Cis and trans modifiers in facioscapulohumeral muscular dystrophy

Sikrová, D.

Cis and *trans* modifiers in facioscapulohumeral muscular dystrophy

Darina Šikrová

***Cis* and *trans* modifiers in facioscapulohumeral muscular dystrophy**
Darina Šikrová
Leiden University Medical Center, The Netherlands

ISBN: 978-94-6458-589-6

Layout & cover design: Publiss
Printing: Ridderprint

The studies described in this thesis have been performed at the Department of Human Genetics, Leiden University Medical Centre, The Netherlands.

Cis and *trans* modifiers in facioscapulohumeral muscular dystrophy

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 14 december 2022
klokke 10.00 uur

door

Darina Šikrová
geboren te Michalovce, Slovakia
in 1992

Promotor:
Prof.dr.ir. S.M. van der Maarel

Co-promotor:
Dr. J. Balog

Promotiecommissie:
Prof.dr. N. Geijsen
Dr. L. Clemens-Daxinger
Prof.dr. B. van Engelen (Radboud University Medical Center)
Prof.dr. J.H.L.M. van Bokhoven (Radboud University Medical Center)

Table of contents

CHAPTER 1

General Introduction

Epigenetic regulation of mammalian genome

The functional output of our genome at any given time is not only determined by the information encoded in its genetic layer, i.e. the DNA sequence itself, but also by different *epigenetic layers* which help in its interpretation. Epigenetic layers mean factors or modifications controlling the stability and inheritance of gene expression patterns across different cell divisions or generations which are not the result of changes in the DNA sequence itself. This genome-wide epigenetic information is known as the *epigenome*. A growing number of diseases stem from mutations that alter different parts of the epigenome. Such mutations can affect chromatin configuration *in cis* or alter either the abundance or activity of epigenetic modifiers leading to epigenetic changes *in trans,* ultimately affecting gene expression. This thesis focuses on one such disease called Facioscapulohumeral muscular dystrophy (FSHD), in which epigenetic deregulation of a specific macrosatellite repeat array favors the expression of its embedded gene. Therefore, the first part of the Introduction is devoted to describing different epigenetic mechanisms involved in transcriptional regulation of the genome, mainly focusing on the epigenetic silencing of repetitive elements. The second part of the Introduction will discuss the genetic and epigenetic etiology of FSHD.

DNA methylation

One of the epigenetic layers that regulates the genome is the direct modification of DNA bases. The most studied DNA modification in mammals is methylation of the 5th carbon of cytosine (*5mC*), generally referred to as DNA methylation. The existence of another form of methylation in genomic DNA, that of the 6th carbon of adenine (*6mA*), while being prevalent in prokaryotes [1], remains disputable in mammals [2]. In the mouse and human genome, around 5% of all cytosines are methylated [3] making 5mC a relatively abundant modification which is therefore sometimes referred to as the fifth DNA base (next to adenine, guanine, cytidine and thymine). The 5mC is usually found in a CpG dinucleotide context resulting in two 5mCs positioned diagonally to each other on opposite DNA strands [4]. The occurrence of 5mC in this symmetrical CpG context allows for faithful reproduction of the methylation information during DNA replication as a 5mC on the mother strand serves as a template for the methylation machinery to methylate the cytosine on a newly replicated daughter strand [5]. In most cell types, except for specific stages during embryogenesis [6] and gametogenesis [7], around 80% of CpGs are methylated [8]. These are typically isolated CpGs, while the remaining unmethylated CpGs are usually clustered in *CpG islands* (*CGIs*), genomic regions which contain a higher density of CpGs than one would expect by chance [9]. These unmethylated CGIs are predominantly associated with promoters of active genes. The exceptions to this are promoter CGIs of three classes of genes for which life-long stable silencing mediated by promoter methylation in somatic tissues is crucial. These include genes on the inactive X chromosome [10], imprinted genes [11] and germline genes [12]. In addition, CGIs found within gene bodies (intragenic) or between genes (intergenic), collectively termed as 'orphan' CGIs, can also become methylated during development or are methylated in a tissue-specific manner [13]. Therefore, the lack of methylation at CGIs is often associated with active transcriptional start sites (TSSs) while their

methylation is associated with gene silencing [14]. In contrast to hypomethylated CGIs of active promoters, the bodies of actively transcribed genes are enriched with methylated CpGs which prevent spurious intragenic transcription initiation events [15,16]. Interestingly, around 20% of gene-associated CGIs in the human genome are absent from the homologous mouse genes and further analysis suggested that both humans and mice are losing CGIs over evolutionary time [17]. This can be explained by the hypermutability of 5mC since it is prone to spontaneous deamination resulting in C to T transitions in the genome, therefore resulting in progressive loss of CpGs through acquired transitions [18].[1]

The DNA methylation patterns are generally stably maintained in somatic cells. However, the somatic epigenome poses a major barrier to sexual reproduction and preparation for a next generation requires its resetting. The reconfiguration of genome-wide DNA methylation patterns happens in two steps during specific developmental time windows (reviewed here [19]). First, somatic methylation signatures are removed in the primordial germ cells (PGCs) and germ cell-specific as well as sex-specific signatures are established during later stages of germ cell development enabling meiotic maturation and subsequent fertilization [20]. After fertilization, the epigenome of a newly formed zygote becomes reprogrammed during subsequent cell divisions to erase gamete-specific signatures inherited from the oocyte and the sperm [21,22]. The DNA methylation erasure is completed at the pre-implantation blastocyst stage after which it is ready for the initiation of the embryonic developmental program and setting up lineage specific methylation profiles.

