Figure 4

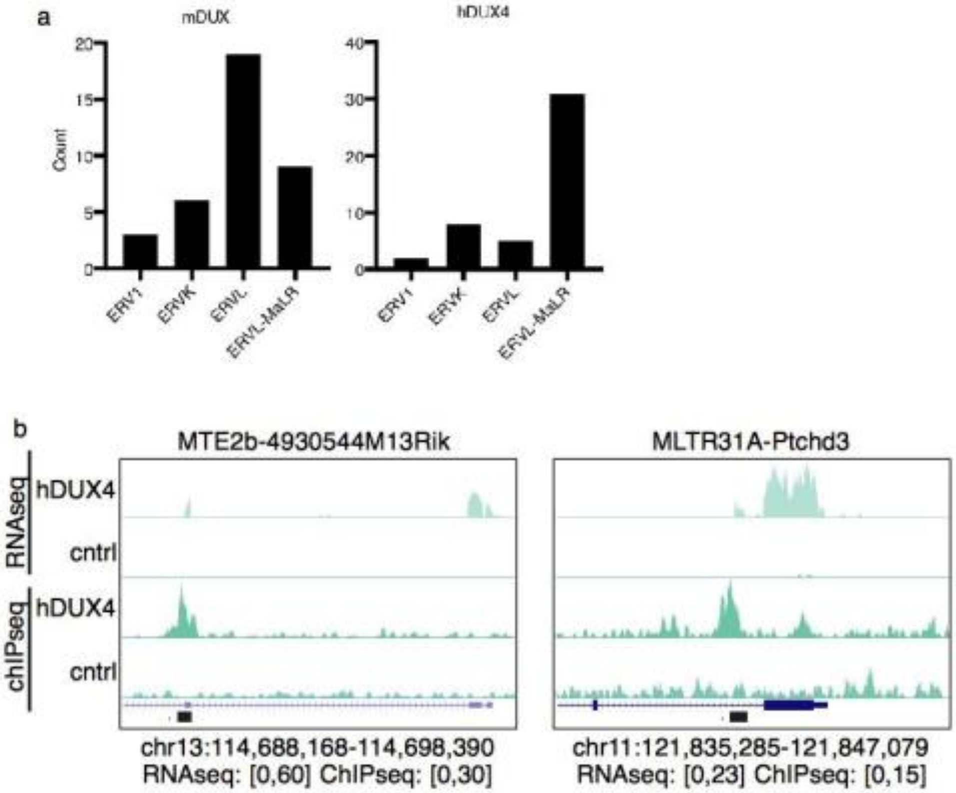
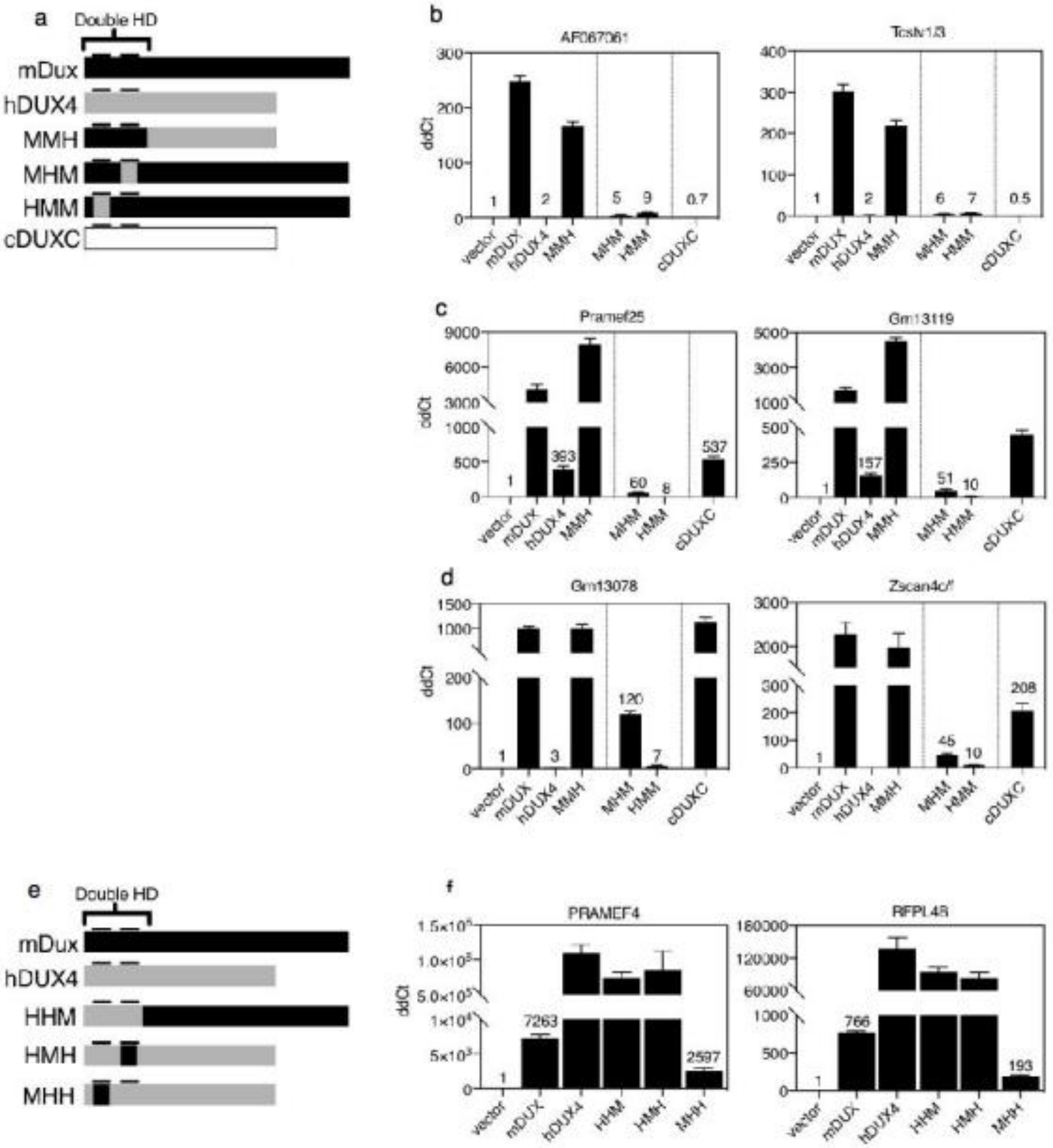
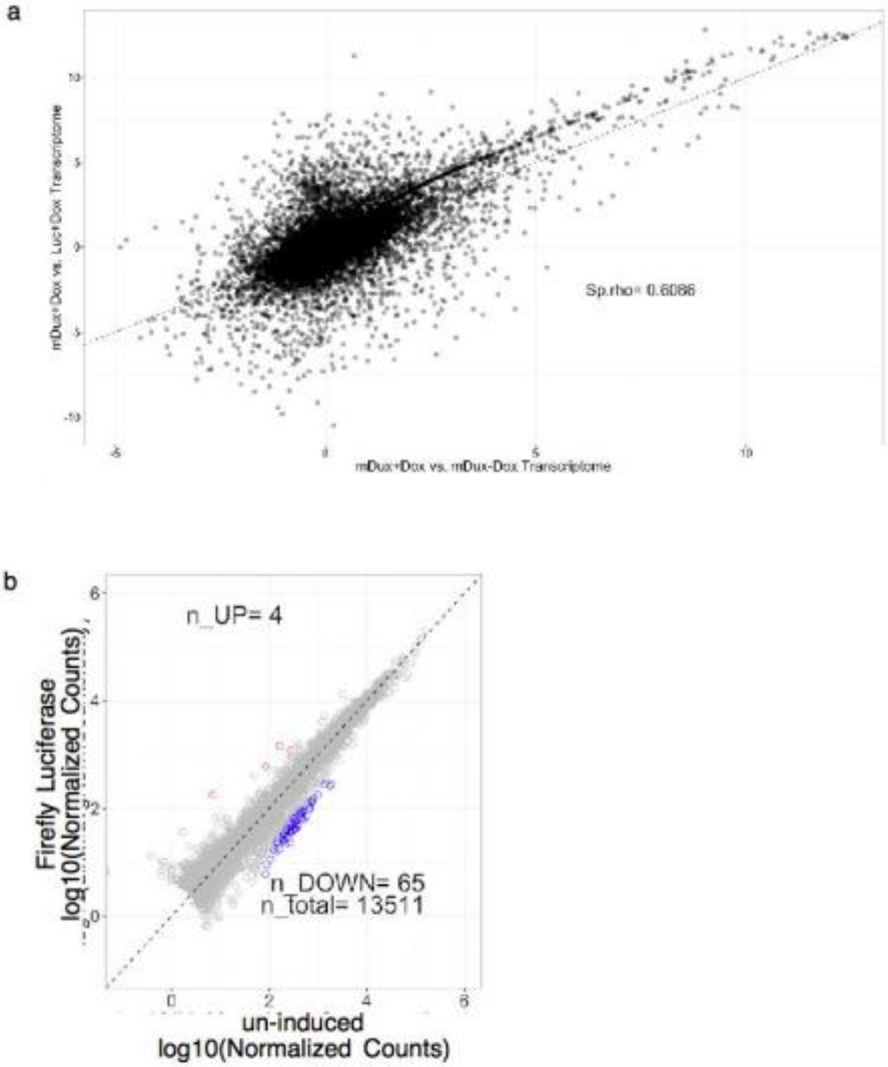

a
mDUX
hDUX4
Count
0
5
10
15
20
0
10
20
30
40
ERV1
ERVK
ERVL
ERVL-MaLR
ERV1
ERVK
ERVL
ERVL-MaLR
b
MTE2b-4930544M13Rik
MLTR31A-Ptchd3
RNAseq
hDUX4
cntrl
chIPseq
hDUX4
cntrl
chr13:114,688,168-114,698,390
RNAseq: [0,60] ChIPseq: [0,30]
chr11:121,835,285-121,847,079
RNAseq: [0,23] ChIPseq: [0,15]

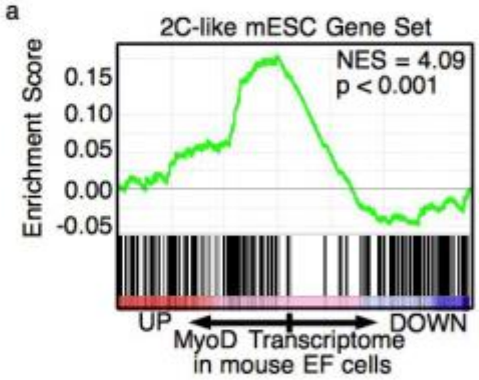
Figure 5


a
Double HD
mDux
hDUX4
MMH
MHM
HMM
cDUXC
b
AF067061
ddCt
300
200
100
0
1
2
5
9
0.7
vector
mDUX
hDUX4
MMH
MHM
HMM
cDUXC
Tcstv1/3
400
300
200
100
0
1
2
6
7
0.5
c
Pramef25
9000
6000
3000
1000
500
0
1
393
60
8
537
Gm13119
5000
3000
1000
500
250
0
1
157
51
10
d
Gm13078
1500
1000
200
100
0
1
3
120
7
Zscan4c/f
3000
2000
300
200
100
0
1
45
10
208
e
Double HD
mDux
hDUX4
HHM
HMH
MHH
f
PRAMEF4
1.5×10^5
1.0×10^5
5.0×10^5
1×10^4
5×10^3
0
1
7263
2597
vector
mDUX
hDUX4
HHM
HMH
MHH
RFPL4B
160000
120000
60000
1000
500
0
1
766
193

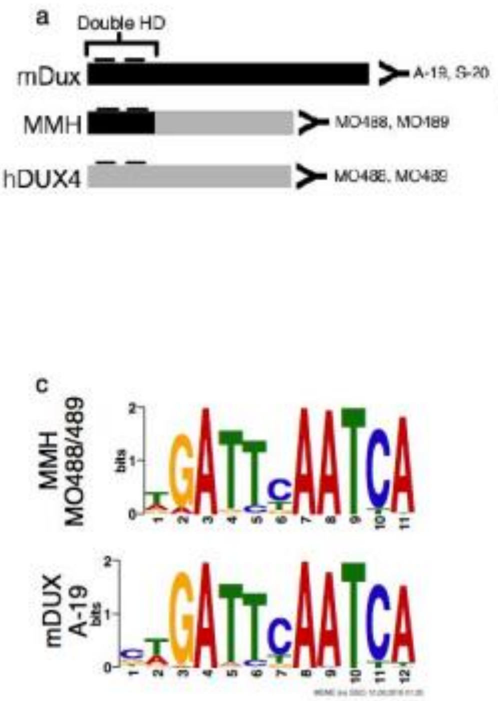
Supplementary Figure 1



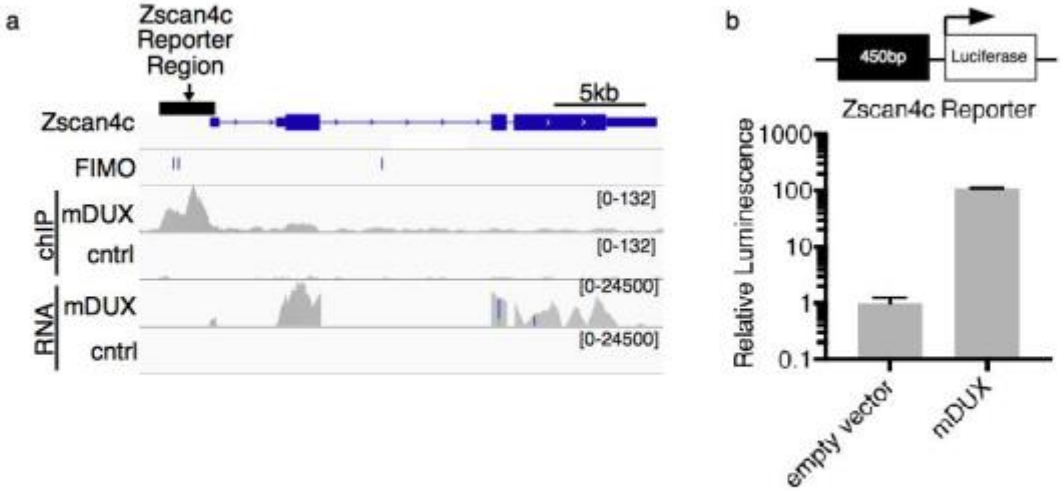
Supplementary Figure 2



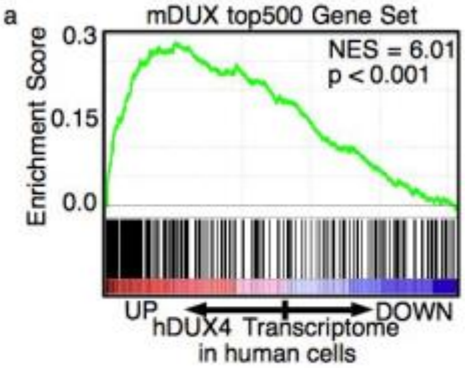
Supplementary Figure 3



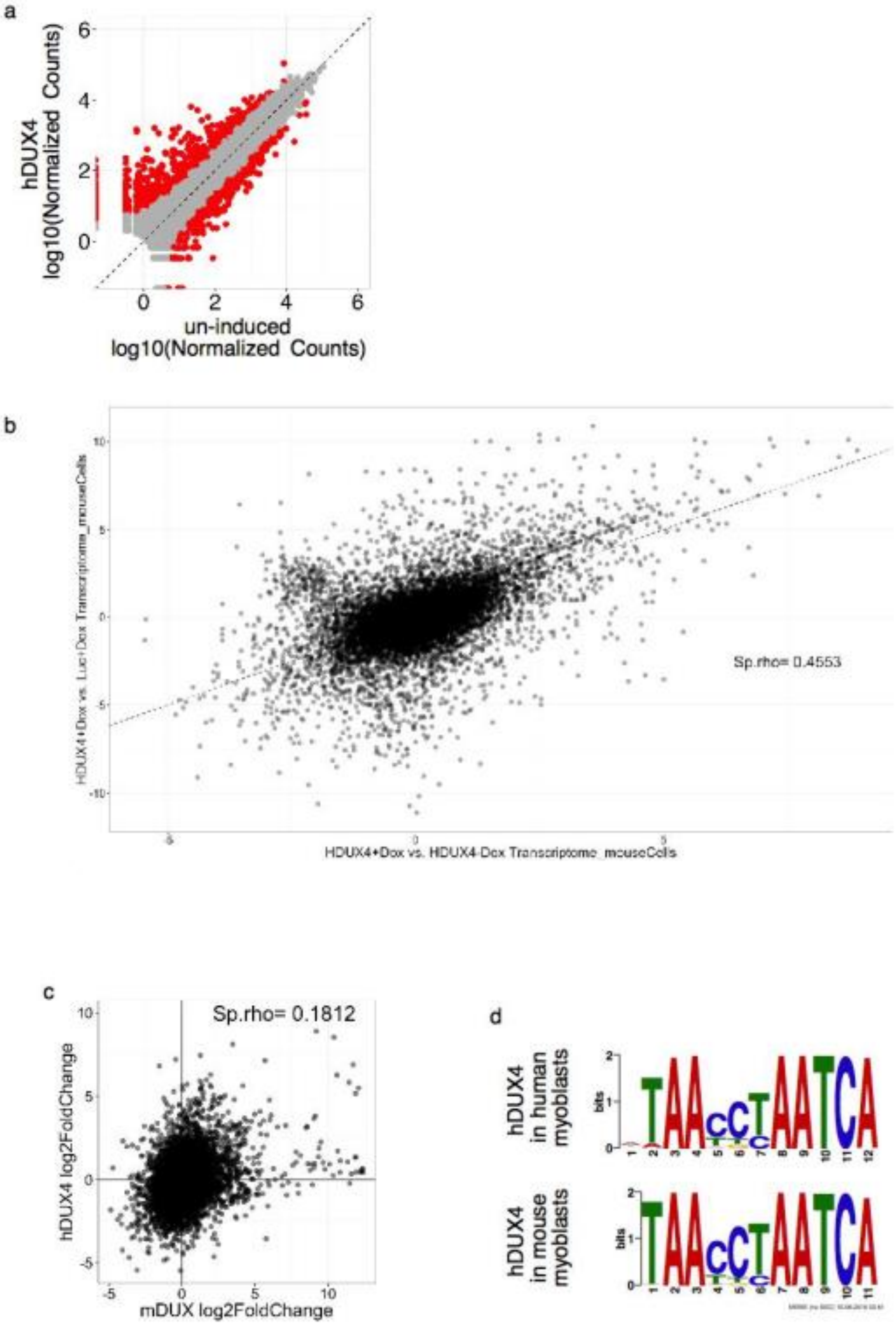
Supplementary Figure 4



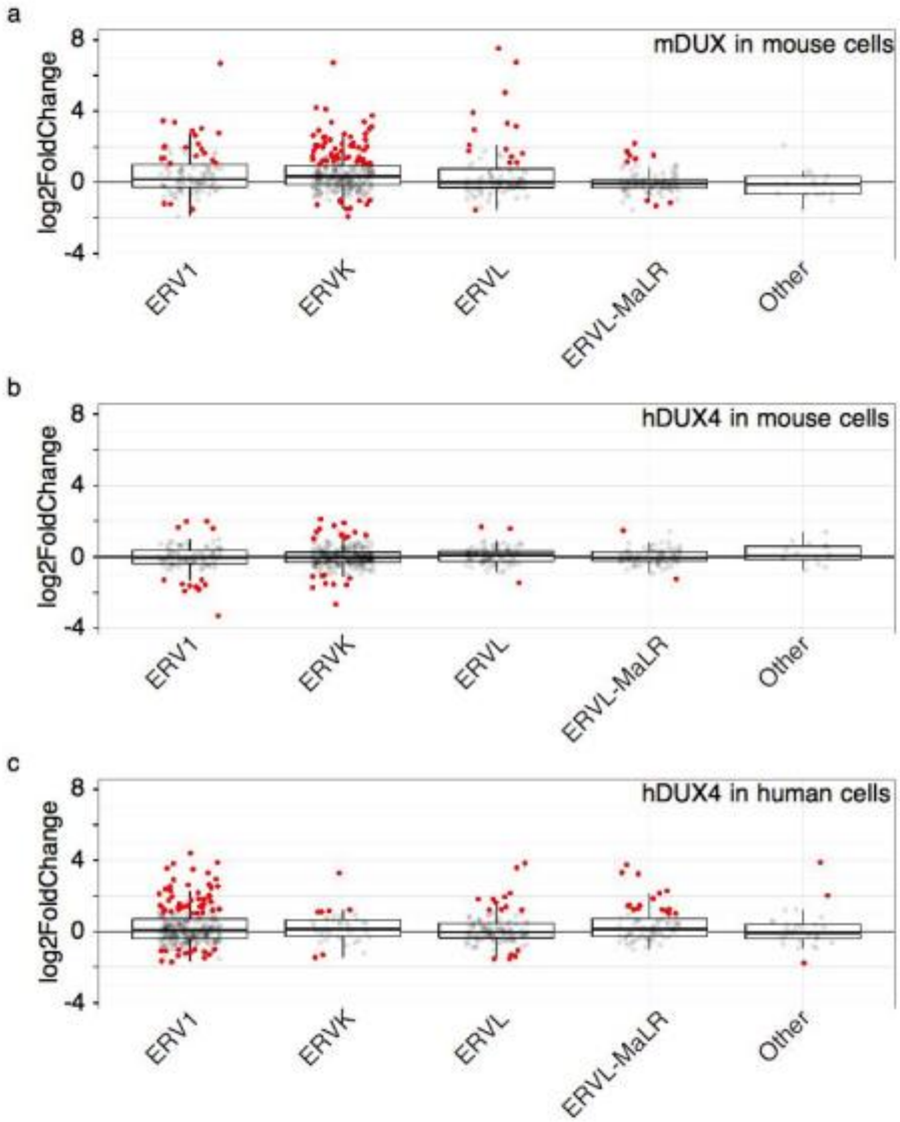
Supplementary Figure 5



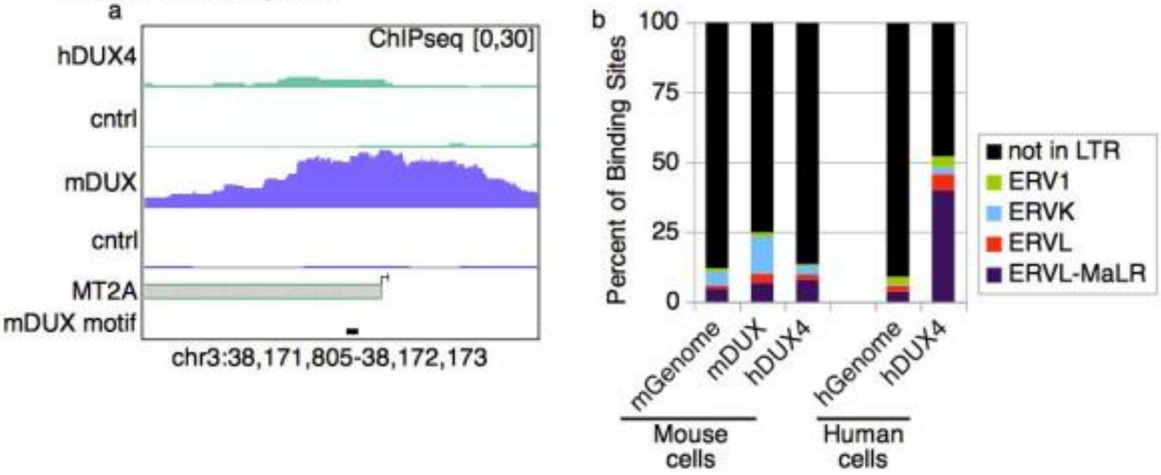
Supplementary Figure 6



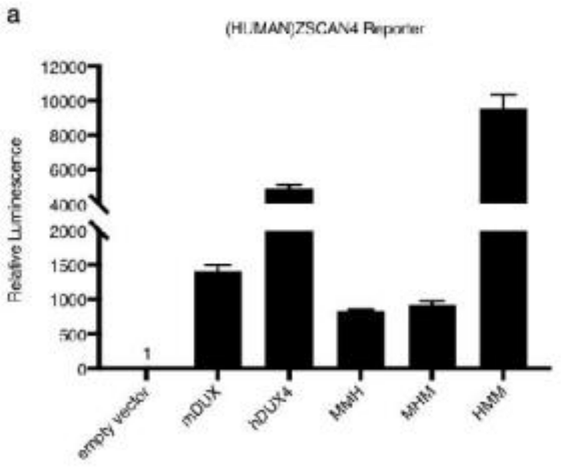
Supplementary Figure 7



Supplementary Figure 8



Supplementary Figure 10



Chapter 3. Functional domains of human DUX4 and mouse DUX

Introduction

This chapter aims to dig deeper into understanding the importance and relative capacities of the individual parts of two DUXC retrogenes, hDUX4 and mDUX. When these factors are compared by sequence, they have three regions of clear homology: homeodomain 1, homeodomain 2 and a stretch of about 50 amino acids at their extreme C-termini. Noticeable, is their extreme dissimilarity in regions outside of these three homologous regions. As if there is selective pressure on these proteins to change as much as possible without completely losing their essential functions, but this hypothesis remains difficult to test in the DUX family. This hypothesis is difficult to test because the DUXC retrogenes (and also DUXC genes outside of rodents and primates) are multi-copy tandemly arrayed genes where gene conversion, array homogenization, presence of multiple array loci and dispersed single-copy DUX pseudogenes (found in the human genome, but not the mouse genome) greatly affects predictions of positive selection.