The life-cycle of DNA methylation is carried out by a collection of enzymes which can be considered based on their action either as the *writers* of this mark (DNA methyltransferases, DNMTs) or *erasers* (ten-eleven translocation enzymes, TETs). In mammals, writers belong to a family of DNMTs consisting of four catalytically active members (DNMT1, DNMT3A, DNMT3B and rodent-specific Dnmt3c) and one catalytically inactive member (DNMT3L), each of which evolved to perform largely non-overlapping functions [23].[2]

DNMT1 was the first DNMT identified [24] and for a long time recognized as a canonical maintenance DNMT because of its high affinity for hemi-methylated DNA [25,26] and its role in re-establishing CpG methylation patterns after DNA replication. However, this longstanding view has been challenged over time as some studies reported that it can also act *in vitro* on unmethylated DNA substrates, albeit with lower efficiency [27,28] and its *de novo* methylation activity was reported in oocytes outside the context of DNA replication [29] as well as during replication-coupled methylation maintenance [30,31]. Whether this *de novo* methylation

1 5mC does not exist in genomes of several widely used model organisms such as *Caenorhabditis elegans*, fission yeasts and bakers' yeasts and is found at very low levels only during early stages of embryonic development of *Drosophila*.

2 After identification of DNMT1, the second candidate mammalian DNMT gene was found which shared a sequence homology with DNMT1 and was named DNMT2 [318]. However, it turned out that it does not have properties that can be expected of a DNA methyltransferase as it has very low affinity towards double stranded DNA and it primarily localizes to the cytoplasm instead of the nucleus. Indeed, it was demonstrated that it is an RNA methyltransferase responsible for methylating the 38th cytosine residue in anticodon loop of certain tRNAs [319].

potential of DNMT1 is biologically relevant or creates only aberrant unspecific byproducts was addressed recently when it was demonstrated that murine Dnmt1 displays *de novo* methylation activity targeted at specific classes of retrotransposons [32]. Similarly, the strict classification of DNMT3A and DNMT3B as purely DNA methylation establishing DNMTs requires fine-tuning. Both DNMT3A and 3B are highly expressed during early embryonic development as well as in mouse embryonic stem cells. Upon differentiation, their expression dramatically declines which is in line with the assumption that they are then dispensable [33–35]. Nevertheless, they are essential for the long-term maintenance of DNA methylation imprints at least in mouse embryonic stem cells [36] and somatic inactivating mutations in DNMT3A have been reported in hematologic malignancies [37]. Moreover, DNMT3B isoforms without catalytic activity can act as accessory factors aiding DNMT1 activity in somatic cells [38]. This suggests that DNMTs could indeed work cooperatively to maintain methylation fidelity and that both DNMT3A and DNMT3B are also important in (some) somatic cells.

Dnmt3c and Dnmt3l, the two most recent evolutionary additions to the family of DNMTs, are involved in mammalian reproduction. *Dnmt3c* arose through a tandem duplication of the *Dnmt3b* gene specifically in the *Muroidea* lineage and is expressed only in male germ cells where it selectively methylates the promoters of evolutionarily young transposable elements thus ensuring their repression. This specialized Dnmt3c activity is required for male fertility [39]. Dnmt3l is a catalytically inactive cofactor that stimulates methyltransferase activities of Dnmt3a and Dnmt3b [40,41]. Similarly to Dnmt3c, its loss leads to male sterility due to the reactivation of certain classes of retrotransposons [42]. In addition, it is also required for proper oogenesis by helping Dnmt3a to establish maternal methylation imprints [43].

As mentioned before, mammalian genomes undergo two rounds of epigenomic resetting during which the majority of 5mC marks are removed. This can be accomplished by passive loss of DNA methylation through replication [44] or by its active removal by the TET family of proteins [45–47]. In mammals, the TET family consists of three members, TET1, TET2 and TET3, all of which catalyze the erasure of the 5mC modification in three sequential oxidation steps by generating 5-hydroxymethylcytosine (5hmC) which is further converted to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [47]. The last two products can be excised from DNA by the base excision repair pathway thus re-installing in unmodified cytosine bases [48,49].

Post-translational modifications of core histone tails

Another epigenetic layer is achieved by organizing DNA into a higher order structure known as *chromatin* which, amongst others, serves as a docking platform for other regulatory molecules. The smallest unit of chromatin is the *nucleosome*, which consists of 147 base pairs of DNA wrapped around two copies of four different histone proteins, usually H2A, H2B, H3 and H4 [50]. Other chromosome region-specific histone subvariants may occur such as centromeric protein A (CENPA) which is a centromere-associated H3 histone variant required for kinetochore assembly and for proper chromosome segregation during cell

division [51]. The N-terminal tails of histones extrude from the nucleosome and can undergo a variety of reversible and dynamic post-translational modifications (PTMs), predominantly at lysine or arginine residues [52]. Specific modifications can either directly influence the chromatin accessibility by changing the local charge or serve to recruit chromatin factors that either condense (repress) or relax (activate) chromatin. Furthermore, the local chromatin composition is a major determinant of the transcriptional activity of a locus.

Chromatin was initially divided into euchromatin and heterochromatin based on a different cytological staining density during interphase, where less compact and brighter stained regions were termed euchromatin and more compact densely stained regions were termed heterochromatin [53]. Nowadays, from a molecular perspective we recognize at least two types of heterochromatin, facultative and constitutive, which have distinct regulatory functions and are enriched for different proteins and protein modifications, but both result in transcriptional attenuation. Facultative heterochromatin is marked by Polycomb Group (PcG) proteins which exist in two separate protein complexes, PRC1 and PRC2, catalyzing monoubiquitination of lysine 119 on histone H2A (H2AK119Ub1) [54] or trimethylation of lysine 27 on histone H3 (H3K27me3) [55], respectively. In mammals, facultative heterochromatin regulates primarily the spatiotemporal expression of developmental genes [56,57] and the formation of the inactive X chromosome in females [58,59]. In contrast, constitutive heterochromatin is mainly formed at gene-poor and repeat-rich regions. A histone mark typical for constitutive heterochromatin is the trimethylation of lysine 9 on histone H3 (H3K9me3), which can be recognized by heterochromatin protein 1 (HP1) homologues promoting further chromatin compaction [60,61]. In mammals, deposition of H3K9me is catalyzed by at least six (five in humans) different H3K9 lysine methyltransferases (KMT) forming three distinct enzymatic systems, namely Suv39h1/Suv39h2, Eset1/Eset2 (Eset1 corresponds to human SETDB1) and G9a/Glp (also known as Ehmt2 and Ehmt1, respectively). Each of them targets different genomic regions. Suv39h1 and Suv39h2 are functionally redundant and primarily responsible for the deposition of H3K9me3 at centromeric and pericentromeric repeats [62,63]. Eset1 is important for silencing of endogenous and newly introduced retroviruses [64,65] and for the establishment of X inactivation [66,67], while G9a/Glp are important for early lineage commitment and permanent silencing of genes driving pluripotency [68–70]. Compound loss of all six H3K9 KMTs in mouse embryonic fibroblasts leads to complete dissolution of heterochromatin, transcriptional de-repression of nearly all families of repeat elements and genomic instability [71], marking the importance of H3K9me3 for transcriptional silencing and maintaining genome integrity. On the other hand, euchromatin represents an accessible chromatin state and contains transcriptionally active genes together with their regulatory elements. Promoters of actively transcribed genes are typically enriched for trimethylated lysine 4 on histone H3 (H3K4me3), while the bodies of actively transcribed genes are enriched for trimethylated lysine 36 on histone H3 (H3K36me3) [72,73].

Crosstalk between DNA methylation and histone modifications

The existence of bidirectional crosstalk between histone modifications and DNA methylation was initially hypothesized based on the observation of genome-wide colocalization of particular histone modifications with DNA methylation [74]. For example, DNA methylation is generally excluded from promoters of actively transcribed genes, whereas the bodies of actively transcribed genes are enriched for DNA methylation. These gene elements are also distinctly enriched for specific histone modifications such as H3K4me3 and H3K36me3, respectively. A mechanistic explanation was subsequently provided for these observations. Both DNMT3A and 3B enzymes contain apart from their methyltransferase domain also two chromatin reader domains, namely the ADD and PWWP domain, which allow for a direct readout of the H3 histone tail and thus help to regulate the deposition of DNA methylation in a chromatin state-aware manner. Specifically, the ADD domain recognizes unmodified H3K4 but is repelled by the increasing number of methyl moieties at K4 [75,76] and thus H3K4me3 acts as a shield against DNA methylation deposition. On the other hand, the PWWP domain directly binds to H3K36me3 and in this way the DNMT3 enzymes are targeted to the bodies of actively transcribed genes [77,78].

A peculiar example is the relationship between DNA methylation and H3K27me3. While they have been shown to co-occupy many CpG-poor regions, they are mutually exclusive at the CGI promoters of PcG target genes [79,80], whose promoters are co-marked by H4K4me3 and H3K27me3 in embryonic stem cells and at the E6.5 epiblast stage before lineage differentiation. This *bivalent* active/repressive state allows for these genes to be readily activated or repressed during lineage specification [81,82]. Furthermore, DNA methylation was shown to interfere with the PRC2 recognition of unmodified as well as H3K27me3-modified nucleosomes *in vitro* [83]. Consistent with this, the loss of DNA methylation results in genome-wide redistribution of H3K27me3 to regions which would be otherwise DNA methylated [80,84], while titrating it away from its native targets leading to their insufficient repression [84]. However, the fact that some regions can adopt both DNA methylation and H3K27me3 suggests that their coexistence might be context-dependent and that under certain unknown circumstances, the avoidance behavior of PRC2 towards methylated DNA sites can be overcome.

Perhaps the tightest cooperative relationship is between DNA methylation and H3K9me3 [74,85]. Together, these modifications enforce a more stable silenced chromatin state and aid each other in its initial establishment and its mitotic propagation. For example, H3K9me3 controls the maintenance of DNA methylation. DNMT1 recognizes H3K9me3 both directly via its RFTS domain [86] as well as indirectly through cooperation with its interacting factor UHRF1 [87,88]. These additional mechanisms, next to hemimethylated DNA itself, boost the fidelity of maintaining DNA methylation patterns in the H3K9me3 context. In addition, DNA methylation at major satellites, which form pericentric heterochromatin in mouse embryonic stem cells, is dependent on Suv39h1/2-mediated deposition of H3K9me3, which is in turn recognized by HP1 proteins facilitating the recruitment of *Dnmt3b* [89]. An earlier study using

immunofluorescence as a readout for H3K9me3 occupancy at major satellites claimed that H3K9me3 is retained at pericentric regions upon loss of DNA methylation in mouse embryonic stem cells lacking all three Dnmts (Dnmt1, Dnmt3a and Dnmt3b) [90]. However, a later study discovered a significant reduction in H3K9me3 in the same cells using quantitative mass spectrometry [91]. The reliance of H3K9me3 on DNA methylation became even more apparent when studying cells lacking DNMT1 which show reduced levels of H3K9me3 at pericentric regions [92]. However, it should be noted that pericentric heterochromatin represents a specific example of crosstalk between H3K9me3 and DNA methylation which cannot be automatically translated to other heterochromatic regions co-enriched for these two marks.