The homeodomain is a well-defined domain. It is 60 amino acids long and comprised of an unstructured N-terminal tail and three alpha helices. Both the N-terminal tail and the third alpha helix (a.k.a. the recognition helix) are thought to make the most contribution to DNA-binding sequence specificity. The N-terminal tail binds the minor groove of DNA while the recognition helix nestles into the major groove. Since a single homeodomain in complex with DNA has been solved, we know that the first and second helices are critical to packing the recognition helix into the major groove by stacking on top of the recognition helix like cord wood, but the first and second helices are not known to contact DNA themselves. Concordantly, DUX homeodomain alignments show more conservation at the N-terminal tail and recognition helix and less conservation of the first and second helices. Although extensive work has been done to determine the "homeodomain code," that is to say the correspondence between amino acids in the recognition helix and their nucleotide-binding preferences. If the "homeodomain code" were fully understood, the promise was the ability to predict the consensus binding motif of DNA-binding proteins *a priori* with just the amino acid sequence of the homeodomain. Unfortunately, the data to support the "homeodomain code" is not consistent leading some groups to argue that there is not a clear grammar between amino acid sequence and nucleotide-binding preferences.

Of particular importance to the double homeodomain factors is whether the 3-dimensional structure of a single homeodomain has any bearing on the 3-dimensional structure of a double homeodomain. One can imagine the structure of a double homeodomain is simply two single homeodomain structures held in a particular orientation by the linker sequence between the two homeodomains. Alternatively, it is formally possible that the two homeodomains coordinate in their 3-dimensional folding to create a structure that is fairly distinct from the known single homeodomain structure. Similarly, differences between homeodomain 1 and homeodomain 2 within one DUX factor argue that the homeodomains likely confer different specificities such that each factor has a preferred orientation of binding. It is unknown whether the orientation is similar between factors such that the N-terminal homeodomain 1 interacts with the 5' part of the consensus motif or vice versa. It is fascinating to speculate in this regard that the mDUX consensus binding motif is palindromic while the hDUX4 consensus binding motif is not palindromic. Note, despite mDUX's palindromic consensus binding motif, the orientation of its homeodomains from N- to C-terminus is the same with the recognition helix of both homeodomain 1 and homeodomain 2 towards the C-terminus. Although our collaborators have tried to crystalize hDUX4 and determine its structure and binding orientation, the structure of any double homeodomain protein remains elusive.

Finally, the C-terminus of these DUX factors is thought to confer transcriptional activation, consistent with the function of an activation domain. Generally, activation domains are thought to have very little conformational specificity, rather they have been characterized colorfully as "acid blobs" or "negative noodles"(70). Consistently, both mDUX and hDUX4 have many acidic glutamic acid residues in their C-termini, offering the hypothesis that the C-termini are interchangeable between mDUX and hDUX4, which the data from this chapter support.

Results

hDUX4 activates several different ages of ERV reporter similarly well

In order to confirm RNA-seq based observations that hDUX4 activated transcription of several endogenous retroviruses, I had 375 basepairs of several ERVs synthesized and cloned these five ERVs individually into luciferase reporters with only a basic promoter. When I transiently transfected the reporters and an hDUX4-expression vector into mouse skeletal muscle cells or human rhabdomyosarcoma cells, I observed increased luciferase signal with hDUX4, but not an empty vector (Figure 11). Another goal of this experiment was to assess how well ERVs that integrated into the genome at various times in the past (i.e. young age == new insertion, old age == ancient insertion) were activated by hDUX4. The big hypothesis was that ER repeats were evolving towards binding hDUX4, such that young ERVs would be more responsive to hDUX4 than old ERVs. In this experiment, THE1-ERVs are the youngest, but they are not the strongest induced by hDUX4. These data do not support our hypothesis.

mDux cannot activate an old-aged MLT1D reporter

I took the repeat that was the oldest insertion that might have been present in the genome of the last common ancestor of mice and humans and asked whether mDUX could activate this old ERV. However, the promoter sequence in MLT1D was not sufficient to drive luciferase expression when mDUX was present (although it was sufficient when hDUX4 was present). In Figure 12a, starting from the left side, compare the green bar (negative control) to the first blue bar (hDUX4 – tall bar) and the first orange bar (mDUX – short bar). mDUX's failure to activate MLT1D reporter could be for a variety of reasons. Two possible explanations are a failure to bind the sequence of MLT1D and a failure to interact with human co-factors necessary for activation. The predicted hDUX4 binding site in MLT1D is TAACTTAATCA. We generated several chimeras and performed the experiment in mouse cells as well as human cells to distinguish between these two explanations. All the data I generated is consistent with a problem in the mDUX double homeodomain region. A chimera that has the mouse homeodomains and human C-terminus cannot activate MLT1D – the mouse homeodomains break hDUX4. Conversely, a chimera with the human homeodomains and the mouse C-terminus can activate MLT1D – the human homeodomains confer activity to mDUX. Given that most homeodomains mediate DNA-binding, mapping the problem to the double homeodomain region strongly suggests a discrepancy in preferred binding motifs between hDUX4 and mDUX. Some homeodomains, however, mediate protein-protein interactions, so this cannot be ruled out as being part of the problem. I also performed this experiment in mouse cells, so that mDUX would have access to any mouse co-factors, but mDUX remained incapable of driving the MLT1D reporter (Figure 12b).

Both mDUX and hDUX4 activate a hZSCAN4 reporter

Given the results with a retrotransposon-based reporter, I next asked if we would get the same results from a promoter that does not include a retrotransposon. I used the strongly hDUX4-responsive reporter based on the putative promoter of the human ZSCAN4 gene, established by previous work in the Tapscott Lab (22). This reporter contains four predicted hDUX4 binding sites, two each of the following sequences: TAATTCAATCA, TAAATCAATCA. In contrast to the results with the MLT1D reporter, hDUX4 and mDUX both activate the hZSCAN4 reporter, although to varying extents across cell types. In human rhabdomyosarcoma cells, mDUX and all chimeras with the mDUX homeodomains activate the hZSCAN4 reporter about 10-fold less than hDUX4, but still 100-fold over background (Figure 13a). In mouse skeletal muscle cells, the levels of activation between mDUX and hDUX4 are not significantly different (Figure 13b). One caveat is that I have not ruled out a cell-type specific result. To bolster support of the species-specific differences hypothesis, one would need to repeat this experiment in another human cell type and observe the 10-fold difference between mDUX and hDUX4 again.

hDUX4 and mDUX C-termini have similar transactivation potential

If the differences between mDUX and hDUX4 I saw in reporter assays are due to differences in the homeodomain region, then controlling for DNA-binding should show similar abilities of the C-termini to activate a reporter. To test this, I created chimeras with the GAL4 DNA-binding domain in place of the double homeodomain regions of mDUX and hDUX4 and tested the chimeras on a GAL4-responsive reporter. Consistent with our hypothesis, mDUX-based and hDUX4-based chimeras have very similar ability to activate the GAL4-reporter (Figure 14a-b).

LEUTX is not a competitive inhibitor of hDUX4 nor mDUX

Taken together, these reporter data strongly implicate differences in the double homeodomain regions as the explanation for the 10-fold difference between mDUX and hDUX4 on the hZSCAN4 reporter in human cells.

When these experiments were done initially, we did not know the consensus binding motif of mDUX, but my ChIP-seq data described in Chapter 2 provided this advance. Revisiting these reporter assay data in light of the ChIP-seq data supports the hypothesis that mDUX and hDUX4 might have different activation abilities on the hZSCAN4 reporter due to differences in binding preferences. While hDUX4 has four predicted binding sites in the hZSCAN4 reporter, mDUX is predicted to bind only two of the four based on scoring the motifs with a position weight matrix of the mDUX consensus binding motif. Number of binding sites has been correlated with transactivation potential for hDUX4 (23), in that genes with more predicted hDUX4 binding sites are more likely to be activated than genes with fewer predicted hDUX4 binding sites. However, this mechanism cannot explain why there is no difference between mDUX and hDUX4 on this same reporter when the experiment occurs in mouse cells. If this is truly a species-specific effect, what is different about human cells and all the human proteins therein and mouse cells and the mouse proteins therein?

Preliminary observations of a DUX-like protein in human rhabdomyosarcoma cells (RD) pointed to a possible explanation. If hDUX4 binds the reporter with higher affinity than mDUX, hDUX4 (and not mDUX) could be capable of evicting a competitive inhibitor that is present in RD cells, but not in mouse cells. Previous work in the Tapscott Lab implicated at least two dominant negative proteins that could compete with hDUX4 and is specific to human cells, both protiens contains the hDUX4 double homeodomains, but lack the transactivating C-terminus (i.e. hDUX4c and hDUX4-short, which is also called hDUX4-spliceA). I also tested LEUTX for a dominant negative effect. LEUTX is a single homeodomain factor that was lost in mice, but retained in humans about which very little is known. From RNA-seq in the Tapscott Lab, LEUTX is bound by hDUX4 near its transcription start site and LEUTX RNA levels increase following hDUX4 expression, consistent with it being a direct target of hDUX4.

As was shown previously, reporter assays that contain equal amounts of both hDUX4 and hDUX4-short resulted in a drop in hZSCAN4-reporter. However, equal amounts of hDUX4 and hDUX4c did not show this effect, but it is unclear whether the explanation is technical or biological. Similarly, co-expression of LEUTX and hDUX4 did not show a dominant negative effect. A similar pattern was with mDUX and these candidate dominant negative proteins. Co-expression of mDUX and hDUX4-short showed inhibition of hZSCAN4-reporter activation, but mDUX activation of the hZSCAN4 reporter was not affected by co-expression of LEUTX. One caveat of the co-expression studies is that wells with one expression vector had twice as much expression vector as wells with two expression vectors, but this technical problem would lead to a false positive result, not a false negative.

Conserved C-Terminal LxxLL Motifs are critical for transactivation

This series of chimeras and reporter assays also aimed to assess the important of the C-terminal LxxLL motifs. Within the double homeobox genes, DUXC genes and retrogenes are defined as those with homology at the C-terminus with other DUXC factors because DUXA and DUXB lack this conserved stretch of residues at the C-terminus(18). This observation begs the question as to the importance and function of these residues, a prominent feature of which are two LxxLL motifs (Figure 16). The LxxLL motif has been described as "a multifunctional binding sequence in transcriptional regulation" (71). Mere presence of an LxxLL pattern does not universally indicate a functional site, but in the case of the DUXC genes and residues, conservation across diverse species is a strong indicator of functionality. Some of the protein-protein interactions that have been described as mediated by LxxLL-motifs are generally transcription factors and their co-activators, but LxxLL motifs can also mediate repression. I deleted the last 55 amino acids from hDUX4 and found that this truncation mutant could not activate the MLT1D reporter nor the hZSCAN4 reporter (Figure 12 and 13; bar labeled "deltaC1C2"). Consistent with a conserved function, adding the homologous residues from mDUX restored transactivation (Figure 12 and 13; bar labeled "mC1C2"). A very recent study demonstrated recruitment of CBP/p300 to hDUX4 via the C-terminus, such that C-terminally deleted hDUX4's did not recruit CBP/p300(72). Further studies will be needed to determine whether the LxxLL motifs specifically mediate this interaction and whether any other protein-protein interactions are mediated by the LxxLL motifs.

Discussion

Homeodomain proteins are intriguing because of their well-established roles in various aspects of development and therefore have been studied extensively. In my mind a glaring exception to this general rule is the family of double homeodomain proteins. Of the three lineages within the DUX family (DUXA, DUXB, DUXC), DUXC has garnered the most attention but only one of its retrogenes: hDUX4. Such that we now have a preponderance of data surrounding hDUX4's mis-expression in skeletal muscle. However, the fundamental question remains as to the normal function of hDUX4 and its connection to other DUXC genes and retrogenes. This chapter took a reductionist approach to determining the functions of hDUX4 and another DUXC retrogene, mDUX. Using a panel of reporters designed with hDUX4 in mind, I compared the activities of each retrogene and created a series of chimeras that swapped various domains between the two proteins to determine which domains were interchangeable between the factors and which were specific to each factor.

One key finding from this work was the discovery that the double homeodomain region conferred specificity while the C-Terminus was quite interchangeable. This strongly implicates DNA-binding differences between the factors as the main determinant of their activities. One possibility that this work has not ruled out is that a species-specific protein-protein interaction is necessary at some transcriptional targets. For example, I observed mDUX activates the hZSCAN4 reporter 10-fold less than hDUX4 in human cells, but this difference is diminished when the experiment occurs in mouse cells. There is a species-specific (formally it could be a cell-type specific) difference whereby mDUX is prevented from its full potential in human cells/RD cells or mDUX is missing a protein-protein interaction partner that is necessary for it to transactivate targets to its full potential. The site of this interaction maps to the double homeodomain region, which in addition to supporting DNA-binding has also been shown to be a site of protein-protein interactions. Since this work was done on reporter plasmids, this particular difference I observed is unlikely to involve differences in binding to chromatin or interacting with chromatin remodelers as reporter plasmids are non-chromatinized. However, other work certainly implicates a close relationship between DUX factors and chromatin as both mDUX and hDUX4 bind and activate genes and repetitive elements in somatic cells that are typically chromatin-inaccessible in somatic cells.

Furthermore, the work in this chapter predicts that C-terminal protein-protein interactions discovered for one factor likely are shared by the other factor. For example, although CBP/p300 was found to interact with the hDUX4 C-terminus, it seems likely that mDUX can interact with mouse CBP/p300 and the human orthologs of CBP/p300 when mDUX is forcibly expressed in a cross-species manner. Further study is needed to determine what other factors may interact with the C-termini of the DUX factors and whether these interactions are specific to the LxxLL motif. In this regard it is interesting to note that many, but not all, DUXC genes and retrogenes have two LxxLL motifs at their extreme C-termini. That the second, more internal, LxxLL is not strictly conserved could indicate that it is redundant in function to the extreme C-terminal LxxLL motif, or it could indicate that different species support different protein-protein interactions with the different number of LxxLL motifs between different species, but further experiments will be needed to distinguish between these possibilities.

Materials and Methods

Cell Culture

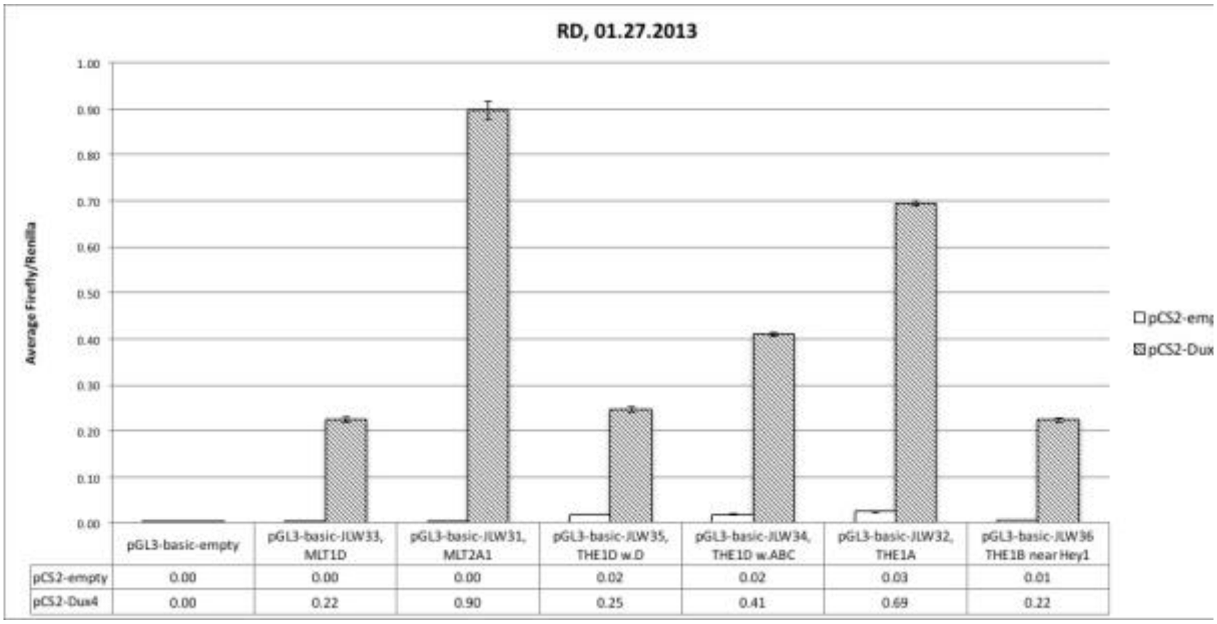
Human rhabdomyosarcoma (RD) cells were grown in DMEM in 10% bovine calf serum (Hyclone) and 1% penicillin/streptomycin. Mouse skeletal muscle cells (C2C12) were grown in DMEM in 10% fetal calf serum (Hyclone) and 1% penicillin/streptomycin.

Luciferase Assay

Transient DNA transfections of RD and C2C12 cells were performed using SuperFect (QIAGEN) according to manufacturer specifications. Briefly, 80,000 C2C12 or 300,000 RD cells were seeded per 6-well dish the day prior to transfection. Cells were co-transfected with pCS2 expression vectors (1 ng/plate) carrying either hDUX4 or mDUX or a chimera of the two and with pGL3-basic reporter vector (1 ng/plate) carrying test promoter fragment upstream of the firefly luciferase gene. Cells were lysed 24-hours post-transfection in Passive Lysis Buffer (Promega). Luciferase activities were quantified using reagents from the Dual-Luciferase Reporter Assay System (Promega) following manufacturer's instructions. Light emission was measured using BioTek Synergy2 luminometer. Luciferase data are given as the averages ± SD of at least triplicates.