Epigenetic regulation of repetitive elements

Repetitive elements, which comprise over half of the human genome [93], can have a profound effect on gene regulation [94,95], chromosome (in)stability (reviewed here [96]), human health (reviewed here [97]) and can even drive species-specific adaptations [98]. Yet, their detailed annotation in human genome assemblies was lacking for a long time due to their repetitiveness. Recent advances in long-read sequencing technologies inspired a new consortium to follow in the footsteps of the Human Genome Project, which mission is to deliver gapless telomere-to-telomere chromosomes assemblies at base pair resolution (hence the name telomere-to-telomere or T2T consortium) and to generate the first complete assembly of the human genome since its first draft was published over 20 years ago [99,100]. The majority of these gaps are comprised of repetitive elements and several pre-prints are already starting to appear delivering comprehensive genetic and epigenetic annotations of previously known as well as newly discovered repetitive elements [93,101,102].

Classification of repetitive elements in mammalian genome

Based on the genomic organization, eukaryotic repeats can be classified into two classes: interspersed repeats and tandem repeats.

Interspersed repeats typically comprise transposable elements (TE) which can be further subdivided based on their mode of moving in the genome. Class I elements or retrotransposons work in a “copy and paste” mechanism in which they replicate themselves by reverse transcription and insert a new copy at the target site. Therefore, their copy number amplifies over time. In contrast, class II elements work in a “cut and paste” mechanism when a specialized enzyme, a transposase or a recombinase usually encoded by the TE itself, mediates its excision from the current position followed by insertion into a new genomic location. Size-wise, TEs are relatively short sequences (50 bp to 12 kb), however, it is their sheer number that can in some extreme cases make up almost 85% of the genome such in the case of wheat ((IWGSC) et al., 2018).[3] Although most TEs have lost their ability

3 One of the largest Class II TEs (up to 100 kb) was recently discovered in a model fungus *Podospora anserina* [320]. The authors whimsically name the new TE *“Enterprise”* as its transported “cargo” is a block of meiotic driver genes termed *Spoks (spore killing)*.

to mobilize further, some of these elements are still capable of "hopping" around causing insertional mutagenesis which can yield a neutral, deleterious or even advantageous outcome (reviewed here [104]). However, there is growing evidence that their main function in the genome is rather their capacity to influence the expression of neighbouring genes. Such function might look selfish at the first glance, however, there are specific instances when the host took advantage of this phenomenon and co-opted it into its own gene regulatory network. One of the most studied occurrences of transposon-mediated regulation of gene expression is during a zygotic genome activation when the embryonic transcriptional program is kickstarted [105]. One particular type of TEs, the murine endogenous retrovirus with leucine tRNA primer (MERVL), has been discovered as being central to this process in mice [106] with its human counterpart human ERVL (HERVL) serving the same function [107,108].

In contrast to interspersed repeats, tandem repeats are comprised of repetitive units which are usually organized in head-to-tail orientation and include multi-copy gene families (such as ribosomal DNA) and satellite repeats.[4] Depending on the length of the satellite unit, satellites can be classified in micro- (2-6 bp), mini- (10-100 bp) or macrosatellites (up to several kb). Tandem repeats often form structural elements of chromosomes which are important for genomic stability such as centromeric [109] and telomeric regions [109] or represent a boundary element driving higher-order chromosome architecture such as the DXZ4 macrosatellite repeat at the inactive X chromosome [110,111]. Furthermore, they show a high degree of polymorphism in their sequence, structure and their copy number [112] all of which can contribute to inter- as well as intra-species phenotypic variation, especially when a tandem repeat in question is formed by gene duplications [113–115]. However, copy number variation of some tandem repeats can also negatively impact human health if they alter the coding region or influence gene expression *in cis*. The most notable examples are microsatellite expansion disorders, in which the microsatellite copy-number increases in successive generations and once it reaches a certain threshold becomes unstable. Over 50 genetic disorders have been linked so far to such repeat expansions (reviewed here [97]). In addition, a reduction in tandem repeat copy number can also be detrimental as is the case for Facioscapulohumeral muscular dystrophy [116] which will be further discussed in the second part of this Introduction.

Regulators of the repeats' epigenetic state

Already in the early 90s, it was observed that integrating an increasing number of gene copies in tandem in plant genomes does not yield higher transcriptional output as compared to the single copy integration event [117,118], a surprisingly counterintuitive result as one would expect. Moreover, multiple tandem insertions are associated with higher DNA methylation, a mark that was as capable of modulating gene expression [119,120]. This phenomenon, when

4 The term satellite DNA was first coined by Pech et al. [321] and was referring to a DNA component that produces a specific *satellite* band that separates from the main DNA band during a caesium chloride density gradient centrifugation. As the density of DNA is a function of its base composition and highly homogeneous or repetitive sequences have this base ratio skewed, this will result in a different migration pattern along the density gradient compared to bulk DNA. The satellite DNA from the Pech paper was later confirmed to belong to centromeric AT-rich alpha satellites.

repetitive regions trigger *cis* heterochromatinization in a copy number-dependent manner, was termed *repeat-induced gene silencing* (RIGS) and was later also confirmed to operate in mammals [112]. Initially, RIGS was proposed to evolve as a protective mechanism of eukaryotic genomes against integration-prone foreign DNA elements such as viruses or transposons [121]. However, RIGS was later also recognized as a natural mechanism for regulation of expression of nearby genes [112] thus representing a case of position effect variegation (PEV) [122].