Figure 11. hDUX4 activates several reporters using ERV sequences

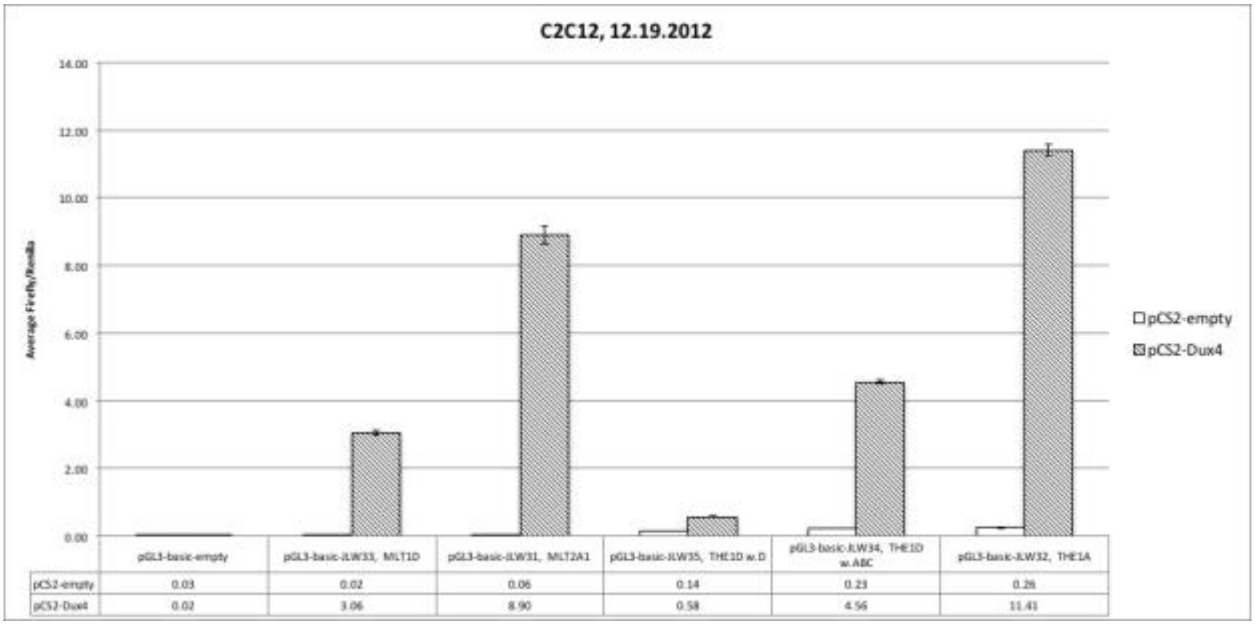
A



RD, 01.27.2013

	pGL3-basic-empty	pGL3-basic-JLW33, MLT1D	pGL3-basic-JLW31, MLT2A1	pGL3-basic-JLW35, THE1D w.D	pGL3-basic-JLW34, THE1D w.ABC	pGL3-basic-JLW32, THE1A	pGL3-basic-JLW36 THE1B near Hey1
pCS2-empty	0.00	0.00	0.00	0.02	0.02	0.03	0.01
pCS2-Dux4	0.00	0.22	0.90	0.25	0.41	0.69	0.22

B



C2C12, 12.19.2012

	pGL3-basic-empty	pGL3-basic-JLW33, MLT1D	pGL3-basic-JLW31, MLT2A1	pGL3-basic-JLW35, THE1D w.D	pGL3-basic-JLW34, THE1D w.ABC	pGL3-basic-JLW32, THE1A
pCS2-empty	0.03	0.02	0.06	0.14	0.23	0.26
pCS2-Dux4	0.02	3.06	8.90	0.58	4.56	11.41

Figure 12. mDUX cannot activate MLT1D reporter

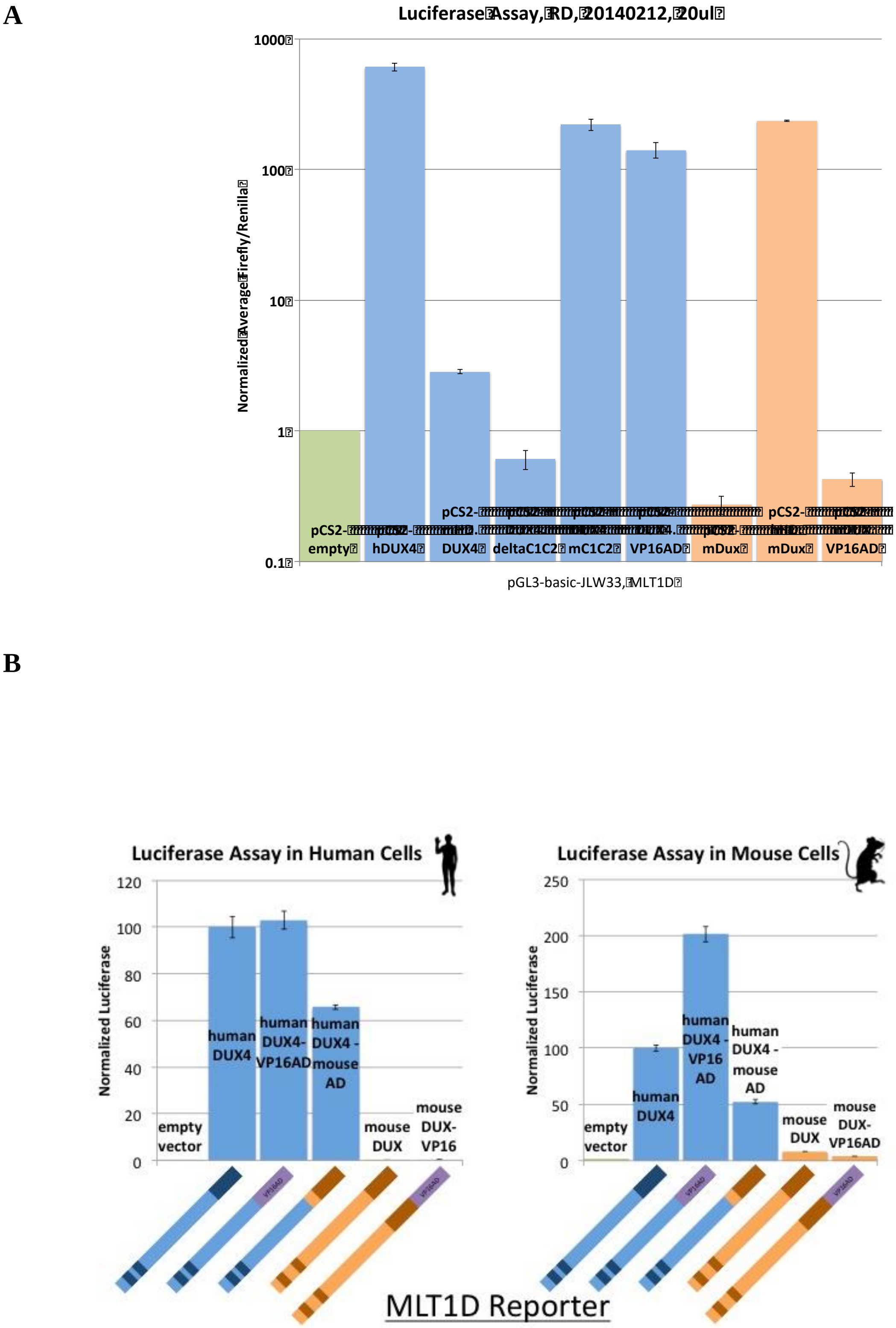
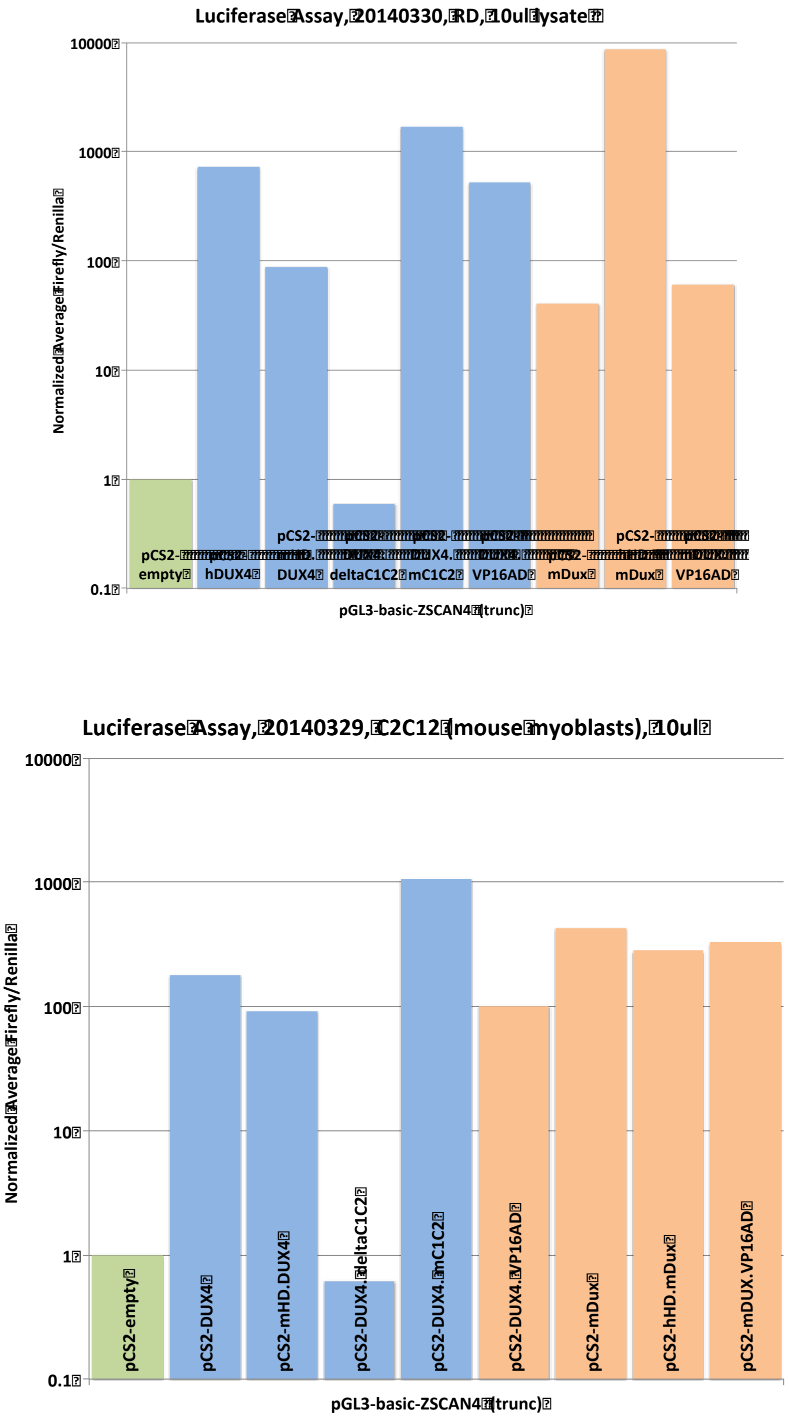


Figure 13. mDUX and hDUX4 both activate hZSCAN4 reporter



Luciferase Assay, 20140330, RD, 10ul lysate

Luciferase Assay, 20140329, C2C12 (mouse myoblasts), 10ul

Figure 14. Without the HD, mDUX and hDUX4 activate Gal4-reporter similarly well

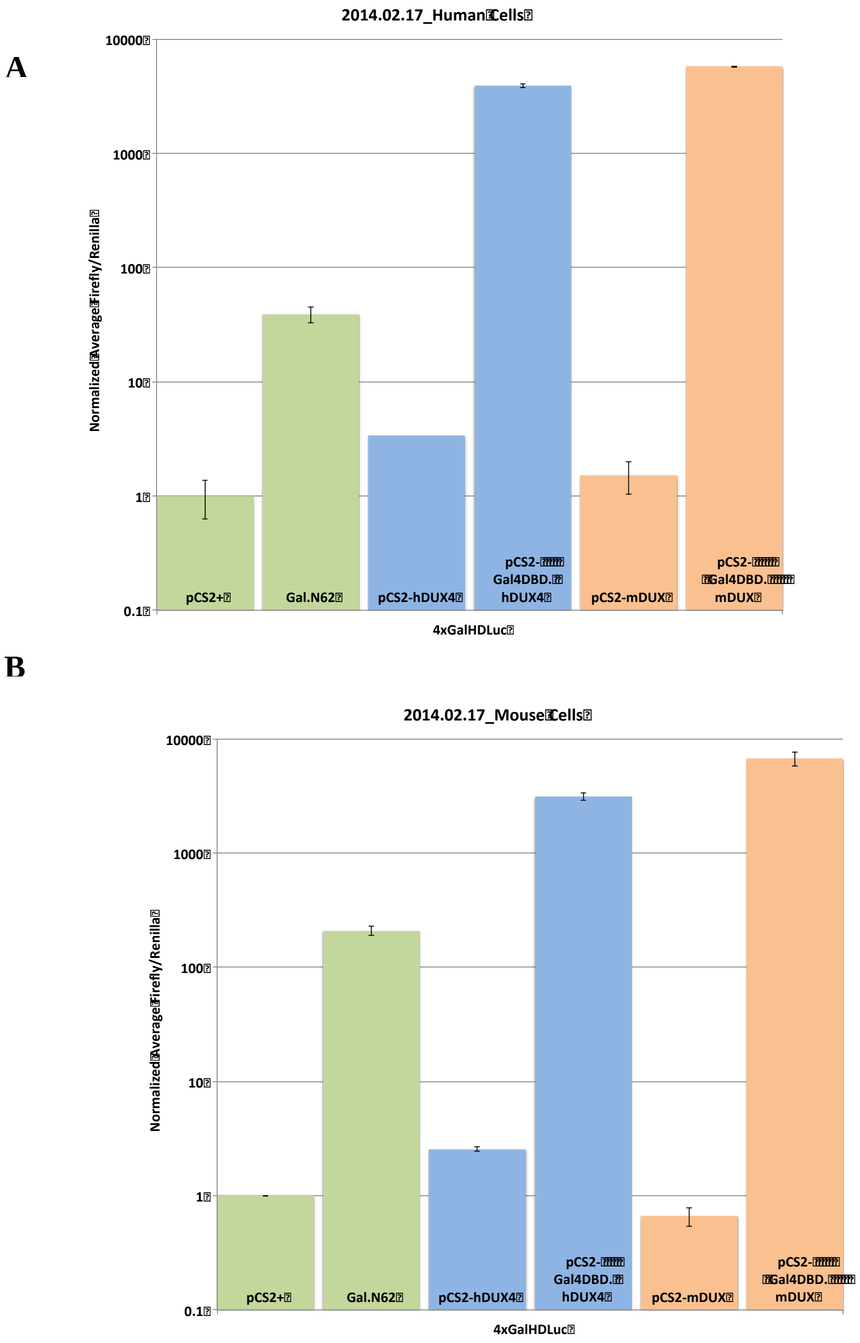
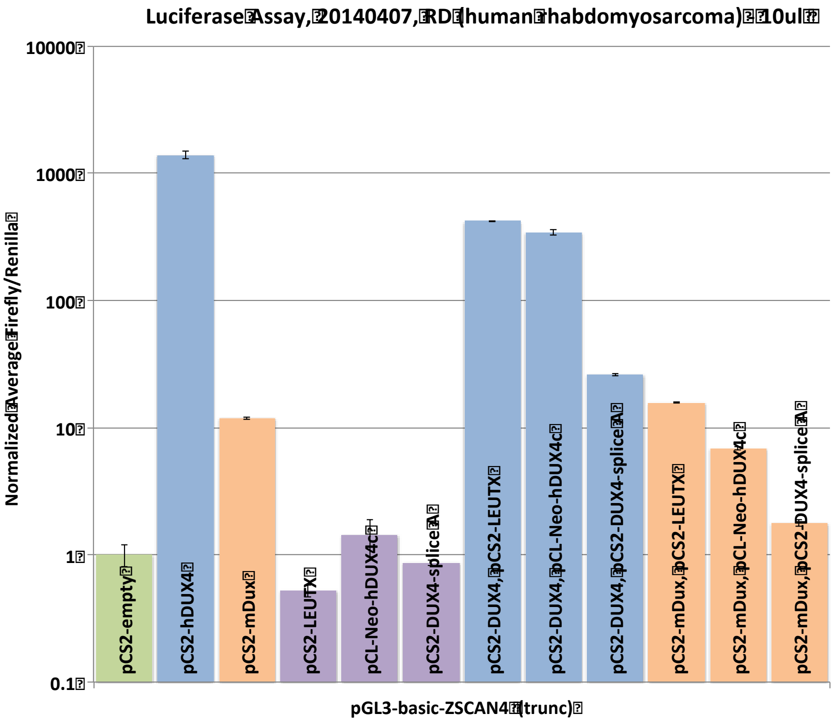


Figure 15. hDUX4-short, but not LEUTX nor hDUX4c, show competitive inhibition of hDUX4

A



B

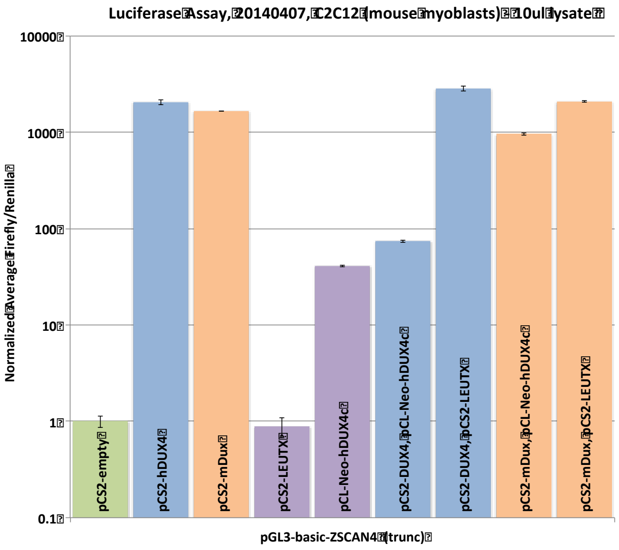
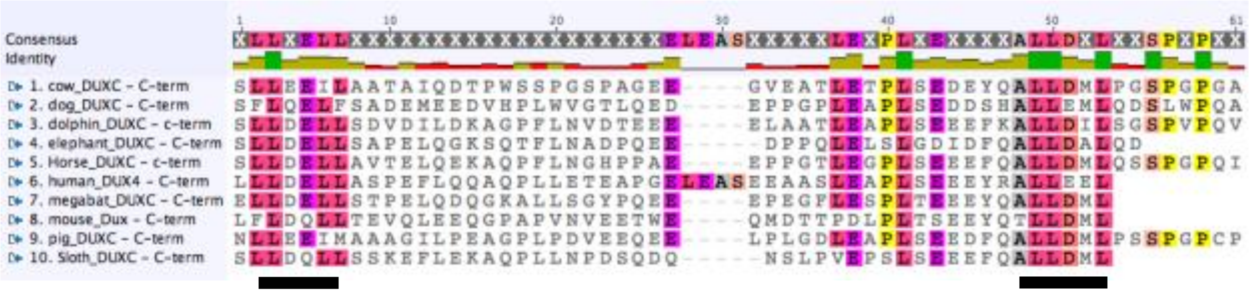


Figure 16 LxxLL motifs in conserved C-termini of DUXC genes and retrogenes



```
                               1        10        20        30        40        50        61
Consensus                      XLLXELLXXXXXXXXXXXXXXXXXXXELEASXXXXXLEXPLXEXXXXALLDXLXXSPXPXX
Identity
1. cow_DUXC - C-term           SLLEEILAATAIQDTPWSSPGSPAGEE----GVEATLETPLSEDEYQALLDMLPGSPGPGA
2. dog_DUXC - C-term           SFLQELFSADEMEEDVHPLWVGTLQED----EPPGPLEAPLSEDDSHALLEMLQDSLWPQA
3. dolphin_DUXC - c-term       SLLDELLSDVDILDKAGPFLNVDTEEE----ELAATLEAPLSEEEFKALLDILSGSPVPQV
4. elephant_DUXC - C-term      SLLDELLSAPELQGKSQTFLNADPQEE-----DPPQLELSLGDIDFQALLDALQD
5. Horse_DUXC - c-term         SLLDELLAVTELQEKAQPFLNGHPPAE----EPPGTLEGPLSEEEFQALLDMLQSSPGPQI
6. human_DUX4 - C-term         LLLDELLASPEFLQQAQPLLETEAPGELEASEEAASLEAPLSEEEYRALLEEL
7. megabat_DUXC - C-term       ELLDELLSTPELQDQGKALLSGYPQEE----EPEGFLESPLTEEEYQALLDML
8. mouse_Dux - C-term          LFLDQLLTEVQLEEQGPAPVNVEETWE----QMDTTPDLPLTSEEYQTLLDML
9. pig_DUXC - C-term           NLLEEIMAAAGILPEAGPLPDVEEQEE----LPLGDLEAPLSEEDFQALLDMLPSSPGPCP
10. Sloth_DUXC - C-term        SLLDQLLSSKEFLEKAQPLLNPDSQDQ-----NSLPVEPSLSEEEFQALLDML
```

Chapter 4. Isoforms of Intron-containing DUXC Genes

Introduction

This chapter is motivated by the observation that a DUXC gene and retrogene have not been found in the same genome, leading to the hypothesis that DUXC genes and DUXC retrogenes are functional homologs. However, no one has characterized any DUXC gene previously. This chapter details work surveying DUXC genes and focuses on canine DUXC as an exemplar of the ancestral DUXC. Inherent in this strategy is a major assumption that remains to be tested, namely that DUXC from canine has not changed markedly from ancestral DUXC and thus can act as a faithful proxy for the ancestral state. In order to test this hypothesis any findings based on canine DUXC should be replicated in additional DUXC from other species, thereby bolstering support for the notion that the findings are general to all DUXC genes and not specific to canine DUXC.

Results

Despite extensive searching, I was unable to identify any DUXC gene or retrogenes in such a species (i.e. lagomorphs – rabbits, squirrels, guinea pig), but this could be simply because their genomes are less well-assembled than primates and rodents and DUXC genes are retrogenes are found in multi-copy arrays that are difficult to assemble(20).

Survey of DUXC Orthologs across the Tree of Life

As a first step to broadly characterizing DUXC genes, I set out to identify the entire open reading frame (ORF) for each species that contains a DUXC gene. The homeodomain sequences for many DUXC genes were published previously (15, 18, 20), but the ORFs were not published. I used tBLASTn searches with the published concatenated homeodomains to identify the putative ORFs and required at least one LxxLL motif in the C-terminus of the predicted protein (Figure 18). This strategy will identify genes that have two homeodomains and the LxxLL motif, but there is some evidence to support alternative splicing in this family that could impact one homeodomain and/or the C-terminus, and this strategy will not identify these isoforms. Thus, this list should be taken as conservative and likely to expand with further studies that are robust to transcripts actually produced by the DUXC loci in these species.

Identification of canine DUXC "interruptedHD" isoform

With the recognition that predicted transcripts and actual transcripts may vary, I focused on amplifying DUXC transcripts from a species for which we had tissue samples from a tissue where it seemed likely to find DUXC transcripts, namely adult canine testes. Recall that hDUX4 is expressed in individuals that are unaffected by FSHD, so we figured that this expression pattern may be conserved in dog and canine testes are readily available as many dogs are neutered. Surprisingly, although I was able to amplify transcripts of several slightly different isoforms from canine testes, none of the isoforms I observed in canine testes ever contained two complete homeodomains (Figure 19). Rather, all isoforms included 180 nucleotides that interrupt the predicted homeodomain sequence; however, the reading frame is not altered by the interruption such that the remainder of the transcript encodes the "correct" amino acids, including the LxxLL motifs at the extreme C-terminus (Figure 20).

This observation raised several intriguing questions. Is the predicted cDUXC_intactHD expressed in a different canine tissue or developmental stage? Is there a difference in function between the cDUXC_intactHD and the cDUXC_interruptedHD? Do other species create multiple isoforms of their DUXC gene, such that this could be a general attribute of DUXC genes? If regulation of an intactHD and an interruptedHD isoform is common to all DUXC genes and thus supposed to be the ancestral state, how did retroposition affect the function of DUXC retrogenes which necessarily cannot toggle between two isoforms as they no longer have introns?

Prediction of DUXC genes with "interruptedHD" isoforms in non-canine species

In order to address the question as to whether the interruptedHD isoform was representative of the ancestral state (inferred from presence in multiple species) or specific to canine and thus not representative of the ancestral state, I used the GenScan algorithm to predict transcripts from the genomic DUXC locus in non-canine species. Since transcript predictions are just that: predictions, any interesting predicted transcripts will need to be confirmed through RNA analysis from tissue samples. First I used tBLASTn searches to identify the DUXC locus in six species with DUXC genes (not retrogenes): armadillo, sloth, pig, horse, dolphin and megabat. I used the concatenated hDUX4 homeodomains and/or the concatenated mDUX homeodomains as queries. Although four of these species were not predicted to encode an intra-homeodomain intron (an intron whose inclusion would create an interruptedHD isoform), two species did: horse and megabat (Figure 21).

In addition to mere presence of an interruptedHD isoform in additional non-canine species, stronger support for ancestral regulation of a interruptedHD DUXC isoform would be to show that all the interruptedHD isoforms predicted in various species have homologous sequence. If this were true, it would strongly support the interrupted isoform being inherited from a common ancestor of these species. On the other hand, the extraordinarily low sequence conservation between mDUX, hDUX4 and cDUXC outside of the homeodomains and 50 residues at the C-terminus might indicate that there is very low selective pressure on parts of the gene outside of these few critical domains. Furthermore, I predicted that the function of the homeodomain interruption is to basically break the first homeodomain's DNA-binding ability and force the protein into a single homedomain binding paradigm, then there may very well be very low pressure to maintain a particular sequence to the interruption. Indeed, the interruptedHD isoform of megabat interrupts the second homeodomain, while dog and horse interrupt the first homeodomain. Additionally, I could not find any sequence homology between the any of the predicted interruptedHD isoforms from horse and megabat and the interruptedHD isoform of canine DUXC. Therefore, it is currently an open question as to whether the interrruptedHD isoform I observed expressed in dog testes was inherited or specific to dogs. Future studies should consider an RT-PCR survey of testes from common domestic or agricultural animals such as cow and/or horse that are castrated on a regular basis.

Canine DUXC intactHD and interruptedHD isoforms have different transactivational abilities

Although the inclusion of 60 amino acids within homeodomain one of cDUXC seems like it would disrupt the conformation necessary for homeomdomain-DNA binding, it is formally possible that the interrupting amino acids could fold out of the way and the so-called "interruptedHD" isoform would bind DNA equally as well as the intactHD isoform. In order to test whether the intactHD and interruptedHD have similar or different DNA-binding capacities and thus would be predicted to activate different (although possibly overlapping) transcriptomes such that a toggle between expressed isoform would confer a functional change, I created two expression constructs that differ only in the inclusion/exclusion of the 60 amino acids that I observed to interrupt the first homeodomain in canine testes. As a first pass, I expressed these constructs individually in mouse skeletal muscle cells and queried loci affected by mDUX expression – effectively this asked two questions simultaneously: 1) Are any transcriptional targets of mDUX conserved with cDUXC? 2) Do cDUXC intactHD and interruptedHD isoforms have similar activities on these targets? Although a negative result would have been fairly uninformative, I found that there was a clear difference in the activities of cDUXC_intactHD and cDUXC_interruptedHD. cDUXC_intactHD was able to transcriptionally activate three mouse genes driven by conventional promoters (i.e. promoters that lack a repetitive element near their TSS), but cDUXC_interruptedHD was incapable of transcriptionally activating of these loci (Figure 22; compare two right-most bars). Although this is a small number of loci, the result is clear – the intactHD and interruptedHD isoforms differed in their transcriptional activity.

Discussion

The overarching question that this chapter addressed was: what are the consequences of intron loss via retroposition? Before this work it was known that some species have DUXC genes (with introns) and some species have DUXC-derived retrogenes (called mDUX in mice and hDUX4 in humans). One curious observation is that these genes and retrogenes have a reciprocal and non-overlapping distribution – no species has been found to have both a DUXC gene and a DUXC retrogene. This seems to imply that whatever the ancestral function of the DUXC gene was, the DUXC retrogene was an acceptable substitute (or it was a non-essential function). This curious distribution of genes and retrogenes does not, however, offer any insights into how retroposition may (or may not) have affected the ancestral function of the DUXC gene beyond its core functionality. That is to say, are the DUXC retrogenes completely redundant in function with DUXC genes or was a new function acquired during retroposition (i.e. neofunctionalization) or was the ancestral function partitioned between two or more factors (i.e. subfunctionalization).

While thinking about how retroposition could have impacted the function of the ancestral DUXC gene, it is interesting to speculate about the mechanism of retroposition of the DUXC genes. Typically, retroposition leads to the insertion of a processed mRNA at a quasi-random location (slight bias for AT-rich regions, otherwise random). If that were true for the retroposition event that created DUXC retrogenes, then in species that contain DUXC retrogenes and only DUXC retrogenes today (e.g. mice and humans), at one time DUXC genes and retrogenes existed in the same genome. If this happened, then the presence of only DUXC retrogenes today might imply that retroposition conferred a selective advantage to the retrogenes over the genes. One hypothesis as to what kind of advantage retroposition could have imparted is connected to the inherent loss of introns during retroposition. If the ancestral gene had introns and multiple isoforms, then retroposition could have impacted the isoforms available as only one isoform retroposed.

The work presented in this chapter established that in canine, at least, the DUXC gene makes an isoform of DUXC in adult testes where 60 amino acids are made that interrupt the first homeodomain. Furthermore, this interrupedHD isoform differs in function to some extent from the predicted "intactHD" isoform of DUXC. It seems likely that the intactHD isoform is expressed in a different tissue or developmental stage as the canine genome has retained the ability to encode the intactHD isoform and this ability likely would have degraded if were not being used in some tissue or developmental stage and thus preserved.

It is important to note that an alternative hypothesis to explain the reciprocal distribution of DUXC genes and retrogenes is that retroposition is unidirectional and thus, if the parental locus were somehow targeted for the insertion of the cDNA made from the processed mRNA, then the locus would have been overwritten. Interestingly, however, species that have DUXC genes also have multi-copy arrays of these genes, and the argument has been made that the ancestral DUXC gene was a multi-copy array. If so, then even retroposition into the parental locus would have only replaced one repeat unit initially such that there would have been both DUXC genes and retrogenes in the locus immediately following retroposition and thus there may still have been competition between repeat units as discussed above.

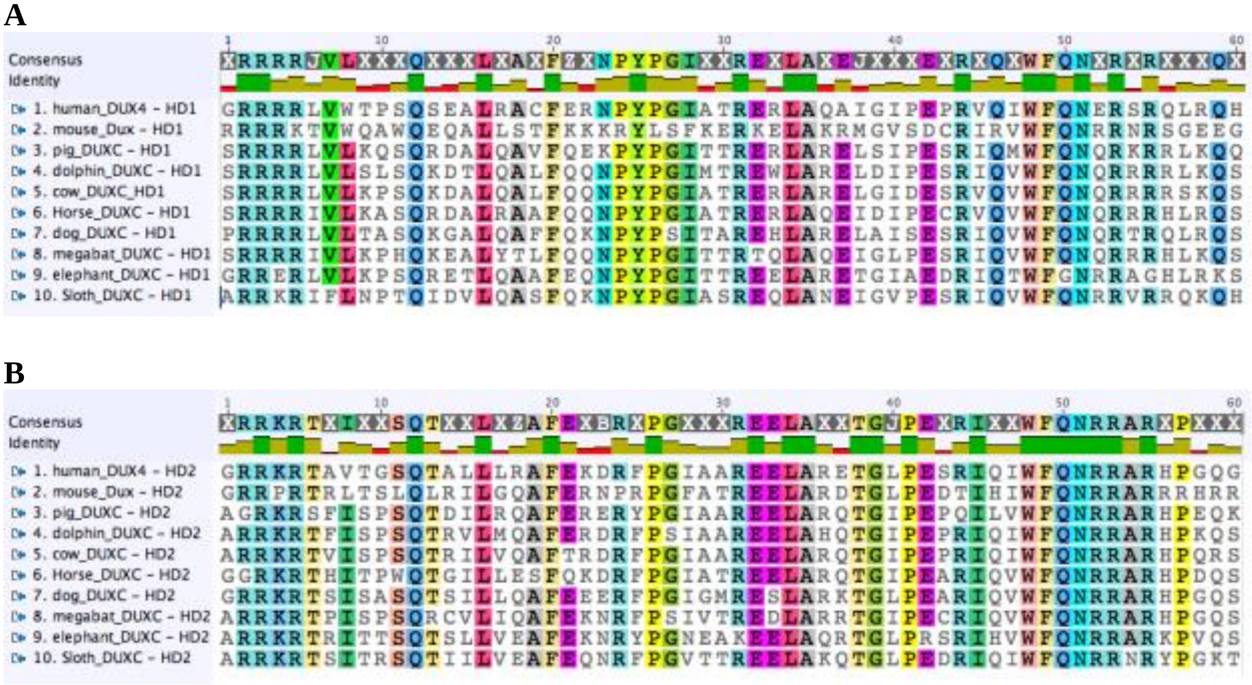
The work in this chapter lays a foundation for future studies. Some studies that will be critical are determining the tissue distribution of these two isoforms: intactHD and interruptedHD. The tissues where each isoform is created will shed light on their functions particularly whether their functions are entirely distinct (likely, if they are never expressed at the same time and place) or whether their functions overlap. Since heritable retroposition must occur in the germline or early embryo, this would be a good place to look for expression. Unfortunately, the canine reproduction is not easily manipulated and thus preimplantation embryos are difficult to acquire. Perhaps similar studies in cow would be more easily achieved. Studies of the binding motifs and transcriptomes of the intactHD and interruptedHD will also be informative and provide more robust comparisons to the functions of DUXC retrogenes.

Materials and Methods

Transient transfection and RT-qPCR

Transient DNA transfections of C2C12 cells were performed using SuperFect (QIAGEN) according to manufacturer specifications. Briefly, 80,000 cells were seeded per well of a 6-well plate the day prior to transfection, 2ug DNA/well and 10ul SuperFect/well. 24hrs post-transfection, total RNA was extracted from whole cells using NucleoSpin RNA kit (Macherey-Nagel) following the manufacturer's instructions. One microgram of total RNA was digested with DNAseI (Invitrogen) and then reverse transcribed into first strand cDNA in a 20 uL reaction using SuperScript III (Invitrogen) and oligo(dT) (Invitrogen). cDNA was diluted and used for RT-qPCR with iTaq Universal SYBR Green Supermix (Bio-Rad). Primer efficiency was determined by standard curve and all primer sets used were >90% efficient. Relative expression levels were normalized to the endogenous control locus Timm17b and empty vector by DeltaDeltaCT.

Figure 17 Homeodomain alignments of DUXC genes and retrogenes



A

```
                         1         10        20        30        40        50        60
Consensus                XRRRRJVLXXXQXXXLXAXFZXNPYPGIXXREXLAXEJXXXEXRXQXWFQNXRXRXXXQX
Identity
1. human_DUX4 - HD1      GRRRRLVWTPSQSEALRACFERNPYPGIATRERLAQAIGIPEPRVQIWFQNERSRQLRQH
2. mouse_Dux - HD1       RRRRKTVWQAWQEQALLSTFKKKRYLSFKERKELAKRMGVSDCRIRVWFQNRRNRSGEEG
3. pig_DUXC - HD1        SRRRRLVLKQSQRDALQAVFQEKPYPGITTRERLARELSIPESRIQMWFQNQRKRRLKQQ
4. dolphin_DUXC - HD1    SRRRRLVLSLSQKDTLQALFQQNPYPGIMTREWLARELDIPESRIQVWFQNQRRRRLKQS
5. cow_DUXC_HD1          SRRRRLVLKPSQKDALQALFQQNPYPGIATRERLARELGIDESRVQVWFQNQRRRRSKQS
6. Horse_DUXC - HD1      SRRRRIVLKASQRDALRAAFQQNPYPGIATRERLAQEIDIPECRVQVWFQNQRRRHLRQS
7. dog_DUXC - HD1        PRRRRLVLTASQKGALQAFFQKNPYPSITAREHLARELAISESRIQVWFQNQRTRQLRQS
8. megabat_DUXC - HD1    SRRRRIVLKPHQKEALYTLFQQNPYPGITTRTQLAQEIGLPESRIQVWFQNQRRRHLKQS
9. elephant_DUXC - HD1   GRRERLVLKPSQRETLQAAFEQNPYPGITTREELARETGIAEDRIQTWFGNRRAGHLRKS
10. Sloth_DUXC - HD1     ARRKRIFLNPTQIDVLQASFQKNPYPGIASREQLANEIGVPESRIQVWFQNRRVRRQKQH
```

B

```
                         1         10        20        30        40        50        60
Consensus                XRRKRTXIXXSQTXXLXZAFEXBRXPGXXXREELAXXTGJPEXRIXXWFQNRRARXPXXX
Identity
1. human_DUX4 - HD2      GRRKRTAVTGSQTALLLRAFEKDRFPGIAAREELARETGLPESRIQIWFQNRRARHPGQG
2. mouse_Dux - HD2       GRRPRTRLTSLQLRILGQAFERNPRPGFATREELARDTGLPEDTIHIWFQNRRARRRHRR
3. pig_DUXC - HD2        AGRKRSFISPSQTDILRQAFERERYPGIAAREELARQTGIPEPQILVWFQNRRARHPEQK
4. dolphin_DUXC - HD2    ARRKRTFISPSQTRVLMQAFERDRFPSIAAREELAHQTGIPEPRIQIWFQNRRARHPKQS
5. cow_DUXC - HD2        ARRKRTVISPSQTRILVQAFTRDRFPGIAAREELARQTGIPEPRIQIWFQNRRARHPQRS
6. Horse_DUXC - HD2      GGRKRTHITPWQTGILLESFQKDRFPGIATREELARQTGIPEARIQVWFQNRRARHPDQS
7. dog_DUXC - HD2        GRRKRTSISASQTSILLQAFEEERFPGIGMRESLARKTGLPEARIQVWFQNRRARHPGQS
8. megabat_DUXC - HD2    ARRKRTPISPSQRCVLIQAFEKNRFPSIVTREDLARRTGIPECRIQVWFQNRRARHPGQS
9. elephant_DUXC - HD2   ARRKRTRITTSQTSLLVEAFEKNRYPGNEAKEELAQRTGLPRSRIHVWFQNRRARKPVQS
10. Sloth_DUXC - HD2     ARRKRTSITRSQTIILVEAFEQNRFPGVTTREELAKQTGLPEDRIQIWFQNRRNRYPGKT
```

Figure 18 Complete ORFs of DUXC Genes and Retrogenes

```
> (HUMAN)DUX4_nucleotide_wildType
##NOTE: NCBI Reference Sequence: NC_000004.12
ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAGG
ACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCT
TTGAGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATC
GGCATTCCGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCT
GAGGCAGCACCGGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAA
GGCCGGCGAAAGCGGACCGCCGTCACCGGATCCCAGACCGCCCTGCTCCTCCGAGC
CTTTGAGAAGGATCGCTTTCCAGGCATCGCCGCCCGGGAGGAGCTGGCCAGAGAGA
CGGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCAC
CCGGGACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCC
CGGCGGGGGTCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGG
AACGGGGCTTCCCGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGG
CTTTCGTGAGCCAGGCAGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCG
CCGGCAGAGGGGATCTCCCAACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGC
CCCGGCTCCTCCGGACGGGGCGCTCTCCCACCCTCAGGCTCCTCGCTGGCCTCCGCA
CCCGGGCAAAAGCCGGGAGGACCGGGACCCGCAGCGCGACGGCCTGCCGGGCCCCT
GCGCGGTGGCACAGCCTGGGCCCGCTCAAGCGGGGCCGCAGGGCCAAGGGGTGCTT
GCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGGCTGGGGCCGGGGTCCCCAGGT
CGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCCACCTCCCCAGCCCGCGC
CCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGGCATCCCGGCGCCC
TCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGGCCTGCTGCTG
GATGAGCTCCTGGCGAGCCCGGAGTTTCTGCAGCAGGCGCAACCTCTCCTAGAAAC
GGAGGCCCCGGGGGAGCTGGAGGCCTCGGAAGAGGCCGCCTCGCTGGAAGCACCCC
TCAGCGAGGAAGAATACCGGGCTCTGCTGGAGGAGCTTTAG
```

```
> (HUMAN)DUX4_nucleotide_codonAltered
ATGGCATTGCCTACACCTTCAGACTCTACGCTGCCTGCAGAGGCTAGGGGAAGAGGT
AGACGGCGGCGATTGGTGTGGACTCCATCACAATCCGAAGCTCTTCGCGCATGCTTC
GAGCGCAATCCCTATCCGGGGATTGCCACAAGGGAGAGGCTTGCACAGGCTATCGG
AATCCCGGAACCGAGAGTGCAGATCTGGTTCCAAAATGAACGCTCTCGGCAGCTCA
GACAGCATCGCAGGGAGTCCCGCCCGTGGCCAGGAAGAAGGGGACCACCTGAAGG
AAGAAGAAAACGCACAGCGGTGACTGGCAGCCAAACGGCTCTGCTGCTCCGCGCTT
TCGAGAAAGATCGGTTCCCCGGAATTGCCGCACGCGAAGAACTCGCCAGAGAAACT
GGGCTCCCAGAATCACGAATACAGATTTGGTTCCAGAACCGCAGAGCAAGACACCC
AGGCCAGGGGGGACGGGCACCTGCTCAGGCCGGTGGACTCTGCTCTGCTGCCCCTG
GGGGCGGCCATCCAGCACCTTCCTGGGTGGCTTTCGCTCATACTGGCGCTTGGGGTA
CCGGGCTGCCTGCTCCGCATGTTCCCTGTGCTCCAGGGGCCCTCCCGCAGGGAGCGT
TTGTTTCCCAGGCAGCTAGGGCTGCACCTGCCCTGCAACCATCACAGGCAGCGCCAG
CTGAAGGCATCAGCCAACCCGCCCCAGCCCGCGGAGATTTTGCTTATGCAGCGCCA
GCACCTCCAGACGGTGCCCTGAGCCACCCCCAAGCCCCCAGATGGCCCCCTCACCCT
GGTAAGTCCCGGGAAGACCGCGATCCCCAACGAGATGGACTGCCCGGTCCTTGCGC
TGTGGCCCAGCCAGGACCTGCTCAAGCCGGCCCTCAGGGGCAAGGAGTGCTGGCCC
CACCTACAAGCCAGGGATCTCCCTGGTGGGGTTGGGGACGCGGACCTCAGGTTGCT
GGAGCCGCTTGGGAGCCTCAGGCCGGAGCTGCACCGCCGCCACAACCGGCCCCTCC
```

CGACGCGTCAGCGTCCGCCCGACAAGGCCAGATGCAGGGAATCCCAGCACCTAGCC
AAGCTCTTCAAGAGCCTGCCCCTTGGAGCGCACTGCCGTGTGGGCTGCTCCTGGATG
AACTCCTGGCTAGCCCAGAATTTCTCCAGCAGGCACAGCCACTCCTGGAAACAGAA
GCTCCGGGAGAGCTCGAAGCCTCCGAAGAAGCAGCAAGCCTGGAGGCACCTCTTTC
CGAGGAGGAGTATAGAGCCCTTCTGGAAGAACTTTGA

> (HUMAN)DUX4_aminoAcid
##NOTE: NCBI Reference Sequence: NC_000004.12
MALPTPSDSTLPAEARGRGRRRRLVWTPSQSEALRACFERNPYPGIATRERLAQAIGIPEP
RVQIWFQNERSRQLRQHRRESRPWPGRRGPPEGRRKRTAVTGSQTALLLRAFEKDRFPG
IAAREELARETGLPESRIQIWFQNRRARHPGQGGRAPAQAGGLCSAAPGGGHPAPSWVA
FAHTGAWGTGLPAPHVPCAPGALPQGAFVSQAARAAPALQPSQAAPAEGISQPAPARG
DFAYAAPAPPDGALSHPQAPRWPPHPGKSREDRDPQRDGLPGPCAVAQPGPAQAGPQG
QGVLAPPTSQGSPWWGWGRGPQVAGAAWEPQAGAAPPPQPAPPDASASARQGQMQGI
PAPSQALQEPAPWSALPCGLLLDELLASPEFLQQAQPLLETEAPGELEASEEAASLEAPLS
EEEYRALLEEL*

hDUX4 Homeodomain #1:
GRRRRLVWTPSQSEALRACFERNPYPGIATRERLAQAIGIPEPRVQIWFQNERSRQLRQH

hDUX4 Homeodomain #2:
GRRKRTAVTGSQTALLLRAFEKDRFPGIAAREELARETGLPESRIQIWFQNRRARHPGQG

hDUX4 Conserved C-terminal domain:
LLLDELLASPEFLQQAQPLLETEAPGELEASEEAASLEAPLSEEEYRALLEEL