PEV refers to a phenomenon when a gene is placed (intentionally or by chance) in proximity to or within a heterochromatic region resulting in its stochastic transcriptional silencing (i.e. variegated expression) due to heterochromatin spreading into the juxtaposed locus. The pioneer of this field was Hermann Joseph Muller in the early 20th century who derived several *Drosophila* mutant lines with different variegated phenotypes due to X-ray induced chromosomal rearrangements [123]. Muller's discovery of PEV kickstarted new studies focusing on how gene expression is influenced by its chromatin environment. Numerous studies revealed many trans-acting modifiers which influence the probability of heterochromatin spreading and thus gene silencing (reviewed here [124]). Factors that increase the mutant phenotype were termed enhancers of variegation, while factors that decrease the mutant phenotype were coined suppressors of variegation. Later, these modifiers have been defined as either structural components of heterochromatin, enzymes that modify chromatin or as nuclear structural components and many of the identified factors were found to be conserved also in mammals [125].

Similarly to genetic screens to identify modifiers of PEV in *Drosophila*, analogous approaches were used to identify factors involved in RIGS in mammals using loci which show variegated phenotypes under genetic homogeneity thus allowing for uncovering factors whose mutation would skew the phenotypic spectrum one or the other way. Such loci, whose epigenetic state can intergenerationally switch from active to repressed, were termed *metastable epialleles* and were studied to capture both 1) the epigenetic basis of the phenotypes associated with these alleles and 2) the stochasticity of their epigenetic state.

The most relevant screen for this thesis is the one conducted in the Emma Whitelaw lab to search for modifiers of variegated multicopy transgene expression [126,127]. This screen used a transgenic inbred mouse line (GFP1 line) carrying a random integration of a transgene array consisting of ~11 copies of a construct in which the α-globin promoter and enhancer drive expression of a GFP reporter resulting in its variegated expression in red blood cells. Importantly, the variegated expression of this transgene is stable throughout generations culminating at around 55% of red blood cells being GFP positive [128]. A shift in the percentage of GFP-expressing red blood cells was used as a read-out in the offspring born to ENU-treated males and mutant alleles which showed enhanced or suppressed variegated expression were designated as *Modifiers of Murine Metastable Epialleles* (*Mommes*) [126]. This screen yielded more than 40 of such dominant mutant alleles (termed MommeDX or ^{MDX}, where "D" denotes a dominant screen and "X" a number referring to an allele in order

in which it was identified) and revealed previously known (e.g. *Dnmt1*, *Dnmt3b*, *Setdb1*, *Suv39h1*) as well as novel genes (e.g. *Smchd1*, *Rlf*, *D14Abb1e, Morc3*) and even genes (e.g. *eIF3h, Hbb*) without a clear link to epigenetic processes (full gene list reviewed here [125]). Therefore, the interpretation of the results should be carried out in light of confounding factors inherent to the screen design such as transgene integration site, tissue-specific phenotypic read-out (potential identification of genes affecting hematopoiesis in this case), the introduction of a foreign DNA sequence which potentially triggers similar host genome responses as retrotransposons or integration-prone viruses, genetic background (i.e. mouse strain) in which the screen was conducted, parent-of-origin effects (screening progenies of ENU-treated males) or the actual structure of the transgene (tandem repeat in this case).

Indeed, several commonalities between retrotransposon and transgene silencing were pointed out previously [129,130]. In line with that, several *MommeD* alleles were found to also modulate the Agouti viable yellow (A^{vy}) locus, in which a spontaneous insertion of an intracisternal A particle (IAP), belonging to a Class II endogenous retrovirus (ERV) family, was shown to modulate the expression of *in cis Agouti* gene responsible for, among others, coat colour [131]. *Agouti* is normally expressed only transiently from a hair cycle-specific promoter and is responsible for the deposition of yellow and black pigment during mouse hair growth [132]. The inserted IAP creates a cryptic promoter that drives continuous expression of *Agouti* leading to a completely yellow coat. However, partial or full silencing of this IAP by e.g. DNA methylation leads to mottled or wild-type-like brown fur color. Specifically, S*mchd1*^{MD1}, *Dnmt1*^{MD2}, *Trim28*^{MD9} and *Setdb1*^{MD13} alleles resulted in a shift to a yellow fur (i.e. failure to repress the IAP), while *Smarca5*^{MD4}, *Rlf*^{MD8} and *Wiz*^{MD30} alleles resulted in a shift to a brown fur [126,127,133]. Interestingly, the resulting phenotypic shifts in the coat color due to these alleles were concordant with their effect on GFP transgene expression suggesting that they play the same role, either repressing or activating, at these two loci. Furthermore, the phenotypic outcome of the coat color and thus *Agouti* gene expression reversely correlated with the DNA methylation status at the 5' long terminal repeat (5' LTR) of the inserted IAP [134]. Similar observation was made also for the methylation status and expression of the GFP transgene and when combined with concrete *MommeD* alleles, namely S*mchd1*^{MD1}, *Rlf*^{MD8}, *Dnmt3b*^{MD14}, *Dnmt1*^{MD32} and *Nrf1*^{MD46} [127,135–137]. However, some *MommeD* alleles such as *Hdac1*^{MD5}, *Baz1b*^{MD10}, *Wiz*^{MD30} and *Rif1*^{MD18} showed no changes in DNA methylation of the GFP transgene and yet showed changes in expression suggesting that these factors are involved in layers of epigenetic regulation unrelated to DNA methylation [127,136].

Follow-up studies employing reverse genetics approaches uncovered that genes underlying *Momme* alleles are involved in epigenetic regulation of diverse endogenous loci including different types of repeats. For instance, Dnmt3b seems to be particularly specialized in the establishment of DNA methylation at pericentromeric [138] and subtelomeric repetitive regions [139] and is also responsible for silencing genes on the inactive X chromosome [140]. Similarly, Suv39h1/Suv39h2 mediate deposition of H3K9me2/me3 at pericentromeric [63,141] and subtelomeric repeats [142]. In contrast, a trio of Mommes (Morc3, Trim28 and Setdb1) is involved in the repression of IAP elements [143–145].

Mutations in *Momme* genes have also been linked to diverse human diseases and syndromes. The most worthy to mention in the context of this thesis are mutations in two genes, *DNMT3B* and *SMCHD1*, as their heterozygous mutations are associated with Facioscapulohumeral muscular dystrophy [146,147]. In addition, biallelic mutations in *DNMT3B* cause the rare Immunodeficiency, Centromeric region instability and Facial anomalies type 1 (ICF1) syndrome [148]. Similarly to *DNMT3B*, mutations in *SMCHD1* can also yield a pleiotropic phenotypic outcome since they are also causative of Bosma Arhinia Microphthalmia Syndrome (BAMS), a very rare condition, with less than 50 patients being reported, characterized by nasal, ocular and reproductive defects [149].

Facioscapulohumeral muscular dystrophy

The FSHD locus

The road to elucidating the root cause of the disease took over 100 years since its first description as a distinct disease entity as FSH type muscular dystrophy by the French physicians Louis Landouzy and Joseph Dejerine in 1884.[5] Studies in the early 90s helped to narrow down the search for the FSHD locus by linking the disease to an *EcoRI* genomic fragment which was polymorphic in length and detected by a DNA probe (p13E-11) mapping to 4q35 [150–152]. Specifically, *EcoRI* fragments usually larger than 28 kb were detected in non-affected individuals, while shorter fragments between 14 – 28 kb co-segregated with FSHD [151]. Interestingly, even after 30 years, a slightly modified approach is being used to this day for routine FSHD diagnostics (Figure 1A) [153]. Soon after, it was shown that the locus in question contains a tandemly repeated sequence dubbed D4Z4[6] which consists of copies of a 3.3 kb repeat unit defined by a *KpnI* restriction site (Figure 1A). Similar repeat sequences map to other locations in the human genome [154,155] with a highly homologous tandem repeat present at 10q26 that can vary between 1-100 units in the population [156,157]. However, the reason why shortening of this particular repeat only on chromosome 4 causes FSHD remained elusive for a long time. The initial hypothesis to explain the chromosome 4 specificity of the disease was inspired by the PEV mechanism and proposed that longer D4Z4 repeats tend to adopt a more heterochromatic structure which would spread *in cis.* In FSHD, due to the reduced D4Z4 copy-number, this heterochromatinization would be partially lost leading to inappropriate expression of nearby gene(s) [154].

5 Initially, FSHD was referred to as Landouzy-Dejerine muscular dystrophy, however, some disputes were raised over who should be acknowledged for the priority of describing this disease as a separate clinical entity [322] as the very first description of the disease was done by the French neurologist Duchenne de Boulogne. The peculiar pattern of muscle weakness first affecting distal leg muscles while skipping proximal leg muscles was first recognized by German neurologist Wilhelm Heinrich Erb. However, it was Landouzy and Dejerine who 'absorbed' prior clinical descriptions of Duchenne and Erb together with observations from their casuistry into one FSH type of muscular dystrophy.

6 The name 'D4Z4' is derived from a nomenclature system which was used for DNA regions of unknown significance during the human genome project: D stands for DNA, 4 stands for chromosome 4, Z indicates a repetitive sequence and 4 is a assigned serial number based on the submission order. Hence, the homologous repeat on chromosome 10 cannot be truly termed D4Z4 and was unfortunately never assigned a D10Z serial number.