> (MOUSE)DUX_nucleotide_wildType
##NOTE: NCBI Reference Sequence: NM_001081954.1
ATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCGGCG
GCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTCA
AGAAGAAGAGATACCTGAGCTTCAAGGAGAGGAAGGAGCTGGCCAAGCGAATGGG
GGTCTCAGATTGCCGCATCCGCGTGTGGTTTCAGAACCGCAGGAATCGCAGTGGAG
AGGAGGGGCATGCCTCAAAGAGGTCCATCAGAGGCTCCAGGCGGCTAGCCTCGCCA
CAGCTCCAGGAAGAGCTTGGATCCAGGCCACAGGGTAGAGGCATGCGCTCATCTGG
CAGAAGGCCTCGCACTCGACTCACCTCGCTACAGCTCAGGATCCTAGGGCAAGCCTT
TGAGAGGAACCCACGACCAGGCTTTGCTACCAGGGAGGAGCTGGCGCGTGACACAG
GGTTGCCCGAGGACACGATCCACATATGGTTTCAAAACCGAAGAGCTCGGCGGCGC
CACAGGAGGGGCAGGCCCACAGCTCAAGATCAAGACTTGCTGGCGTCACAAGGGTC
GGATGGGGCCCCTGCAGGTCCGGAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAAC
TTGTTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACCTCGAGCCCTAGCGA
CTTGCCCTCCTTCTGCGGAGAGTCCCAGCCTTTCCAAGTGGCACAGCCCCGTGGAGC
AGGCCAACAAGAGGCCCCCACTCGAGCAGGCAACGCAGGCTCTCTGGAACCCCTCC
TTGATCAGCTGCTGGATGAAGTCCAAGTAGAAGAGCCTGCTCCAGCCCCTCTGAATT
TGGATGGAGACCCTGGTGGCAGGGTGCATGAAGGTTCCCAGGAGAGCTTTTGGCCA
CAGGAAGAAGCAGGAAGTACAGGCATGGATACTTCTAGCCCCAGCGACTCAAACTC

```
CTTCTGCAGAGAGTCCCAGCCTTCCCAAGTGGCACAGCCCTGTGGAGCGGGCCAAG
AAGATGCCCGCACTCAAGCAGACAGCACAGGCCCTCTGGAACTCCTCCTCCTTGATC
AACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGTCCCACTGGATTGGGGT
AGAAATCCTGGCAGCAGGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGA
GGAAGCAGTAAATTCGGGCATGGATACCTCGATCCCTAGCATCTGGCCAACCTTCTG
CAGAGAATCCCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGG
CCCCCACTCAAGGTGGGAACACGGACCCCCTGGAGCTCTTCCTCTATCAACTGTTGG
ATGAAGTCCAAGTAGAAGAGCATGCTCCAGCCCCTCTGAATTGGGATGTAGATCCTG
GTGGCAGGGTGCATGAAGGTTCGTGGGAGAGCTTTTGGCCACAGGAAGAAGCAGGA
AGTACAGGCCTGGATACTTCAAGCCCCAGCGACTCAAACTCCTTCTTCAGAGAGTCC
AAGCCTTCCCAAGTGGCACAGCGCCGTGGAGCGGGCCAAGAAGATGCCCGCACTCA
AGCAGACAGCACAGGCCCTCTGGAACTCCTCCTCTTTGATCAACTGCTGGACGAAGT
CCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCA
TGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGCAGCAAATTCG
GGCAGGGATACCTCGATCCCTAGCATCTGGCCAGCCTTCTGCAGAAAATCCCAGCCT
CCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCCCCCATTCAAGGTGG
GAACACGGACCCCCTGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGA
GGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATGGACACA
ACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTCTGA

> (MOUSE)DUX_nucleotide_codonAltered
ATGGCTGAGGCTGGCTCTCCAGTGGGAGGATCTGGAGTGGCCAGAGAATCAAGGAG
AAGGAGGAAAACTGTCTGGCAAGCTTGGCAGGAACAGGCACTCCTGAGCACATTTA
AGAAAAAAAGGTATCTGTCCTTTAAAGAAAGAAAGGAACTGGCAAAAAGGATGGG
AGTTTCTGATTGCAGGATCAGAGTCTGGTTCCAGAATAGGAGAAATAGGTCTGGGG
AGGAAGGACATGCAAGCAAGAGAAGCATAAGAGGTTCCAGGAGGCTGGCATCCCCT
CAACTTCAGGAGGAACTGGGAAGTAGGCCCCAAGGCAGGGGCATGAGGTCCTCAGG
GAGGAGACCCAGAACCAGGCTGACAAGTCTGCAGCTGAGAATCCTTGGTCAGGCTT
TTGAAAGGAATCCAAGGCCAGGATTTGCCACCAGAGAGGAACTGGCCAGGGATACA
GGCCTTCCTGAGGATACTATCCATATCTGGTTCCAGAACAGGAGGGCCAGGAGAAG
GCACAGAAGGGGAAGACCTACAGCCCAGGACCAGGACCTCCTGGCTTCCCAGGGTT
CTGATGGAGCACCTGCTGGGCCTGAAGGTAGAGAGAGAGAAGGAGCACAGGAAAA
TTTGCTGCCCCAGGAGGAGGCAGGATCAACAGGGATGGACACCTCAAGCCCTTCTG
ACCTCCCTTCATTCTGTGGTGAATCACAGCCCTTTCAGGTGGCCCAGCCCAGGGGAG
CTGGACAGCAGGAGGCTCCCACAAGGGCAGGGAATGCTGGATCATTGGAGCCACTG
TTGGACCAGCTCTTGGATGAGGTCCAGGTGGAGGAACCTGCCCCAGCTCCACTCAAC
CTGGATGGTGATCCTGGGGGGAGGGTTCATGAGGGTAGTCAGGAGTCCTTCTGGCCC
CAGGAGGAGGCTGGTTCTACTGGAATGGACACTTCTTCACCCTCTGACAGCAATAGC
TTTTGCAGGGAGAGTCAACCCTCTCAGGTAGCTCAGCCTTGTGGGGCTGGCCAGGAG
GATGCTAGGACCCAGGCTGACTCAACAGGGCCCTTGGAGCTGTTGCTGCTGGACCA
GCTCCTGGATGAGGTACAGAAGGAGGAACATGTACCAGTGCCCCTGGACTGGGGGA
GGAACCCTGGAAGCAGAGAACATGAGGGTAGTCAGGATTCTCTCCTTCCTCTGGAA
GAGGCTGTGAATTCTGGAATGGACACTAGTATACCAAGTATTTGGCCTACATTTTGC
AGGGAGTCACAACCCCCACAGGTGGCTCAGCCTTCAGGACCTGGGCAGGCCCAGGC
TCCTACCCAAGGGGGTAATACAGACCCACTGGAACTCTTTCTGTATCAGCTGCTGGA
TGAGGTCCAGGTGGAGGAACATGCCCCAGCTCCACTCAACTGGGATGTGGATCCAG
```

GGGGCAGAGTCCATGAGGGTTCCTGGGAGTCATTCTGGCCCCAGGAGGAGGCAGGC
TCTACAGGACTGGACACAAGCTCCCCTAGTGACAGCAACTCATTCTTTAGGGAGAGT
AAGCCCTCTCAGGTTGCTCAAAGGAGGGGAGCTGGGCAAGAGGATGCCAGGACTCA
GGCTGACAGTACAGGACCCCTGGAGCTGCTGTTGTTTGACCAGCTCCTGGATGAAGT
GCAGAAGGAGGAACATGTTCCAGCTCCCCTGGACTGGGGAAGGAACCCTGGTTCTA
TGGAACATGAGGGCTCTCAGGACTCTCTCTTGCCTCTGGAAGAAGCTGCTAATAGTG
GCAGAGATACAAGTATCCCAAGCATTTGGCCTGCCTTTTGCAGGAAAAGCCAGCCA
CCCCAGGTAGCCCAGCCTAGTGGACCTGGACAGGCTCAGGCACCTATACAAGGAGG
CAACACTGACCCATTGGAGTTGTTTCTGGACCAGCTGCTCACTGAGGTGCAACTGGA
GGAACAAGGGCCAGCACCTGTCAATGTTGAAGAGACCTGGGAACAGATGGATACCA
CTCCAGACTTGCCACTGACTTCTGAAGAGTACCAGACCCTTCTTGACATGCTGTAA

> (MOUSE)DUX_aminoAcid
##NOTE: NCBI Reference Sequence: NM_001081954.1
MAEAGSPVGGSGVARESRRRRKTVWQAWQEQALLSTFKKKRYLSFKERKELAKRMGV
SDCRIRVWFQNRRNRSGEEGHASKRSIRGSRRLASPQLQEELGSRPQGRGMRSSGRRPR
TRLTSLQLRILGQAFERNPRPGFATREELARDTGLPEDTIHIWFQNRRARRRHRRGRPTA
QDQDLLASQGSDGAPAGPEGREREGAQENLLPQEEAGSTGMDTSSPSDLPSFCGESQPF
QVAQPRGAGQQEAPTRAGNAGSLEPLLDQLLDEVQVEEPAPAPLNLDGDPGGRVHEGS
QESFWPQEEAGSTGMDTSSPSDSNSFCRESQPSQVAQPCGAGQEDARTQADSTGPLELL
LLDQLLDEVQKEEHVPVPLDWGRNPGSREHEGSQDSLLPLEEAVNSGMDTSIPSIWPTFC
RESQPPQVAQPSGPGQAQAPTQGGNTDPLELFLYQLLDEVQVEEHAPAPLNWDVDPGG
RVHEGSWESFWPQEEAGSTGLDTSSPSDSNSFFRESKPSQVAQRRGAGQEDARTQADST
GPLELLLFDQLLDEVQKEEHVPAPLDWGRNPGSMEHEGSQDSLLPLEEAANSGRDTSIPS
IWPAFCRKSQPPQVAQPSGPGQAQAPIQGGNTDPLELFLDQLLTEVQLEEQGPAPVNVEE
TWEQMDTTPDLPLTSEEYQTLLDML*

mDUX Homeodomain #1:
RRRRKTVWQAWQEQALLSTFKKKRYLSFKERKELAKRMGVSDCRIRVWFQNRRNRSG
EEG

mDUX Homeodomain #2:
GRRPRTRLTSLQLRILGQAFERNPRPGFATREELARDTGLPEDTIHIWFQNRRARRRHRR

mDUX Conserved C-terminal domain:
LFLDQLLTEVQLEEQGPAPVNVEETWEQMDTTPDLPLTSEEYQTLLDML

> (CANFA)DUXC_nucleotide_wildType
##NOTE: CANFA indicates domesticated dog.
ATGGCCTCCAGCAGCACCCCCGGCGGCCCACTCCCTCGAGCACCCCGACGAAGGAG
GCTCGTGTTGACGGCAAGCCAGAAGGGGGCCCTGCAGGCATTCTTCCAGAAGAACC
CTTACCCCAGCATCACTGCCAGAGAACACCTGGCCCGAGAGCTGGCCATCTCCGAGT
CTAGAATCCAGGTCTGGTTCCAAAACCAGAGAACGAGACAGCTAAGGCAGAGCCGC
CGACTGGACTCCAGAATTCCCCAAGGAGAAGGGCCACCGAATGGAAAGGCACAGCC
TCCAGGTCGAGTCCCGAAGGAAGGCAGGAGAAAACGGACATCCATTTCTGCATCCC
AAACCAGTATCCTCCTTCAAGCCTTTGAGGAGGAGCGGTTTCCTGGCATTGGTATGA

GGGAAAGCCTGGCCAGAAAAACAGGCCTTCCAGAAGCCAGAATTCAGGTTTGGTTT
CAGAACAGAAGAGCTCGGCACCCAGGGCAGAGCCCAAGTGGGCCCGAGAATGCTTT
GGCGGCAAACCACAAACCCAGTCCTCGCGGGACGGTCCCATTGGACCAAAGCCACC
TGTCAAGGGTCCCCAGGAGCTCTCCAAATCTGGCTCCCTTCGATCCCTTGGGAAGCA
TGCAGACGCAGGCTGCAGGGACACCTCCTGTCTCCTCCGTGGTTGTTGTCCCTCCAG
TTTCTTGTGGGGGCTTTGGGCGCCTGATTCCGGGGGCCTGCCTGGTCACACCAACCT
TAGGTGGGCAAGGAGGAATCGCTGCTGCTCCCAGAGTCCTGGGGAGCCGATGCTGC
CCAGAACTGACTCCAGGAGGGGGCCTCTCACCAGGTCATGCTGACCTTGGCCTCCCC
TCCCCTGGGAGATGCCAGCAGCCGAAAGAGCACCCCAGCAAGGCGCCCCTGCCCTC
GCAAGTTGGCCCGCGGCCTCCGCCTGTTGATCCTCCTCAACACTGGGGTCATGCAGG
TCCCCCGGGCACCGGTCAGGCCACGCCGAGGAGGGGCCAAAGTTCCCAGGCAGTCA
TGGGCACAGCAGGGTCCCAGGATGGGACAGGGCAGCAGCCCGCCCCCGGGGAGAG
CCCCGCTTGGTGGCAACAGCCTCCCCCTCCTGCAGGGCCATGTGTCCCGCTGCCCCC
ACAACACCAGCTGTGTGCGGACACCTCCAGTTTCCTACAAGAGCTTTTCTCAGCCGA
TGAGATGGAAGAAGATGTCCACCCCTTGTGGGTGGGGACTCTGCAGGAGGACGAAC
CTCCAGGACCCCTGGAAGCACCCCTCAGCGAGGACGATTCTCACGCTCTGCTGGAA
ATGCTACAGGACTCCTTGTGGCCTCAGGCCTAG

> (CANFA)DUXC_aminoAcid
##NOTE: CANFA indicates domesticated dog.
MASSSTPGGPLPRAPRRRRLVLTASQKGALQAFFQKNPYPSITAREHLARELAISESRIQV
WFQNQRTRQLRQSRRLDSRIPQGEGPPNGKAQPPGRVPKEGRRKRTSISASQTSILLQAF
EEERFPGIGMRESLARKTGLPEARIQVWFQNRRARHPGQSPSGPENALAANHKPSPRGT
VPLDQSHLSRVPRSSPNLAPFDPLGSMQTQAAGTPPVSSVVVVPPVSCGGFGRLIPGACL
VTPTLGGQGGIAAAPRVLGSRCCPELTPGGGLSPGHADLGLPSPGRCQQPKEHPSKAPLP
SQVGPRPPPVDPPQHWGHAGPPGTGQATPRRGQSSQAVMGTAGSQDGTGQQPAPGESP
AWWQQPPPPAGPCVPLPPQHQLCADTSSFLQELFSADEMEEDVHPLWVGTLQEDEPPGP
LEAPLSEDDSHALLEMLQDSLWPQA*
(CANFA)DUXC Homeodomain #1:
PRRRRLVLTASQKGALQAFFQKNPYPSITAREHLARELAISESRIQVWFQNQRTRQLRQS

(CANFA)DUXC Homeodomain #2:
GRRKRTSISASQTSILLQAFEEERFPGIGMRESLARKTGLPEARIQVWFQNRRARHPGQS

(CANFA)DUXC Conserved C-terminal domain:
SFLQELFSADEMEEDVHPLWVGTLQEDEPPGPLEAPLSEDDSHALLEMLQDSLWPQA

>Horse_DUXC_nucleotides
ATGGCCTGTGCGGAGACGGTCCTGGGCGCTGTCAAGAGGCCCTGGCTGTCGTGCCC
GCAGACGGCGGCTGCCGCTCAGGGAAACCACCTGCAGACGAGGCGTCCTGGTGGCA
GCGGTGGAGGCGTGGCAGCTGGCCCGCATCAGAGAGGATCCCGACGCAGGAGGATT
GTTTTGAAGGCGAGTCAGAGGGACGCTCTGCGAGCAGCGTTTCAACAGAACCCTTA
CCCTGGGATCGCCACCAGAGAACGCCTGGCCCAAGAGATTGACATTCCGGAATGCA
GAGTCCAGGTTTGGTTTCAAAACCAACGCAGAAGACATCTAAGGCAGAGCCGGTCG
GGCTCGGCGAGCTCCGTGGGAGAAGGGCAATCGCCTGGAGAGGAGCAGCCCCAAG
CTCGGGCCGCAGAAGGCGGAAGAAAGCGGACACACATCACTCCGTGGCAAACCGG

GATCCTCCTTGAGAGCTTCCAGAAGGACCGATTTCCTGGCATCGCTACCAGGGAAGA
ACTGGCCAGACAAACGGGCATCCCAGAGGCGAGAATTCAGGTGTGGTTTCAGAACC
GAAGAGCTCGGCACCCAGACCAGAGTGGAAGCGGCCCGGTGAATGCCTTGGCGGAA
GGCCCCAGTCCCAGGGCTCCCCTGACTGCCCTCCAGGACCAAGCCAACCTGTCCTCT
GTCCCCAGCAGCTCTCCGCATCTGCCTCCCTGGAACCCTCCTGGGCTCTTGCCATCGC
CCGCGACAGCCGCTCCTCCACTCTGCCCGGTGTTCTTCGTTCCTTGGGTTCCCTCTGG
GGCCTGTGTGGGCCGGCCACCGGAGCCCCTGGTGGTCATGACAGCCCAGCCTGTGCT
GGGAAAGGAGAACGTTCACCCTCCTTGGACACTTCTGTGTCCCTGCTCAACCGGGCC
GCCTCTGGCAGGCGGTCTCTCAGCGATGCAGCCTCCTCTCCGGCCCACGCCCGGAGG
AAAATGCCAGGAGCACGACGGGCACGCTGGCGGGAGGGGGCTGCCCTTCCCACACT
CCCCTCAGCCTCACCCTGACCGTCCTCAGCAACAGTGGCAGCACCTGGGTGGGCCAG
GAGCCTTCCCCGCTATGCAGCCTTGGGGCGAGTGGCCTCAGGTCCTCCCGGCCCCAG
AGGAGCCTCAGGGAAGGGCGGTTCAGCAGTCTGCGCACCCTGACACACACGTGTGG
CCATGGGAGGAGCCATCAGCCGGAGAGCCCTCTGCTCAGCCGGGCCCACAGCAGCA
GCACTCTGCGCAAACCCCCAGCCTCCTAGATGAGCTGCTCGCAGTCACAGAGCTGCA
GGAAAAGGCACAGCCGTTCCTGAACGGGCATCCGCCGGCAGAGGAGCCTCCGGGAA
CACTGGAAGGTCCCCTCAGCGAGGAGGAATTTCAGGCTCTGCTCGACATGCTGCAA
AGCTCACCAGGGCCTCAGATTTAG

>Horse_DUXC_aminoAcids
MACAETVLGAVKRPWLSCPQTAAAAQGNHLQTRRPGGSGGGVAAGPHQRGSRRRRIV
LKASQRDALRAAFQQNPYPGIATRERLAQEIDIPECRVQVWFQNQRRRHLRQSRSGSAS
SVGEGQSPGEEQPQARAAEGGRKRTHITPWQTGILLESFQKDRFPGIATREELARQTGIPE
ARIQVWFQNRRARHPDQSGSGPVNALAEGPSPRAPLTALQDQANLSSVPSSSPHLPPWN
PPGLLPSPATAAPPLCPVFFVPWVPSGACVGRPPEPLVVMTAQPVLGKENVHPPWTLLC
PCSTGPPLAGGLSAMQPPLRPTPGGKCQEHDGHAGGRGLPFPHSPQPHPDRPQQQWQH
LGGPGAFPAMQPWGEWPQVLPAPEEPQGRAVQQSAHPDTHVWPWEEPSAGEPSAQPG
PQQQHSAQTPSLLDELLAVTELQEKAQPFLNGHPPAEEPPGTLEGPLSEEEFQALLDMLQ
SSPGPQI*

Horse_DUXC_Homeodomain1:
SRRRRIVLKASQRDALRAAFQQNPYPGIATRERLAQEIDIPECRVQVWFQNQRRRHLRQ
S

Horse_DUXC_Homeodomain2:
GGRKRTHITPWQTGILLESFQKDRFPGIATREELARQTGIPEARIQVWFQNRRARHPDQS

Horse_DUXC_Conserved C-terminal domain:
SLLDELLAVTELQEKAQPFLNGHPPAEEPPGTLEGPLSEEEFQALLDMLQSSPGPQI

> Pig_DUXC_nucleotides
ATGCCCCTCAAGTTGGCAGTGTTGGCTCTTTGCTTGGCCTCATGCCAGCAATCATTTT
TCCTAATGGGCTCACTTTCTAGAGGATCACGGAGAAGGAGGCTTGTTCTGAAACAGA
GTCAGCGGGATGCTCTGCAAGCAGTCTTTCAAGAGAAGCCCTACCCTGGTATAACGA
CCAGAGAACGACTGGCCAGAGAACTTAGCATCCCAGAAAGCCGAATTCAGATGTGG
TTCCAAAACCAAAGAAAACGACGTCTCAAGCAGCAGAGCAGAGGGCCACCTGAGA

CTATCCCCCAACCAGGGCCACCACAGCGGGAGCAACAGCTTCAGACTTCTCCCACTC
CTGCAATCCCAAAAGAGGCTGGGAGAAAGCGGTCATTCATCTCTCCCTCACAAACA
GACATCCTTCGGCAAGCCTTTGAGCGGGAACGATACCCAGGCATTGCCGCCAGGGA
AGAACTGGCACGTCAAACAGGGATTCCAGAACCTCAGATTCTGGTGTGGTTTCAGA
ACCGACGAGCTCGGCACCCAGAGCAGAAGGGAAGTGGGTCTGCCAATGTGCCCGGA
GTAGACCCCAATTCTGCAAAAGGCCTACCACTTCCATCGGACCAGGGCATGCCAAC
CACTGCCCACAGCAGCCCTACTCACAGTGCTCCTCCTCCTCCCTCTAACCCACCAAG
GGAGAACATGCTGTCCATCACCCCCATGGTGGCCACTGCTGCGATCGCCCCCAAATT
CATAGTTCCTGGGGCTCCCACAGCAGGCTGTGAGGGCCAGAGCCTGCCCATGATCTT
CATCATGGCCCAGCCAAGTCCAGTTCTGCAGGCAATAGTGAACCCTCCCATGCTTTG
GACGCTTCCTCTGACTCAGTCCTCACCAGGGCCAATGCCCATTCCTGCAGGGGGTCT
CACACCTATTCACACAGGGCTCTGGCCAACATCCCAAGAAGGACCATGGCAGGAGA
ACAATCTGCACACTATGCCAGCAGAAAAATGCCTCCCACACATCCCTCAGCCACCCC
TTGCCAGTCGTGCAGAGCCCCTGCCACTGCTGGACCCAGTGAAGACCTGCACTTATG
CCAGGCCAGAATGGGCCCAGGCATCCTCAGCTCAAGTCACCAGTGGGAAGCCTGTG
CATGGGGCCATGCTGCAGCCTGCACAGGCTGACACACTTATCTGCCCCTCTCATCTG
GCCCCCTCAAATGAAGAGCTGTGCCCTCCCATTGACCTGCAGCAGAACAAGCCCTCA
GCCTTCCAGGGCTCATCAAACCTCCTTGAGGAAATTATGGCAGCTGCAGGCATTCTG
CCTGAGGCAGGGCCTCTTCCAGACGTGGAGGAACAGGAAGAGCTTCCCCTAGGAGA
CCTGGAAGCACCCCTCAGTGAGGAAGATTTCCAGGCCCTCCTCGACATGCTGCCAAG
CTCCCCAGGTCCTTGTCCTTAG

> Pig_DUXC_aminoAcids
MPLKLAVLALCLASCQQSFFLMGSLSRGSRRRRLVLKQSQRDALQAVFQEKPYPGITTR
ERLARELSIPESRIQMWFQNQRKRRLKQQSRGPPETIPQPGPPQREQQLQTSPTPAIPKEA
GRKRSFISPSQTDILRQAFERERYPGIAAREELARQTGIPEPQILVWFQNRRARHPEQKGS
GSANVPGVDPNSAKGLPLPSDQGMPTTAHSSPTHSAPPPPSNPPRENMLSITPMVATAAI
APKFIVPGAPTAGCEGQSLPMIFIMAQPSPVLQAIVNPPMLWTLPLTQSSPGPMPIPAGGL
TPIHTGLWPTSQEGPWQENNLHTMPAEKCLPHIPQPPLASRAEPLPLLDPVKTCTYARPE
WAQASSAQVTSGKPVHGAMLQPAQADTLICPSHLAPSNEELCPPIDLQQNKPSAFQGSS
NLLEEIMAAAGILPEAGPLPDVEEQEELPLGDLEAPLSEEDFQALLDMLPSSPGPCP*

Pig_DUXC_Homeodomain1:
SRRRRLVLKQSQRDALQAVFQEKPYPGITTRERLARELSIPESRIQMWFQNQRKRRLKQ
Q

Pig_DUXC_Homeodomain2:
AGRKRSFISPSQTDILRQAFERERYPGIAAREELARQTGIPEPQILVWFQNRRARHPEQK

Pig_DUXC_Conserved C-terminal domain:
NLLEEIMAAAGILPEAGPLPDVEEQEELPLGDLEAPLSEEDFQALLDMLPSSPGPCP

>Elephant_DUXC_nucleotides
ATGGATCCGACCGGCGCTTCGAGTCGCTCTCAAAATCCACGAGGCCGACGAGAGAG
GTTGGTTTTGAAGCCCAGTCAAAGAGAGACCCTGCAAGCAGCGTTTGAACAGAACC
CCTACCCTGGTATAACTACCAGAGAAGAACTCGCCAGAGAAACCGGCATCGCGGAG