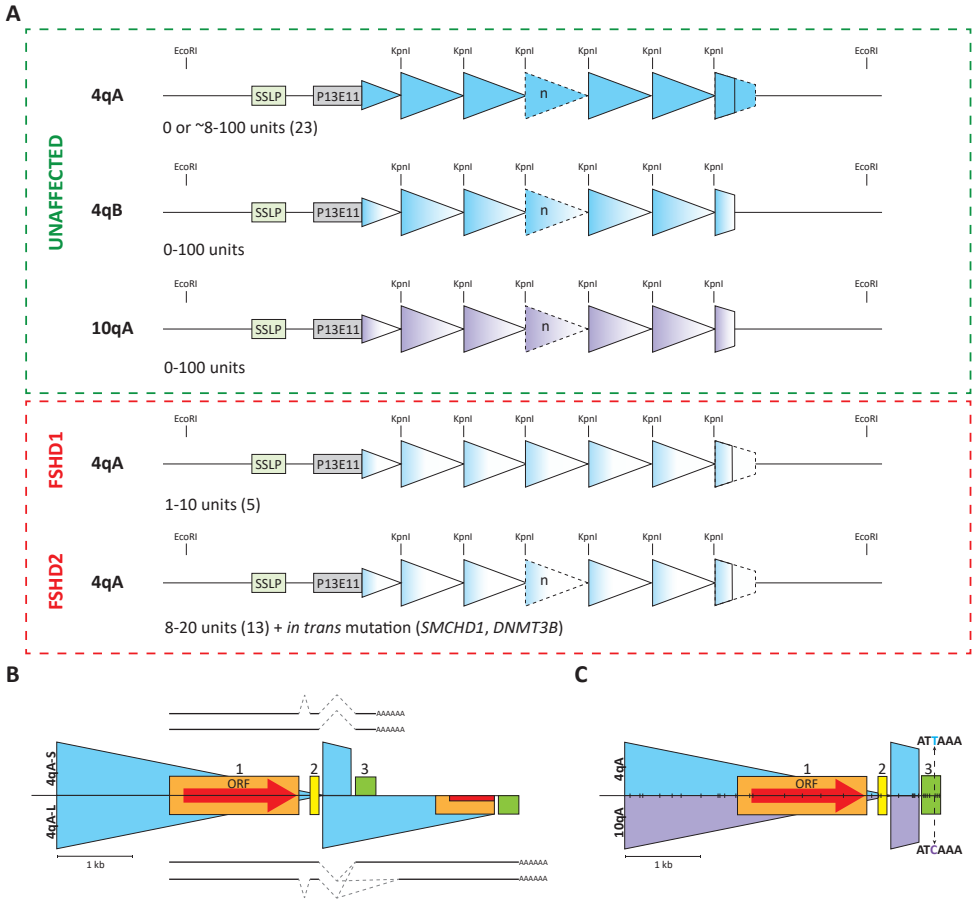


Figure 1. Genetics and epigenetics of D4Z4 in FSHD. A) Schematic representation of homologous D4Z4 repeats present on the long arms of chromosomes 4 and 10 as illustrated by blue (4q) and violet (10q) triangles organized in tandem. Each triangle represents one repeat unit as defined by a *KpnI* restriction site and the first and the last repeat units are incomplete. Proximal to the repeat are two sequence elements that are utilized for the assessment of D4Z4 haplotype (SSLP) and for the determination of copy-number by Southern blot (P13E11). The 4qA/B typing is performed with Southern blotting using probes that hybridize further downstream of the distal *EcoRI* site and are thus not depicted. Unaffected individuals either carry variably sized 4qB D4Z4 repeats, whose epigenetic state is not relevant for FSHD or D4Z4 repeats on the 4qA background whose length is sufficient for proper epigenetic silencing of the repeat (usually more than 8 units). Similarly to 4qB D4Z4 repeats, any copy-number and its associated epigenetic state of 10qA D4Z4 repeats is irrelevant for FSHD pathogenesis. Epigenetic dysregulation of 4qA D4Z4 repeats is caused either by its reduction in copy-number (FSHD1; 1-10 units) or by co-inheritance of intermediately-sized (8-20 units) repeat together with a mutation in (at least) one of its *trans* modifiers (FSHD2). Numbers in the brackets next to the designated repeat ranges refer to a median size of the repeat based on their prevalence in the European population. The color gradient of triangles represents varying levels of 4q/10q D4Z4 epigenetic repression found in healthy and FSHD individuals (the lighter the color the lower the DNA methylation levels and thus repression). **B)** Specifically the 4qA161 D4Z4 haplotype, which is the most frequent haplotype in FSHD individuals, can end in two different forms (4qA-S or 4qA-L) depending on the break-point in the most distal incomplete repeat unit. The two forms give rise to different *DUX4* mRNA isoforms differing in their 3'UTR. The *DUX4* mRNA isoforms are further diversified by the optional splicing of exon 2. Three *DUX4* exons are represented by brackets of different colors (orange, yellow and green) with the DUX4 ORF being fully contained within exon

1 is represented by the thick red arrow. **C)** Comparison of the terminal 4qA (blue) and 10qA (violet) repeat unit depicted as a mirror image. The position of the *DUX4* PAS in exon 3 is depicted by a dashed arrow line highlighting the 4qA/10qA SNP in the *DUX4* PAS. Other distinguishing SNPs between 4qA and 10qA D4Z4 are marked on the line that separates 4q from 10q.

In line with this hypothesis, it was shown that D4Z4 repeat contractions leads to DNA hypomethylation of this locus [158] which is accompanied by reduced levels of H3K9me3 [159] and thus possibly affecting the regulation of several candidate FSHD genes proximal to 4q D4Z4 repeat, including *ANT1, FRG1*, *TUBB4Q* and *FRG2* [160–163]. However, while one study documented upregulation of some candidate genes in FSHD muscle biopsies [160], other studies reported no changes in mRNA expression of these genes between FSHD and control cases arguing against the PEV hypothesis [164–167]. Furthermore, it was shown that at least one D4Z4 unit is required for disease development [168] suggesting that FSHD is tightly associated with the D4Z4 repeat itself rather than its surrounding chromosomal region. Indeed, every D4Z4 unit was found to contain an open reading frame encoding for a putative double homeobox protein termed *DUX4* [169–171]. Thus, another hypothesis was put forward suggesting that the epigenetic de-repression of contracted D4Z4 repeats leads to the expression of this repeat-encoded *DUX4* gene [171]. But it was not until 2010 that a unifying mechanism for FSHD-associated *DUX4* expression was presented, which confirmed the latter hypothesis (Figure 1B) [172]. Furthermore, the possible involvement of other candidate genes on chromosome 4 was challenged by describing FSHD individuals having atypical D4Z4 rearrangements. This includes cases with large proximal deletions occurring *in cis* to the contracted 4q D4Z4 repeat sometimes encompassing *FRG2* and *TUBB4Q* [173,174], as well as FSHD cases with interchromosomal rearrangements between 4q35 and 10q26 resulting in a hybrid, contracted D4Z4 repeat at 10q26 and leading to a physical separation of the contracted D4Z4 repeat partially of 4q origin and other 4q FSHD candidate genes [175].

Nowadays, two genetically distinct forms of FSHD are recognized, FSHD1 (OMIM #158900) and FSHD2 (OMIM #158901), however, both involve epigenetic de-repression of the 4q D4Z4 repeat associated with *DUX4* expression in skeletal muscle. They do, however, arise by distinct genetic mechanisms (Figure 1A). While in FSHD1 it is the contraction of the D4Z4 repeat that causes loss of its heterochromatinization, in FSHD2, it is mutation(s) *in trans* in genes which play a role in establishing or maintaining the heterochromatic state of D4Z4 [146,147,176]. Another notable difference is that in the latter case, the chromatin state of the 10q26 D4Z4 repeat is also affected whereas in FSHD1 the chromatin changes are constricted to the contracted 4q allele only [158,177].