GATCGCATTCAGACTTGGTTTGGAAACCGCAGAGCAGGTCACCTAAGGAAGAGCCG
CTCGGCCTCTGGACAGGCCTCCGAAGAAGAGCCGTCCCAGGGACAGGGAGAGCCTC
AGCCTTGGTCTCCGGAAAATTTCCCCAAAGCGGCCAGACGAAAACGCACACGCATC
ACCACATCGCAAACGAGTCTCCTAGTCGAGGCCTTCGAGAAGAACCGGTACCCTGG
TAACGAGGCCAAGGAAGAACTGGCTCAACGAACTGGCCTTCCGCGATCCCGAATTC
ACGTATGGTTTCAGAACCGAAGAGCTCGGAAGCCGGTGCAGAGCGCGAGTGCACCG
CCGAAGTCCTTGGCAGACAGCCCGACTCCTGCGGCCACGCTTCCACTCGACCAAAGC
GACCTGTCCTCTGTACAGAGCACCTACCCTCTCGGCCCACCCTCCCATCCTTCTAGCA
GCAACCAAGCCATCCTACCTGTTCTCACTGAGTCCCGTACACCATTTCTTCCTTCGGA
ACCCACCCAGGGCTGTGCCGGCCAAGCACCGGGTGCCGTGTTGGACCAGCCCGCCC
TGATTGTGAAGAAGACAGCAGAGACCTCTCACGCGCCGGGGACACACCTGAACCAA
TCGCCAACAGGACCCACTGTGGGAGACAGGCTGTCAGACCCTCAGGCTCCTTTCTGG
CCCCAATACCCAGGAAATTACCAGGATCGCGACCAACATGCTGTCTCGGCAGGGTG
GCTCGCCCAAGACCCTTCTCGGCCTGACAATTCAAAGACGCAAGGGCAGGTTCCGG
CTCAGCAAGTCACAGCTCCCTTCACGCAATGGGGCTGTGAGGTGGCCCAGGGTGTG
ACCGCCCGATGGGAACCCAGCCAAGAGACACTCCAGCAGCCCGGACACTCCGAGGC
ACACCTGTGGCCAGAGCCGGCACAATCGGCTCAAGAGTCATCTCATCCACCAGACC
AAGACTGCCAGGAAACCGAGAGCCTTTTAGATGAACTCCTCTCCGCCCCAGAGTTGC
AGGGAAAGTCCCAAACCTTTCTGAACGCGGATCCACAGGAGGAGGACCCTCCACAA
CTCGAACTCTCCCTCGGCGACATTGACTTTCAGGCTCTGCTTGACGCGCTGCAAGAT
TGA

>Elephant_DUXC_aminoAcids
MDPTGASSRSQNPRGRRERLVLKPSQRETLQAAFEQNPYPGITTREELARETGIAEDRIQ
TWFGNRRAGHLRKSRSASGQASEEEPSQGQGEPQPWSPENFPKAARRKRTRITTSQTSL
LVEAFEKNRYPGNEAKEELAQRTGLPRSRIHVWFQNRRARKPVQSASAPPKSLADSPTP
AATLPLDQSDLSSVQSTYPLGPPSHPSSSNQAILPVLTESRTPFLPSEPTQGCAGQAPGAV
LDQPALIVKKTAETSHAPGTHLNQSPTGPTVGDRLSDPQAPFWPQYPGNYQDRDQHAV
SAGWLAQDPSRPDNSKTQGQVPAQQVTAPFTQWGCEVAQGVTARWEPSQETLQQPGH
SEAHLWPEPAQSAQESSHPPDQDCQETESLLDELLSAPELQGKSQTFLNADPQEEDPPQL
ELSLGDIDFQALLDALQD*

Elephant_DUXC_Homeodomain1:
GRRERLVLKPSQRETLQAAFEQNPYPGITTREELARETGIAEDRIQTWFGNRRAGHLRKS

Elephant _DUXC_Homeodomain2:
ARRKRTRITTSQTSLLVEAFEKNRYPGNEAKEELAQRTGLPRSRIHVWFQNRRARKPVQ
S

Elephant _DUXC_Conserved C-terminal domain:
SLLDELLSAPELQGKSQTFLNADPQEEDPPQLELSLGDIDFQALLDALQD

>Sloth_DUXC_nucleotides
ATGCGGATGACCCGAATCGCCATCTCCCTGGTGTCCGCTGATGACAGCCTTCCAAGT
ACCCTGAAAGGAGTGGCCCGAAGAAAGAGGATCTTTTTGAACCCAACTCAAATTGA
TGTCCTGCAAGCATCGTTTCAAAAGAACCCCTACCCTGGTATAGCTTCCAGGGAACA

ACTGGCTAATGAAATTGGTGTTCCAGAGTCTCGAATTCAGGTTTGGTTTCAGAACCG
GAGAGTAAGACGCCAAAAGCAGCATCAACCGCAGTCTGGATCCTGCTCAGAAGATT
GTTTACCCAAAGAAGCCCGTCGTAAGCGCACATCCATCACCAGATCCCAAACCATC
ATTCTGGTTGAGGCCTTTGAGCAGAACCGATTCCCTGGTGTTACAACCAGAGAAGAA
CTTGCTAAACAAACAGGCCTTCCAGAAGATAGAATTCAGATATGGTTTCAGAATCGG
AGAAATCGGTACCCAGGGAAGACACCAAGCGGACACAGAAATTCCGCGGCAGGTG
CCCCAAATCGGAGGCCTCATCTGACCATTGGGCAGGAGAAAACTCACCTGATCACT
GTCCCAAGAAGGCCCCATCATCTTGCTTCCTGCAATATTTTCCACGAGACATGCATA
ATTCCCTCCACTATTCTTTTGTGCCTCACAACCTCTGCTCTTAAGGATTCAAATGTGA
ACTGCATGAGTCAGGCACCCCATTTCCTGGAGGCCCAGCCCACACTGACTGCACAG
GCAGGGGCAAACGCTTACCCCACACAGACTATTATCAGTCACTGCCCAGCAGAGCA
ACCTCTGGGAATGGGGTTCTCAGATAAGCCAAATAATTTCAAGCTCCCTTTCCAGGG
AAAATGCCAGGATCAAGATGAATCCACTGGAAGGGGAGTGGTGCAGTTGAAAGACA
ATCCCCTGACACAAACTGACAATGAAAAACAACAATTACATGATGTTGGTCGGGCA
GACACATCTCACAACATGCAGTGGTGCAGCGAGGAGTTGCAAAGTGTGAATGCAGA
AGGAGAAACTCCTGAAGGGAAACTTCATCAGCCTAGACACTCTGAGATGCAGCCAG
GGCAGCAGCAGGCAGAATCAGCTGAAGAGCCATCACTTCCCCCTGCCCAGGAGCAC
CAGCAAGATCTGGAGTCCTGGAGCCTTCTGGACCAACTGCTGTCGAGCAAAGAATTT
CTGGAAAAGGCCCAACCTCTTCTCAATCCAGATTCCCAGGACCAGAATTCTCTACCA
GTTGAACCATCCCTCAGTGAGGAAGAGTTTCAGGCTCTGCTTGACATGCTGTGA

>Sloth_DUXC_aminoAcids
MRMTRIAISLVSADDSLPSTLKGVARRKRIFLNPTQIDVLQASFQKNPYPGIASREQLANE
IGVPESRIQVWFQNRRVRRQKQHQPQSGSCSEDCLPKEARRKRTSITRSQTIILVEAFEQN
RFPGVTTREELAKQTGLPEDRIQIWFQNRRNRYPGKTPSGHRNSAAGAPNRRPHLTIGQE
KTHLITVPRRPHHLASCNIFHETCIIPSTILLCLTTSALKDSNVNCMSQAPHFLEAQPTLTA
QAGANAYPTQTIISHCPAEQPLGMGFSDKPNNFKLPFQGKCQDQDESTGRGVVQLKDNP
LTQTDNEKQQLHDVGRADTSHNMQWCSEELQSVNAEGETPEGKLHQPRHSEMQPGQQ
QAESAEEPSLPPAQEHQQDLESWSLLDQLLSSKEFLEKAQPLLNPDSQDQNSLPVEPSLS
EEEFQALLDML*

Sloth_DUXC_Homeodomain1:
ARRKRIFLNPTQIDVLQASFQKNPYPGIASREQLANEIGVPESRIQVWFQNRRVRRQKQH

Sloth _DUXC_Homeodomain2:
ARRKRTSITRSQTIILVEAFEQNRFPGVTTREELAKQTGLPEDRIQIWFQNRRNRYPGKT

Sloth _DUXC_Conserved C-terminal domain:
SLLDQLLSSKEFLEKAQPLLNPDSQDQNSLPVEPSLSEEEFQALLDML

Chimeras:
> MMH_nucleotide
##NOTE: MMH is mDUX homeodomains and the hDUX4 carboxy-terminus
ATGGCTGAGGCTGGCTCTCCAGTGGGAGGATCTGGAGTGGCCAGAGAATCAAGGAG
AAGGAGGAAAACTGTCTGGCAAGCTTGGCAGGAACAGGCACTCCTGAGCACATTTA
AGAAAAAAAGGTATCTGTCCTTTAAAGAAAGAAAGGAACTGGCAAAAAGGATGGG

```
AGTTTCTGATTGCAGGATCAGAGTCTGGTTCCAGAATAGGAGAAATAGGTCTGGGG
AGGAAGGACATGCAAGCAAGAGAAGCATAAGAGGTTCCAGGAGGCTGGCATCCCCT
CAACTTCAGGAGGAACTGGGAAGTAGGCCCCAAGGCAGGGGCATGAGGTCCTCAGG
GAGGAGACCCAGAACCAGGCTGACAAGTCTGCAGCTGAGAATCCTTGGTCAGGCTT
TTGAAAGGAATCCAAGGCCAGGATTTGCCACCAGAGAGGAACTGGCCAGGGATACA
GGCCTTCCTGAGGATACTATCCATATCTGGTTCCAGAACAGGAGGGCCAGGAGAAG
GCACAGAAGGGGAAGACCTCCTGCTCAGGCCGGTGGACTCTGCTCTGCTGCCCCTG
GGGGCGGCCATCCAGCACCTTCCTGGGTGGCTTTCGCTCATACTGGCGCTTGGGGTA
CCGGGCTGCCTGCTCCGCATGTTCCCTGTGCTCCAGGGGCCCTCCCGCAGGGAGCGT
TTGTTTCCCAGGCAGCTAGGGCTGCACCTGCCCTGCAACCATCACAGGCAGCGCCAG
CTGAAGGCATCAGCCAACCCGCCCCAGCCCGCGGAGATTTTGCTTATGCAGCGCCA
GCACCTCCAGACGGTGCCCTGAGCCACCCCCAAGCCCCCAGATGGCCCCCTCACCCT
GGTAAGTCCCGGGAAGACCGCGATCCCCAACGAGATGGACTGCCCGGTCCTTGCGC
TGTGGCCCAGCCAGGACCTGCTCAAGCCGGCCCTCAGGGGCAAGGAGTGCTGGCCC
CACCTACAAGCCAGGGATCTCCCTGGTGGGGTTGGGGACGCGGACCTCAGGTTGCT
GGAGCCGCTTGGGAGCCTCAGGCCGGAGCTGCACCGCCGCCACAACCGGCCCCTCC
CGACGCGTCAGCGTCCGCCCGACAAGGCCAGATGCAGGGAATCCCAGCACCTAGCC
AAGCTCTTCAAGAGCCTGCCCCTTGGAGCGCACTGCCGTGTGGGCTGCTCCTGGATG
AACTCCTGGCTAGCCCAGAATTTCTCCAGCAGGCACAGCCACTCCTGGAAACAGAA
GCTCCGGGAGAGCTCGAAGCCTCCGAAGAAGCAGCAAGCCTGGAGGCACCTCTTTC
CGAGGAGGAGTATAGAGCCCTTCTGGAAGAACTTTGA

> MMH_aminoAcid
##NOTE: MMH is mDUX homeodomains and the hDUX4 carboxy-terminus
MAEAGSPVGGSGVARESRRRRKTVWQAWQEQALLSTFKKKRYLSFKERKELAKRMGV
SDCRIRVWFQNRRNRSGEEGHASKRSIRGSRRLASPQLQEELGSRPQGRGMRSSGRRPR
TRLTSLQLRILGQAFERNPRPGFATREELARDTGLPEDTIHIWFQNRRARRRHRRGRPPA
QAGGLCSAAPGGGHPAPSWVAFAHTGAWGTGLPAPHVPCAPGALPQGAFVSQAARAA
PALQPSQAAPAEGISQPAPARGDFAYAAPAPPDGALSHPQAPRWPPHPGKSREDRDPQR
DGLPGPCAVAQPGPAQAGPQGQGVLAPPTSQGSPWWGWGRGPQVAGAAWEPQAGAA
PPPQPAPPDASASARQGQMQGIPAPSQALQEPAPWSALPCGLLLDELLASPEFLQQAQPL
LETEAPGELEASEEAASLEAPLSEEEYRALLEEL*

> MHM_nucleotide
##NOTE: MHM is the second hDUX4 homeodomain introduced into mDUX in place of the
mDUX second homeodomain
ATGGCTGAGGCTGGCTCTCCAGTGGGAGGATCTGGAGTGGCCAGAGAATCAAGGAG
AAGGAGGAAAACTGTCTGGCAAGCTTGGCAGGAACAGGCACTCCTGAGCACATTTA
AGAAAAAAAGGTATCTGTCCTTTAAAGAAAGAAAGGAACTGGCAAAAAGGATGGG
AGTTTCTGATTGCAGGATCAGAGTCTGGTTCCAGAATAGGAGAAATAGGTCTGGGG
AGGAAGGACATGCAAGCAAGAGAAGCATAAGAGGTTCCAGGAGGCTGGCATCCCCT
CAACTTCAGGAGGAACTGGGAAGTAGGCCCCAAGGCAGGGGCATGAGGTCCTCAGG
AAGAAGAAAACGCACAGCGGTGACTGGCAGCCAAACGGCTCTGCTGCTCCGCGCTT
TCGAGAAAGATCGGTTCCCCGGAATTGCCGCACGCGAAGAACTCGCCAGAGAAACT
GGGCTCCCAGAATCACGAATACAGATTTGGTTCCAGAACCGCAGAGCAAGACACCC
AGGCCAGGGGGGAAGACCTACAGCCCAGGACCAGGACCTCCTGGCTTCCCAGGGTT
```

CTGATGGAGCACCTGCTGGGCCTGAAGGTAGAGAGAGAGAAGGAGCACAGGAAAA
TTTGCTGCCCCAGGAGGAGGCAGGATCAACAGGGATGGACACCTCAAGCCCTTCTG
ACCTCCCTTCATTCTGTGGTGAATCACAGCCCTTTCAGGTGGCCCAGCCCAGGGGAG
CTGGACAGCAGGAGGCTCCCACAAGGGCAGGGAATGCTGGATCATTGGAGCCACTG
TTGGACCAGCTCTTGGATGAGGTCCAGGTGGAGGAACCTGCCCCAGCTCCACTCAAC
CTGGATGGTGATCCTGGGGGGAGGGTTCATGAGGGTAGTCAGGAGTCCTTCTGGCCC
CAGGAGGAGGCTGGTTCTACTGGAATGGACACTTCTTCACCCTCTGACAGCAATAGC
TTTTGCAGGGAGAGTCAACCCTCTCAGGTAGCTCAGCCTTGTGGGGCTGGCCAGGAG
GATGCTAGGACCCAGGCTGACTCAACAGGGCCCTTGGAGCTGTTGCTGCTGGACCA
GCTCCTGGATGAGGTACAGAAGGAGGAACATGTACCAGTGCCCCTGGACTGGGGGA
GGAACCCTGGAAGCAGAGAACATGAGGGTAGTCAGGATTCTCTCCTTCCTCTGGAA
GAGGCTGTGAATTCTGGAATGGACACTAGTATACCAAGTATTTGGCCTACATTTTGC
AGGGAGTCACAACCCCCACAGGTGGCTCAGCCTTCAGGACCTGGGCAGGCCCAGGC
TCCTACCCAAGGGGGTAATACAGACCCACTGGAACTCTTTCTGTATCAGCTGCTGGA
TGAGGTCCAGGTGGAGGAACATGCCCCAGCTCCACTCAACTGGGATGTGGATCCAG
GGGGCAGAGTCCATGAGGGTTCCTGGGAGTCATTCTGGCCCCAGGAGGAGGCAGGC
TCTACAGGACTGGACACAAGCTCCCCTAGTGACAGCAACTCATTCTTTAGGGAGAGT
AAGCCCTCTCAGGTTGCTCAAAGGAGGGGAGCTGGGCAAGAGGATGCCAGGACTCA
GGCTGACAGTACAGGACCCCTGGAGCTGCTGTTGTTTGACCAGCTCCTGGATGAAGT
GCAGAAGGAGGAACATGTTCCAGCTCCCCTGGACTGGGGAAGGAACCCTGGTTCTA
TGGAACATGAGGGCTCTCAGGACTCTCTCTTGCCTCTGGAAGAAGCTGCTAATAGTG
GCAGAGATACAAGTATCCCAAGCATTTGGCCTGCCTTTTGCAGGAAAAGCCAGCCA
CCCCAGGTAGCCCAGCCTAGTGGACCTGGACAGGCTCAGGCACCTATACAAGGAGG
CAACACTGACCCATTGGAGTTGTTTCTGGACCAGCTGCTCACTGAGGTGCAACTGGA
GGAACAAGGGCCAGCACCTGTCAATGTTGAAGAGACCTGGGAACAGATGGATACCA
CTCCAGACTTGCCACTGACTTCTGAAGAGTACCAGACCCTTCTTGACATGCTGTAA

> MHM_aminoAcid
##NOTE: MHM is the second hDUX4 homeodomain introduced into mDUX in place of the mDUX second homeodomain
MAEAGSPVGGSGVARESRRRRKTVWQAWQEQALLSTFKKKRYLSFKERKELAKRMGV
SDCRIRVWFQNRRNRSGEEGHASKRSIRGSRRLASPQLQEELGSRPQGRGMRSSGRRKR
TAVTGSQTALLLRAFEKDRFPGIAAREELARETGLPESRIQIWFQNRRARHPGQGGRPTA
QDQDLLASQGSDGAPAGPEGREREGAQENLLPQEEAGSTGMDTSSPSDLPSFCGESQPF
QVAQPRGAGQQEAPTRAGNAGSLEPLLDQLLDEVQVEEPAPAPLNLDGDPGGRVHEGS
QESFWPQEEAGSTGMDTSSPSDSNSFCRESQPSQVAQPCGAGQEDARTQADSTGPLELL
LLDQLLDEVQKEEHVPVPLDWGRNPGSREHEGSQDSLLPLEEAVNSGMDTSIPSIWPTFC
RESQPPQVAQPSGPGQAQAPTQGGNTDPLELFLYQLLDEVQVEEHAPAPLNWDVDPGG
RVHEGSWESFWPQEEAGSTGLDTSSPSDSNSFFRESKPSQVAQRRGAGQEDARTQADST
GPLELLLFDQLLDEVQKEEHVPAPLDWGRNPGSMEHEGSQDSLLPLEEAANSGRDTSIPS
IWPAFCRKSQPPQVAQPSGPGQAQAPIQGGNTDPLELFLDQLLTEVQLEEQGPAPVNVEE
TWEQMDTTPDLPLTSEEYQTLLDML*

> HMM_nucleotide
##NOTE: HMM is the first hDUX4 homeodomain introduced into mDUX in place of the mDUX first homeodomain

```
ATGGCTGAGGCTGGCTCTCCAGTGGGAGGATCTGGAGTGGCCAGAGAATCAGGTAG
ACGGCGGCGATTGGTGTGGACTCCATCACAATCCGAAGCTCTTCGCGCATGCTTCGA
GCGCAATCCCTATCCGGGGATTGCCACAAGGGAGAGGCTTGCACAGGCTATCGGAA
TCCCGGAACCGAGAGTGCAGATCTGGTTCCAAAATGAACGCTCTCGGCAGCTCAGA
CAGCATCATGCAAGCAAGAGAAGCATAAGAGGTTCCAGGAGGCTGGCATCCCCTCA
ACTTCAGGAGGAACTGGGAAGTAGGCCCCAAGGCAGGGGCATGAGGTCCTCAGGGA
GGAGACCCAGAACCAGGCTGACAAGTCTGCAGCTGAGAATCCTTGGTCAGGCTTTT
GAAAGGAATCCAAGGCCAGGATTTGCCACCAGAGAGGAACTGGCCAGGGATACAG
GCCTTCCTGAGGATACTATCCATATCTGGTTCCAGAACAGGAGGGCCAGGAGAAGG
CACAGAAGGGGAAGACCTACAGCCCAGGACCAGGACCTCCTGGCTTCCCAGGGTTC
TGATGGAGCACCTGCTGGGCCTGAAGGTAGAGAGAGAGAAGGAGCACAGGAAAAT
TTGCTGCCCCAGGAGGAGGCAGGATCAACAGGGATGGACACCTCAAGCCCTTCTGA
CCTCCCTTCATTCTGTGGTGAATCACAGCCCTTTCAGGTGGCCCAGCCCAGGGGAGC
TGGACAGCAGGAGGCTCCCACAAGGGCAGGGAATGCTGGATCATTGGAGCCACTGT
TGGACCAGCTCTTGGATGAGGTCCAGGTGGAGGAACCTGCCCCAGCTCCACTCAAC
CTGGATGGTGATCCTGGGGGGAGGGTTCATGAGGGTAGTCAGGAGTCCTTCTGGCCC
CAGGAGGAGGCTGGTTCTACTGGAATGGACACTTCTTCACCCTCTGACAGCAATAGC
TTTTGCAGGGAGAGTCAACCCTCTCAGGTAGCTCAGCCTTGTGGGGCTGGCCAGGAG
GATGCTAGGACCCAGGCTGACTCAACAGGGCCCTTGGAGCTGTTGCTGCTGGACCA
GCTCCTGGATGAGGTACAGAAGGAGGAACATGTACCAGTGCCCCTGGACTGGGGGA
GGAACCCTGGAAGCAGAGAACATGAGGGTAGTCAGGATTCTCTCCTTCCTCTGGAA
GAGGCTGTGAATTCTGGAATGGACACTAGTATACCAAGTATTTGGCCTACATTTTGC
AGGGAGTCACAACCCCCACAGGTGGCTCAGCCTTCAGGACCTGGGCAGGCCCAGGC
TCCTACCCAAGGGGGTAATACAGACCCACTGGAACTCTTTCTGTATCAGCTGCTGGA
TGAGGTCCAGGTGGAGGAACATGCCCCAGCTCCACTCAACTGGGATGTGGATCCAG
GGGGCAGAGTCCATGAGGGTTCCTGGGAGTCATTCTGGCCCCAGGAGGAGGCAGGC
TCTACAGGACTGGACACAAGCTCCCCTAGTGACAGCAACTCATTCTTTAGGGAGAGT
AAGCCCTCTCAGGTTGCTCAAAGGAGGGGAGCTGGGCAAGAGGATGCCAGGACTCA
GGCTGACAGTACAGGACCCCTGGAGCTGCTGTTGTTTGACCAGCTCCTGGATGAAGT
GCAGAAGGAGGAACATGTTCCAGCTCCCCTGGACTGGGGAAGGAACCCTGGTTCTA
TGGAACATGAGGGCTCTCAGGACTCTCTCTTGCCTCTGGAAGAAGCTGCTAATAGTG
GCAGAGATACAAGTATCCCAAGCATTTGGCCTGCCTTTTGCAGGAAAAGCCAGCCA
CCCCAGGTAGCCCAGCCTAGTGGACCTGGACAGGCTCAGGCACCTATACAAGGAGG
CAACACTGACCCATTGGAGTTGTTTCTGGACCAGCTGCTCACTGAGGTGCAACTGGA
GGAACAAGGGCCAGCACCTGTCAATGTTGAAGAGACCTGGGAACAGATGGATACCA
CTCCAGACTTGCCACTGACTTCTGAAGAGTACCAGACCCTTCTTGACATGCTGTAA
```

> HMM_aminoAcid

##NOTE: HMM is the first hDUX4 homeodomain introduced into mDUX in place of the mDUX first homeodomain

```
MAEAGSPVGGSGVARESGRRRRLVWTPSQSEALRACFERNPYPGIATRERLAQAIGIPEP
RVQIWFQNERSRQLRQHHASKRSIRGSRRLASPQLQEELGSRPQGRGMRSSGRRPRTRL
TSLQLRILGQAFERNPRPGFATREELARDTGLPEDTIHIWFQNRRARRRHRRGRPTAQDQ
DLLASQGSDGAPAGPEGREREGAQENLLPQEEAGSTGMDTSSPSDLPSFCGESQPFQVA
QPRGAGQQEAPTRAGNAGSLEPLLDQLLDEVQVEEPAPAPLNLDGDPGGRVHEGSQESF
WPQEEAGSTGMDTSSPSDSNSFCRESQPSQVAQPCGAGQEDARTQADSTGPLELLLLDQ
```

LLDEVQKEEHVPVPLDWGRNPGSREHEGSQDSLLPLEEAVNSGMDTSIPSIWPTFCRESQ
PPQVAQPSGPGQAQAPTQGGNTDPLELFLYQLLDEVQVEEHAPAPLNWDVDPGGRVHE
GSWESFWPQEEAGSTGLDTSSPSDSNSFFRESKPSQVAQRRGAGQEDARTQADSTGPLE
LLLFDQLLDEVQKEEHVPAPLDWGRNPGSMEHEGSQDSLLPLEEAANSGRDTSIPSIWP
AFCRKSQPPQVAQPSGPGQAQAPIQGGNTDPLELFLDQLLTEVQLEEQGPAPVNVEETW
EQMDTTPDLPLTSEEYQTLLDML*

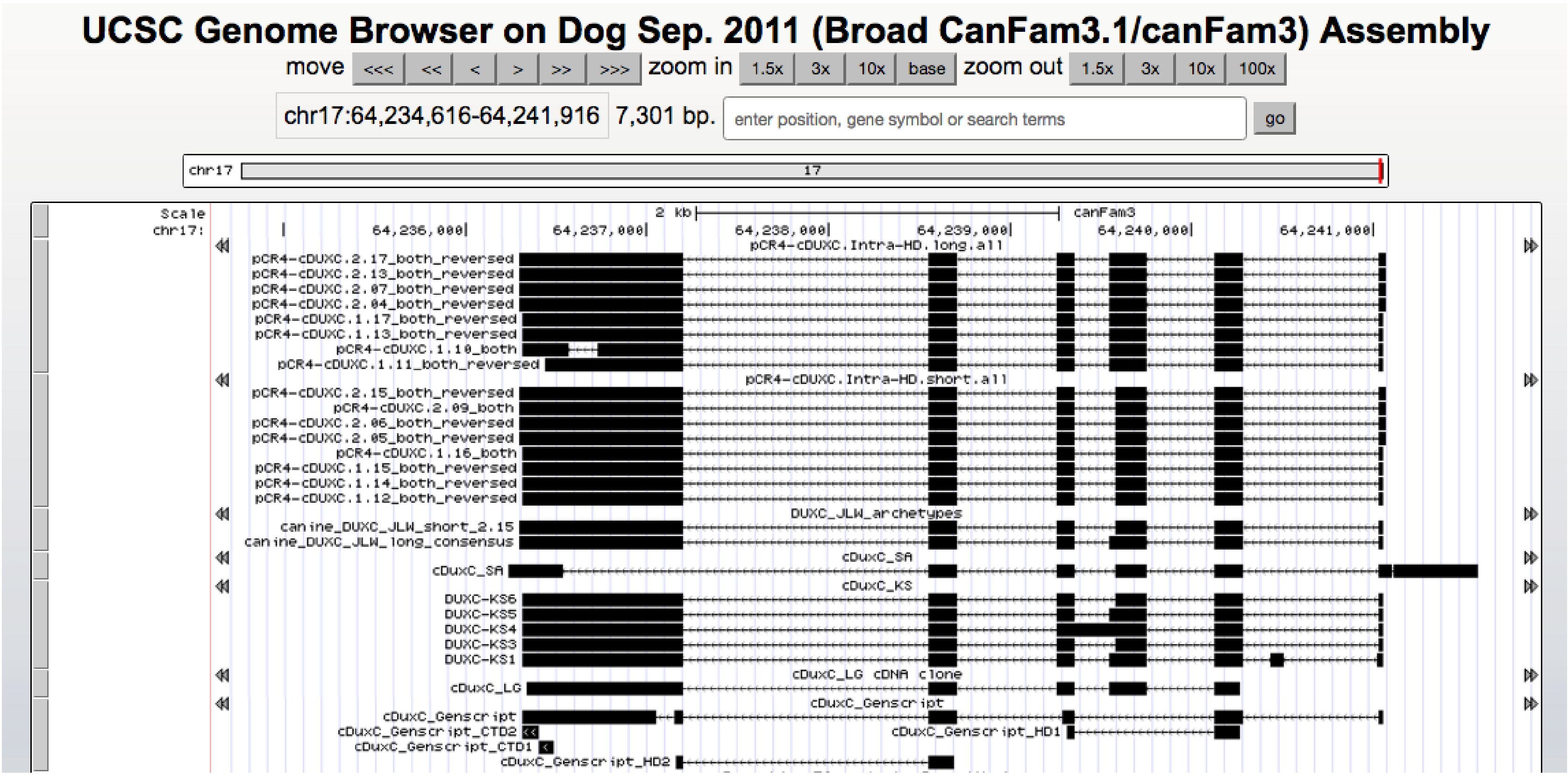
> HMH_nucleotide
##NOTE: HMH is the second mDUX homeodomain introduced into hDUX4 in place of the hDUX4 second homeodomain
ATGGCATTGCCTACACCTTCAGACTCTACGCTGCCTGCAGAGGCTAGGGGAAGAGGT
AGACGGCGGCGATTGGTGTGGACTCCATCACAATCCGAAGCTCTTCGCGCATGCTTC
GAGCGCAATCCCTATCCGGGGATTGCCACAAGGGAGAGGCTTGCACAGGCTATCGG
AATCCCGGAACCGAGAGTGCAGATCTGGTTCCAAAATGAACGCTCTCGGCAGCTCA
GACAGCATCGCAGGGAGTCCCGCCCGTGGCCAGGAAGAAGGGGACCACCTGAAGG
GAGGAGACCCAGAACCAGGCTGACAAGTCTGCAGCTGAGAATCCTTGGTCAGGCTT
TTGAAAGGAATCCAAGGCCAGGATTTGCCACCAGAGAGGAACTGGCCAGGGATACA
GGCCTTCCTGAGGATACTATCCATATCTGGTTCCAGAACAGGAGGGCCAGGAGAAG
GCACAGAAGGGGACGGGCACCTGCTCAGGCCGGTGGACTCTGCTCTGCTGCCCCTG
GGGGCGGCCATCCAGCACCTTCCTGGGTGGCTTTCGCTCATACTGGCGCTTGGGGTA
CCGGGCTGCCTGCTCCGCATGTTCCCTGTGCTCCAGGGGCCCTCCCGCAGGGAGCGT
TTGTTTCCCAGGCAGCTAGGGCTGCACCTGCCCTGCAACCATCACAGGCAGCGCCAG
CTGAAGGCATCAGCCAACCCGCCCCAGCCCGCGGAGATTTTGCTTATGCAGCGCCA
GCACCTCCAGACGGTGCCCTGAGCCACCCCCAAGCCCCCAGATGGCCCCCTCACCCT
GGTAAGTCCCGGGAAGACCGCGATCCCCAACGAGATGGACTGCCCGGTCCTTGCGC
TGTGGCCCAGCCAGGACCTGCTCAAGCCGGCCCTCAGGGGCAAGGAGTGCTGGCCC
CACCTACAAGCCAGGGATCTCCCTGGTGGGGTTGGGGACGCGGACCTCAGGTTGCT
GGAGCCGCTTGGGAGCCTCAGGCCGGAGCTGCACCGCCGCCACAACCGGCCCCTCC
CGACGCGTCAGCGTCCGCCCGACAAGGCCAGATGCAGGGAATCCCAGCACCTAGCC
AAGCTCTTCAAGAGCCTGCCCCTTGGAGCGCACTGCCGTGTGGGCTGCTCCTGGATG
AACTCCTGGCTAGCCCAGAATTTCTCCAGCAGGCACAGCCACTCCTGGAAACAGAA
GCTCCGGGAGAGCTCGAAGCCTCCGAAGAAGCAGCAAGCCTGGAGGCACCTCTTTC
CGAGGAGGAGTATAGAGCCCTTCTGGAAGAACTTTGA

> HMH_aminoAcid
##NOTE: HMH is the second mDUX homeodomain introduced into hDUX4 in place of the hDUX4 second homeodomain
MALPTPSDSTLPAEARGRGRRRRLVWTPSQSEALRACFERNPYPGIATRERLAQAIGIPEP
RVQIWFQNERSRQLRQHRRESRPWPGRRGPPEGRRPRTRLTSLQLRILGQAFERNPRPGF
ATREELARDTGLPEDTIHIWFQNRRARRRHRRGRAPAQAGGLCSAAPGGGHPAPSWVA
FAHTGAWGTGLPAPHVPCAPGALPQGAFVSQAARAAPALQPSQAAPAEGISQPAPARG
DFAYAAPAPPDGALSHPQAPRWPPHPGKSREDRDPQRDGLPGPCAVAQPGPAQAGPQG
QGVLAPPTSQGSPWWGWGRGPQVAGAAWEPQAGAAPPPQPAPPDASASARQGQMQGI
PAPSQALQEPAPWSALPCGLLLDELLASPEFLQQAQPLLETEAPGELEASEEAASLEAPLS
EEEYRALLEEL*

```
> MHH_nucleotide
##NOTE: MHH is the first mDUX homeodomain introduced into hDUX4 in place of the
hDUX4 first homeodomain
ATGGCATTGCCTACACCTTCAGACTCTACGCTGCCTGCAGAGGCTAGGGGAAGAAG
GAGAAGGAGGAAAACTGTCTGGCAAGCTTGGCAGGAACAGGCACTCCTGAGCACAT
TTAAGAAAAAAAGGTATCTGTCCTTTAAAGAAAGAAAGGAACTGGCAAAAAGGATG
GGAGTTTCTGATTGCAGGATCAGAGTCTGGTTCCAGAATAGGAGAAATAGGTCTGG
GGAGGAAGGACGCAGGGAGTCCCGCCCGTGGCCAGGAAGAAGGGGACCACCTGAA
GGAAGAAGAAAACGCACAGCGGTGACTGGCAGCCAAACGGCTCTGCTGCTCCGCGC
TTTCGAGAAAGATCGGTTCCCCGGAATTGCCGCACGCGAAGAACTCGCCAGAGAAA
CTGGGCTCCCAGAATCACGAATACAGATTTGGTTCCAGAACCGCAGAGCAAGACAC
CCAGGCCAGGGGGGACGGGCACCTGCTCAGGCCGGTGGACTCTGCTCTGCTGCCCC
TGGGGGCGGCCATCCAGCACCTTCCTGGGTGGCTTTCGCTCATACTGGCGCTTGGGG
TACCGGGCTGCCTGCTCCGCATGTTCCCTGTGCTCCAGGGGCCCTCCCGCAGGGAGC
GTTTGTTTCCCAGGCAGCTAGGGCTGCACCTGCCCTGCAACCATCACAGGCAGCGCC
AGCTGAAGGCATCAGCCAACCCGCCCCAGCCCGCGGAGATTTTGCTTATGCAGCGC
CAGCACCTCCAGACGGTGCCCTGAGCCACCCCCAAGCCCCCAGATGGCCCCCTCACC
CTGGTAAGTCCCGGGAAGACCGCGATCCCCAACGAGATGGACTGCCCGGTCCTTGC
GCTGTGGCCCAGCCAGGACCTGCTCAAGCCGGCCCTCAGGGGCAAGGAGTGCTGGC
CCCACCTACAAGCCAGGGATCTCCCTGGTGGGGTTGGGGACGCGGACCTCAGGTTG
CTGGAGCCGCTTGGGAGCCTCAGGCCGGAGCTGCACCGCCGCCACAACCGGCCCCT
CCCGACGCGTCAGCGTCCGCCCGACAAGGCCAGATGCAGGGAATCCCAGCACCTAG
CCAAGCTCTTCAAGAGCCTGCCCCTTGGAGCGCACTGCCGTGTGGGCTGCTCCTGGA
TGAACTCCTGGCTAGCCCAGAATTTCTCCAGCAGGCACAGCCACTCCTGGAAACAG
AAGCTCCGGGAGAGCTCGAAGCCTCCGAAGAAGCAGCAAGCCTGGAGGCACCTCTT
TCCGAGGAGGAGTATAGAGCCCTTCTGGAAGAACTTTGA
```

```
> MHH_aminoAcid
##NOTE: MHH is the first mDUX homeodomain introduced into hDUX4 in place of the
hDUX4 first homeodomain
MALPTPSDSTLPAEARGRRRRRKTVWQAWQEQALLSTFKKKRYLSFKERKELAKRMG
VSDCRIRVWFQNRRNRSGEEGRRESRPWPGRRGPPEGRRKRTAVTGSQTALLLRAFEKD
RFPGIAAREELARETGLPESRIQIWFQNRRARHPGQGGRAPAQAGGLCSAAPGGGHPAP
SWVAFAHTGAWGTGLPAPHVPCAPGALPQGAFVSQAARAAPALQPSQAAPAEGISQPA
PARGDFAYAAPAPPDGALSHPQAPRWPPHPGKSREDRDPQRDGLPGPCAVAQPGPAQA
GPQGQGVLAPPTSQGSPWWGWGRGPQVAGAAWEPQAGAAPPPQPAPPDASASARQGQ
MQGIPAPSQALQEPAPWSALPCGLLLDELLASPEFLQQAQPLLETEAPGELEASEEAASL
EAPLSEEEYRALLEEL*
```

Figure 19. Canine DUXC Isoforms expressed in Testes



UCSC Genome Browser on Dog Sep. 2011 (Broad CanFam3.1/canFam3) Assembly

chr17:64,234,616-64,241,916 7,301 bp.

move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
100x
enter position, gene symbol or search terms
go
chr17
17
Scale
chr17:
2 kb
canFam3
64,236,000
64,237,000
64,238,000
64,239,000
64,240,000
64,241,000
pCR4-cDUXC.Intra-HD.long.all
pCR4-cDUXC.2.17_both_reversed
pCR4-cDUXC.2.13_both_reversed
pCR4-cDUXC.2.07_both_reversed
pCR4-cDUXC.2.04_both_reversed
pCR4-cDUXC.1.17_both_reversed
pCR4-cDUXC.1.13_both_reversed
pCR4-cDUXC.1.10_both
pCR4-cDUXC.1.11_both_reversed
pCR4-cDUXC.Intra-HD.short.all
pCR4-cDUXC.2.15_both_reversed
pCR4-cDUXC.2.09_both
pCR4-cDUXC.2.06_both_reversed
pCR4-cDUXC.2.05_both_reversed
pCR4-cDUXC.1.16_both
pCR4-cDUXC.1.15_both_reversed
pCR4-cDUXC.1.14_both_reversed
pCR4-cDUXC.1.12_both_reversed
DUXC_JLW_archetypes
canine_DUXC_JLW_short_2.15
canine_DUXC_JLW_long_consensus
cDuxC_SA
cDuxC_SA
cDuxC_KS
DUXC-KS6
DUXC-KS5
DUXC-KS4
DUXC-KS3
DUXC-KS1
cDuxC_LG cDNA clone
cDuxC_LG
cDuxC_Genscript
cDuxC_Genscript
cDuxC_Genscript_CTD2
cDuxC_Genscript_CTD1
cDuxC_Genscript_HD1
cDuxC_Genscript_HD2

Figure 20. Comparison between predicted and observed canine DUXC isoforms

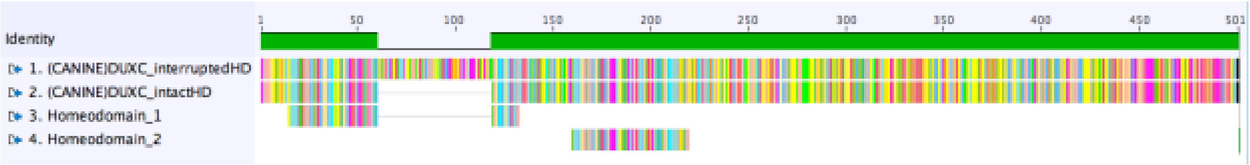
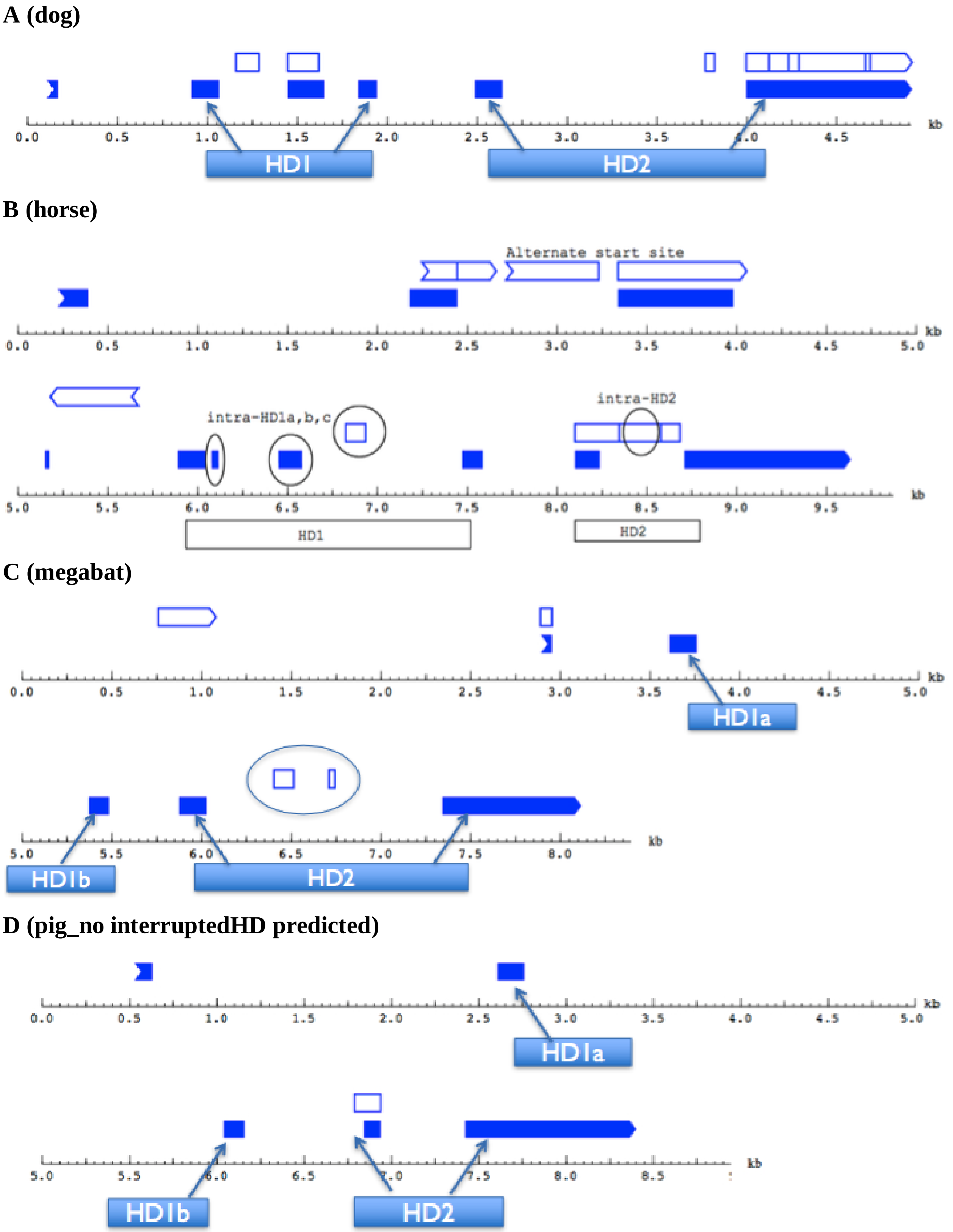


Figure 21. Predicted "interruptedHD" isoforms in non-canine species



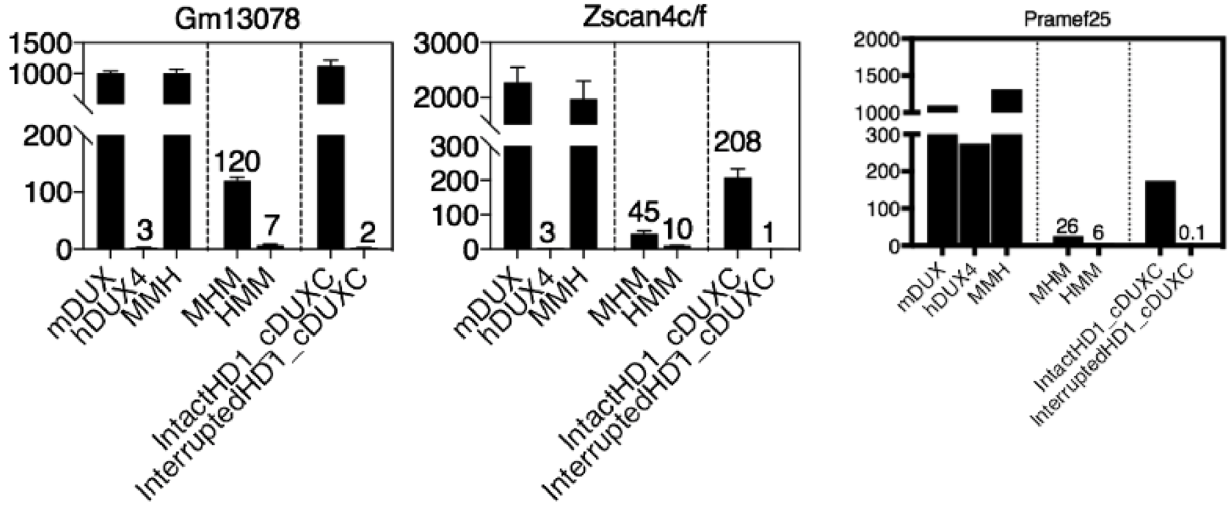
A (dog)

B (horse)

C (megabat)

D (pig_no interruptedHD predicted)

Figure 22. cDUXC_interruptedHD isoform does not activate three mDUX target genes



Chapter 5. Discussion

The goal of my thesis was to test the hypothesis that the mouse double homeobox protein, mDUX, is a functional homolog of the disease-causing human double homeobox protein, hDUX4. I found that mDUX and hDUX4 regulate similar transcriptomes in their correct-species contexts, which suggests functional homology. In this regard, it is interesting to note that both factors bind and activate endogenous retroviruses (ERVs), but that many of these ERVs are lineage-specific and not shared between mice and humans.

My work to characterize the mDUX transcriptome following RNA-sequencing lead to the discovery that the mDUX transcriptome in mouse skeletal muscle cells is strikingly similar to the mouse two-cell embryo transcriptome. Subsequent work by our collaborators immediately corroborated my finding and extended it to hDUX4, such that we now know that hDUX4 and mDUX are both expressed in cleavage stage embryos and activate transcriptomes reminiscent of this early developmental time point when mis-expressed in models of early embryos (human induced pluripotent stem cells and mouse embryonic stem cells) or cultured skeletal muscle cells. The cleavage stage embryo is remarkable because it is totipotent and exists at a time when the embryonic genome is transcribed for the first time. Further work will need to determine whether mDUX is necessary for development such that a mDUX knockout would be embryonic lethal or not. Additional studies should also investigate the interplay between DUX4-family genes and chromatin modifiers maternally deposited and transcribed at EGA and single homeodomain transcription factors activated by the DUX genes.

To better understand the extent to which DUX4-family factors are truly functional homologs, I asked whether they could function interchangeably by expressing hDUX4 in mouse cells and found that while they are not interchangeable, they do share the ability to activate some early developmental genes. This partial cross-species functional conservation is consistent with the partial conservation, partial divergence of the two factors' consensus binding motifs. As DUX factors contain two DNA-binding homeodomains, I next tested the hypothesis that the homeodomains function modularly and the similarities and differences of the consensus binding motifs reflect the similarities and differences of the proteins' homeodomains. I found that there is some modularity at conventional promoters (i.e. promoters that do not contain ERVs) particularly the second and more-conserved homeodomain, but there is no modularity at ERV-promoted genes. My thesis work increased our understanding of the evolutionary history of DUX genes, the physiological functions of DUX genes and the role mDUX can play in furthering our understanding of and treatment options for the hDUX4-caused human disease Facioscapulohumeral Muscular Dystrophy (FSHD).

Retroposition and functional evolution of DUX genes

In studying both DUXC genes and DUXC retrogenes, why work also addressed the consequences of retroposition. While retroposition can affect a gene in many ways such as changing its promoter sequence leading to changes in cell type expression and introducing mutations leading to premature stop codons, I focused on the impact of intron loss and thus the fixation of a particular isoform. DUXC genes have not been found to co-exist in genomes with DUXC retrogenes, so that retroposition not only locked in one isoform, it seems to have been at the expense of the intron-containing DUXC gene. DUXC genes with introns can use alternative splicing to create different isoforms and indeed, we have evidence of the canine DUXC gene creating at least two different isoforms – one we observe expressed in adult canine testes (i.e. "interruptedHD" isoform) and one we predict to exist in some tissue at some developmental time point because it is a robust transcriptional activator of some early embryo genes when mis-expressed in mouse skeletal muscle cells (i.e. "intactHD" isoform).

A key question to address in future studies will be the location and/or timing of expression of the intactHD isoform and the differences in consensus binding motifs and target genes between the intactHD and interruptedHD isoforms. DUXC retrogenes are often identified by the presence of two intact homeodomains, it remains an open question whether all DUXC retrogenes were created from the intactHD isoform of DUXC or we only have identified DUXC retrogenes from the intactHD isoform because of the search criteria used, which required two intact homeodomains to call a putative retrogene as a DUXC retrogene. It will be critical to determine whether these isoforms identified in canine DUXC are representative of the ancestral DUXC by assessing the presence of similar isoforms in additional Laurasiatherian species. Nevertheless, one hypothesis to test is that mDUX and hDUX4 (both DUXC retrogenes from the intactHD isoform) are more similar in function to the intactHD isoform than the interruptedHD isoform. One possibility is that the interruptedHD isoform functions as a single homeodomain factor at a subset of targets affected by the intactHD isoform, which is a double homeodomain factor. In vitro experiments that measure protein-DNA binding affinity, such as the electromobility shift assay (EMSA), will be necessary to fully explore whether mDUX and hDUX4 retain the ability to bind DNA as a single homeodomain factor or whether single homeodomain binding could be a function that was lost during isoform selection during retroposition. Alternative, if mDUX and hDUX4 cannot bind DNA as a single homeodomain factor, perhaps single homeodomain factors that they activate now serve this function and indeed we observe several single homeodomain factors increase in expression following mDUX and hDUX4 expression.

In a similar vein, it will be interesting to determine whether retroposition of the intactHD isoform changed the DNA-binding paradigm used by these factors: dimers or singletons. One hypothesis is that the interruptedHD isoform bound DNA as a dimer while the intactHD isoform bound DNA as a singleton such that following retroposition, the retrogene-encoded factor was locked into the singleton DNA-binding paradigm. Dimerization state can be determined using EMSA's as well and preliminary data did not show hDUX4 binding as a dimer, but the DNA-binding paradigms of the canine DUXC intactHD and interruptedHD isoforms are currently unknown.

Homeodomain duplication, placentation and ERVs

Double homeobox genes have only been found in placental mammals, such that the prevailing hypothesis is that a single homeobox gene duplicated in the last common ancestor of placental mammals to create the double homeobox genes. In support of this hypothesis, a single homeobox gene has been identified in the *Duxbl*-syntentic locus in species outside of placental mammals (i.e. chicken and opossum). One hypothesis for the mechanism of homeodomain duplication is that a local duplication of the entire single homeodomain gene happened first and then there was a recombination event between these two genes to create a single gene with two homeodomains and in fact, opossum has multiple copies of the putative ancestral single homeodomain gene.

Given that double homeodomain genes were created at the same time as the placenta in evolutionary history, a natural question is whether the two events were related. Put a different way, what benefit did homeodomain duplication confer such that double homeodomain genes were maintained and was this benefit related to placentation?

Limited data suggests that hDUX4 has a role in the placenta as its transcript, protein and target gene mRNAs can be observed there. Future studies will be needed to characterize the nature of hDUX4's role in the placenta. However, it is interesting to note that the placenta is a site of endogenous retrovirus (ERV) transcription and proteins originally encoded by ERVs have been 'domesticated' for use in the cell-cell fusion necessary for placenta formation (i.e. syncytins).

My work on mDUX established a relationship between mDUX and endogenous retroviruses and previous work established a parallel relationship between hDUX4 and ERVs. In future studies, it will be interesting to determine whether all DUX4-family genes bind and activate transcription at ERVs. If so, these data would support the hypothesis that the relationship between DUX4-family genes and ERVs was established at the time of homeodomain duplication. One relevant feature of ancestral endogenous retroviruses is that their transcription could have led to mobilization and amplification in copy number using a 'copy and paste' mechanism; this ability has been retained in some lineages (e.g. mouse) and lost in others (e.g. humans). However, if an ancestral DUX gene led to the amplification of ERVs in the last common ancestor of placental mammals, we might expect DUX4-family proteins to regulate a shared set of very old ERVs (i.e. ERVs that integrated into the genome a long time ago), but such a shared set of ERVs has not been identified between hDUX4-regulated ERVs and mDUX-regulated ERVs. The lack of a shared set of ERVs identified might be due to the complexities of assigning orthology to ERVs between species due to their prevalence and high sequence similarity and thus, it remains an open question.

In conclusion, mDUX is a DUXC retrogene that shares core transcriptional features with the human DUXC retrogene (hDUX4) in that they both activate transcription of genes and retrotransposons highly specific to cleavage stage embryos. These data and data from our collaborators indicate that mDUX and hDUX4 share the conserved physiological role of transcriptional activation in totipotent embryos. Limited data from the canine DUXC gene show that this functional conservation might extend to all DUXC genes and retrogenes. These findings have broad implications for the co-evolutionary history of DUX4-family genes and ERVs and opens alternative avenues for developing animal models of FSHD based on mDUX and canine DUXC expression.

References

1 Mostacciuolo, M.L., Pastorello, E., Vazza, G., Miorin, M., Angelini, C., Tomelleri, G., Galluzzi, G. and Trevisan, C.P. (2009) Facioscapulohumeral muscular dystrophy: epidemiological and molecular study in a north-east Italian population sample. *Clin Genet*, **75**, 550-555.

2 Sposito, R., Pasquali, L., Galluzzi, F., Rocchi, A., Solito, B., Soragna, D., Tupler, R. and Siciliano, G. (2005) Facioscapulohumeral muscular dystrophy type 1A in northwestern Tuscany: a molecular genetics-based epidemiological and genotype-phenotype study. *Genet Test*, **9**, 30-36.

3 Flanigan, K.M., Coffeen, C.M., Sexton, L., Stauffer, D., Brunner, S. and Leppert, M.F. (2001) Genetic characterization of a large, historically significant Utah kindred with facioscapulohumeral dystrophy. *Neuromuscul Disord*, **11**, 525-529.

4 Deenen, J.C., Arnts, H., van der Maarel, S.M., Padberg, G.W., Verschuuren, J.J., Bakker, E., Weinreich, S.S., Verbeek, A.L. and van Engelen, B.G. (2014) Population-based incidence and prevalence of facioscapulohumeral dystrophy. *Neurology*, **83**, 1056-1059.

5 Orrell, R.W. (2011) Facioscapulohumeral dystrophy and scapuloperoneal syndromes. *Handb Clin Neurol*, **101**, 167-180.

6 (1997) A prospective, quantitative study of the natural history of facioscapulohumeral muscular dystrophy (FSHD): implications for therapeutic trials. The FSH-DY Group. *Neurology*, **48**, 38-46.

7 Landouzy L, D.J. (1885) De la myopathie atrophique progressive. *Rev Med Franc* in press., 81–253.

8 Wijmenga, C., Frants, R.R., Brouwer, O.F., Moerer, P., Weber, J.L. and Padberg, G.W. (1990) Location of facioscapulohumeral muscular dystrophy gene on chromosome 4. *Lancet*, **336**, 651-653.

9 Hewitt, J.E., Lyle, R., Clark, L.N., Valleley, E.M., Wright, T.J., Wijmenga, C., van Deutekom, J.C., Francis, F., Sharpe, P.T., Hofker, M. *et al.* (1994) Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Hum Mol Genet*, **3**, 1287-1295.

10 Snider, L., Asawachaicharn, A., Tyler, A.E., Geng, L.N., Petek, L.M., Maves, L., Miller, D.G., Lemmers, R.J., Winokur, S.T., Tawil, R. *et al.* (2009) RNA transcripts, miRNA-sized

fragments and proteins produced from D4Z4 units: new candidates for the pathophysiology of facioscapulohumeral dystrophy. *Hum Mol Genet*, **18**, 2414-2430.

11 Dixit, M., Ansseau, E., Tassin, A., Winokur, S., Shi, R., Qian, H., Sauvage, S., Matteotti, C., van Acker, A.M., Leo, O. *et al.* (2007) DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proc Natl Acad Sci U S A*, **104**, 18157-18162.

12 Rijkers, T., Deidda, G., van Koningsbruggen, S., van Geel, M., Lemmers, R.J., van Deutekom, J.C., Figlewicz, D., Hewitt, J.E., Padberg, G.W., Frants, R.R. *et al.* (2004) FRG2, an FSHD candidate gene, is transcriptionally upregulated in differentiating primary myoblast cultures of FSHD patients. *J Med Genet*, **41**, 826-836.

13 Gabellini, D., Green, M.R. and Tupler, R. (2002) Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell*, **110**, 339-348.

14 Alexiadis, V., Ballestas, M.E., Sanchez, C., Winokur, S., Vedanarayanan, V., Warren, M. and Ehrlich, M. (2007) RNAPol-ChIP analysis of transcription from FSHD-linked tandem repeats and satellite DNA. *Biochim Biophys Acta*, **1769**, 29-40.

15 Clapp, J., Mitchell, L.M., Bolland, D.J., Fantes, J., Corcoran, A.E., Scotting, P.J., Armour, J.A. and Hewitt, J.E. (2007) Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. *Am J Hum Genet*, **81**, 264-279.

16 Lemmers, R.J., van der Vliet, P.J., Klooster, R., Sacconi, S., Camano, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W. *et al.* (2010) A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*, **329**, 1650-1653.

17 Lek, A., Rahimov, F., Jones, P.L. and Kunkel, L.M. (2015) Emerging preclinical animal models for FSHD. *Trends Mol Med*, **21**, 295-306.

18 Leidenroth, A. and Hewitt, J.E. (2010) A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evol Biol*, **10**, 364.

19 Kaessmann, H., Vinckenbosch, N. and Long, M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet*, **10**, 19-31.

20 Leidenroth, A., Clapp, J., Mitchell, L.M., Coneyworth, D., Dearden, F.L., Iannuzzi, L. and Hewitt, J.E. (2012) Evolution of DUX gene macrosatellites in placental mammals. *Chromosoma*, **121**, 489-497.

21 Bosnakovski, D., Daughters, R.S., Xu, Z., Slack, J.M. and Kyba, M. (2009) Biphasic myopathic phenotype of mouse DUX, an ORF within conserved FSHD-related repeats. *PLoS One*, **4**, e7003.

22 Geng, L.N., Yao, Z., Snider, L., Fong, A.P., Cech, J.N., Young, J.M., van der Maarel, S.M., Ruzzo, W.L., Gentleman, R.C., Tawil, R. *et al.* (2012) DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev Cell*, **22**, 38-51.

23 Young, J.M., Whiddon, J.L., Yao, Z., Kasinathan, B., Snider, L., Geng, L.N., Balog, J., Tawil, R., van der Maarel, S.M. and Tapscott, S.J. (2013) DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. *PLoS Genet*, **9**, e1003947.

24 Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J. *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol*, **20**, 1131-1139.

25 Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E. *et al.* (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, **500**, 593-597.

26 Tohonen, V., Katayama, S., Vesterlund, L., Jouhilahti, E.M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Burglin, T.R. *et al.* (2015) Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat Commun*, **6**, 8207.

27 Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D. and Knowles, B.B. (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell*, **7**, 597-606.

28 Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D. and Pfaff, S.L. (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, **487**, 57-63.

29 Zalzman, M., Falco, G., Sharova, L.V., Nishiyama, A., Thomas, M., Lee, S.L., Stagg, C.A., Hoang, H.G., Yang, H.T., Indig, F.E. *et al.* (2010) Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature*, **464**, 858-863.

30 Amano, T., Hirata, T., Falco, G., Monti, M., Sharova, L.V., Amano, M., Sheer, S., Hoang, H.G., Piao, Y., Stagg, C.A. *et al.* (2013) Zscan4 restores the developmental potency of embryonic stem cells. *Nat Commun*, **4**, 1966.

31 Akiyama, T., Xin, L., Oda, M., Sharov, A.A., Amano, M., Piao, Y., Cadet, J.S., Dudekula, D.B., Qian, Y., Wang, W. *et al.* (2015) Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. *DNA Res*, **22**, 307-318.

32 Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B. *et al.* (2011) Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev*, **25**, 594-607.

33 Wasson, J.A., Simon, A.K., Myrick, D.A., Wolf, G., Driscoll, S., Pfaff, S.L., Macfarlan, T.S. and Katz, D.J. (2016) Maternally provided LSD1/KDM1A enables the maternal-to-zygotic transition and prevents defects that manifest postnatally. *Elife*, **5**.

34 Ishiuchi, T., Enriquez-Gasca, R., Mizutani, E., Boskovic, A., Ziegler-Birling, C., Rodriguez-Terrones, D., Wakayama, T., Vaquerizas, J.M. and Torres-Padilla, M.E. (2015) Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat Struct Mol Biol*, **22**, 662-671.

35 Takahashi, K. and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663-676.

36 Tawil, R., van der Maarel, S.M. and Tapscott, S.J. (2014) Facioscapulohumeral dystrophy: the path to consensus on pathophysiology. *Skelet Muscle*, **4**, 12.

37 Snider, L., Geng, L.N., Lemmers, R.J., Kyba, M., Ware, C.B., Nelson, A.M., Tawil, R., Filippova, G.N., van der Maarel, S.M., Tapscott, S.J. *et al.* (2010) Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS Genet*, **6**, e1001181.

38 Das, S. and Chadwick, B.P. (2016) Influence of Repressive Histone and DNA Methylation upon D4Z4 Transcription in Non-Myogenic Cells. *PLoS One*, **11**, e0160022.

39 Tawil, R. and Van Der Maarel, S.M. (2006) Facioscapulohumeral muscular dystrophy. *Muscle Nerve*, **34**, 1-15.

40 Lemmers, R.J., Tawil, R., Petek, L.M., Balog, J., Block, G.J., Santen, G.W., Amell, A.M., van der Vliet, P.J., Almomani, R., Straasheijm, K.R. *et al.* (2012) Digenic inheritance of an

SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet*, **44**, 1370-1374.

41 Wallace, L.M., Garwick, S.E., Mei, W., Belayew, A., Coppee, F., Ladner, K.J., Guttridge, D., Yang, J. and Harper, S.Q. (2011) DUX4, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. *Ann Neurol*, **69**, 540-552.

42 Krom, Y.D., Thijssen, P.E., Young, J.M., den Hamer, B., Balog, J., Yao, Z., Maves, L., Snider, L., Knopp, P., Zammit, P.S. *et al.* (2013) Intrinsic epigenetic regulation of the D4Z4 macrosatellite repeat in a transgenic mouse model for FSHD. *PLoS Genet*, **9**, e1003415.

43 Dandapat, A., Bosnakovski, D., Hartweck, L.M., Arpke, R.W., Baltgalvis, K.A., Vang, D., Baik, J., Darabi, R., Perlingeiro, R.C., Hamra, F.K. *et al.* (2014) Dominant lethal pathologies in male mice engineered to contain an X-linked DUX4 transgene. *Cell Rep*, **8**, 1484-1496.

44 Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, 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.

45 Yao, Z., Snider, L., Balog, J., Lemmers, R.J., Van Der Maarel, S.M., Tawil, R. and Tapscott, S.J. (2014) DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum Mol Genet*, **23**, 5342-5352.

46 Jagannathan, S., Shadle, S., Resnick, R., Snider, L., Tawil, R.N., van der Maarel, S.M., Bradley, R.K. and Tapscott, S.J. (2016) Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. *Hum Mol Genet*, in press.

47 Falco, G., Lee, S.L., Stanghellini, I., Bassey, U.C., Hamatani, T. and Ko, M.S. (2007) Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Dev Biol*, **307**, 539-550.

48 Zhang, W., Walker, E., Tamplin, O.J., Rossant, J., Stanford, W.L. and Hughes, T.R. (2006) Zfp206 regulates ES cell gene expression and differentiation. *Nucleic Acids Res*, **34**, 4780-4790.

49 Coordinators, N.R. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **44**, D7-19.

50 Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, **37**, W202-208.

51 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.

52 Blond, J.L., Lavillette, D., Cheynet, V., Bouton, O., Oriol, G., Chapel-Fernandes, S., Mandrand, B., Mallet, F. and Cosset, F.L. (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *J Virol*, **74**, 3321-3329.

53 Assou, S., Boumela, I., Haouzi, D., Monzo, C., Dechaud, H., Kadoch, I.J. and Hamamah, S. (2012) Transcriptome analysis during human trophectoderm specification suggests new roles of metabolic and epigenetic genes. *PLoS One*, **7**, e39306.

54 Bosnakovski, D., Xu, Z., Gang, E.J., Galindo, C.L., Liu, M., Simsek, T., Garner, H.R., Agha-Mohammadi, S., Tassin, A., Coppee, F. *et al.* (2008) An isogenetic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. *EMBO J*, **27**, 2766-2779.

55 Eidahl, J.O., Giesige, C.R., Domire, J.S., Wallace, L.M., Fowler, A.M., Guckes, S., Garwick-Coppens, S., Labhart, P. and Harper, S.Q. (2016) Mouse Dux is myotoxic and shares partial functional homology with its human paralog DUX4. *Hum Mol Genet*, in press.

56 Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105-1111.

57 Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

58 Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J.P. (2006) GenePattern 2.0. *Nat Genet*, **38**, 500-501.

59 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

60 Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*, **44**, D336-342.

61 Conerly, M.L., Yao, Z., Zhong, J.W., Groudine, M. and Tapscott, S.J. (2016) Distinct Activities of Myf5 and MyoD Indicate Separate Roles in Skeletal Muscle Lineage Specification and Differentiation. *Dev Cell*, **36**, 375-385.

62 Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L. *et al.* (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell*, **18**, 662-674.

63 Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

64 Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**, 207-210.

65 Choi, J., Costa, M.L., Mermelstein, C.S., Chagas, C., Holtzer, S. and Holtzer, H. (1990) MyoD converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc Natl Acad Sci U S A*, **87**, 7988-7992.

66 Davis, R.L., Weintraub, H. and Lassar, A.B. (1987) Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, **51**, 987-1000.

67 Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M.A., Lassar, A.B. and Miller, A.D. (1989) Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A*, **86**, 5434-5438.

68 Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, **29**, 24-26.

69 Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, **14**, 178-192.

70 Sigler, P.B. (1988) Transcriptional activation. Acid blobs and negative noodles. *Nature*, **333**, 210-212.

71 Plevin, M.J., Mills, M.M. and Ikura, M. (2005) The LxxLL motif: a multifunctional binding sequence in transcriptional regulation. *Trends Biochem Sci*, **30**, 66-69.

72 Choi, S.H., Gearhart, M.D., Cui, Z., Bosnakovski, D., Kim, M., Schennum, N. and Kyba, M. (2016) DUX4 recruits p300/CBP through its C-terminus and induces global H3K27 acetylation changes. *Nucleic Acids Res*, **44**, 5161-5173.

VITA

Jennifer L. Whiddon was born in Midland, Texas, but spent most of her youth in The Woodlands, Texas. In 2008, she graduated from the University of Texas at Austin with a Bachelor of Science in Psychology as well as a Bachelor of Arts in Plan II. Following graduation, Jenn spent two years teaching science to middle school students at KIPP: Austin College Prep. After taking a year off to travel domestically and in Southeast Asia, she earned a Doctor of Philosophy at the University of Washington in Molecular and Cellular Biology. She performed her doctorate work at Fred Hutchinson Cancer Research Center in the laboratory of Dr. Stephen Tapscott. In 2016, she returned to Texas to begin her postdoctoral studies.