Clinical presentation

FSHD is regarded one of the more common muscular dystrophies in adults with an estimated prevalence ranging between 0.8 and 4.6 per 100,000 [178]. From a clinical perspective, FSHD1 and FSHD2 cases are indistinguishable [179,180]. Age at onset as well as clinical severity varies extensively from patient to patient with one-fifth of individuals with an FSHD-sized D4Z4

repeat within FSHD1 families remaining asymptomatic [181]. However, this number might be lower as an important factor for disease presentation is age. While FSHD affects both sexes, some FSHD1 family studies reported that females are more likely to be less severely affected or asymptomatic than males despite carrying an identical D4Z4 repeat array [182–184]. Evidence for such sex bias in FSHD2 families is weaker, but this can be also due to the relatively small sample sizes as compared to studies in FSHD1 families [179,180]. In "classical" FSHD cases, the first symptoms become apparent in the second decade. Being a slowly progressive disease, individuals are often diagnosed relatively late in life with a median age at diagnosis of around 40 [185]. Although FSHD patients typically have a normal life expectancy, their fitness decreases over time with almost one-fourth of cases requiring a wheelchair by the age of 50 [186,187]. In addition, there is an infantile form of the disease, representing around 4% of all FSHD cases, with more severe symptoms and faster progression [188,189].

The typical clinical presentation of FSHD includes early involvement of the muscles of the face, shoulder girdle, and upper arms, often in an asymmetric manner. As the disease progresses, lower extremities can also become affected, starting with the distal muscles and later involving more proximal leg muscles. Apart from muscle involvement, extramuscular manifestations have been reported in FSHD. These include ophthalmological abnormalities [190,191], high-frequency hearing loss [192–194] and CNS abnormalities like epilepsy and mental retardation. These extramuscular manifestations are more prominent in the more severe infant onset form of the disease [189,195,196].

Cis modifiers in FSHD

FSHD1 is due to a contraction of at least one 4q D4Z4 repeat to a size of 1 – 10 units. However, for a 4q D4Z4 contraction to result in FSHD, it needs to occur on a specific 4q subtelomeric genetic background [197,198]. Based on sequence variations immediately distal to D4Z4, 4q subtelomeres were initially subgrouped into two main allelic variants, 4qA and 4qB, which are equally common in the European population [199]. Interestingly, contractions of D4Z4 on 4qB alleles have not been observed to cause FSHD suggesting that some 4qA-specific sequences underlie 4qA pathogenicity or that 4qB alleles contain protective genetic elements [197,200]. The most noteworthy difference between 4qA and 4qB alleles is the presence of a 260 bp sequence immediately distal to D4Z4 on the 4qA background, termed pLAM which is followed by a β-satellite repeat [200]. Furthermore, the 10q subtelomere shows a high degree of sequence homology (98%) to 4qA [199] and thus is referred to as 10qA. Yet, D4Z4 contractions at 10qA are not pathogenic [156]. In addition, even though all FSHD D4Z4 alleles are of the 4qA type, not all contracted 4qA D4Z4 alleles result in FSHD [198]. A worldwide population study further revealed nine subtelomeric 4qA haplotypes based on several sequence polymorphisms found within and flanking the repeat [201]. One of the main sequence features defining the haplotype is a simple sequence length polymorphism (SSLP) located approximately 3 kb proximal to both 4q and 10q D4Z4 repeats (Figure 1A). All haplotypes are thus defined by their chromosomal location (4 or 10), distal variant (A or B) and SSLP (between 157 and 182 bp). Out of nine defined 4qA haplotypes, only three (4A159, 4A161

and 4A168) have been indisputably associated with FSHD [172], while the classification of 4A166 as FSHD-permissive haplotype remains inconclusive due to contradicting findings [198,202,203]. The most predominant haplotype found in FSHD in Europe is unsurprisingly 4A161 [198] as it is also the most frequent FSHD-permissive haplotype found in the control European population [201]. In addition, this haplotype shows another degree of variability in its distal end. The most distal unit in 4qA haplotypes is incomplete and usually formed by a proximal 0.33 kb of the D4Z4 unit. However, the 4A161 haplotype can instead of this short (S) end, also terminate with a longer incomplete unit of 1.6 kb which is then referred to as the long (L) variant (Figure 1B) [172,204]. Nevertheless, in regards to FSHD, contractions of either 4A161 variant (4A161S or 4A161L) are disease-causing [172,204]. A near-perfect explanation for the 4qA linkage of FSHD came with a seminal study providing a functional explanation for the pathogenicity of certain 4qA haplotypes [172]. Two earlier studies already showed that it is almost exclusively the most distal full-length D4Z4 unit that can express fully processed and stabilized *DUX4* mRNA by using a polyadenylation signal (PAS) in the pLAM region (Figure 1B) [165,205]. Extending on that, Lemmers et al. showed that the 10qA pLAM region contains a SNP in the sequence corresponding to the *DUX4* PAS sequence found on disease-permissive 4qA haplotypes (4qA: AT**T**AAA -> 10qA: AT**C**AAA) corrupting its functionality (Figure 1C) [172]. The only haplotype that currently remains unresolved is 4A166 as its disease association remains unclear. While it does contain a functional 4qA *DUX4* PAS sequence, the majority of other SNPs in its pLAM region are more 10qA-like [172]. Therefore, more exhaustive population studies, as well as functional dissection of the effects of 4qA/10qA sequence polymorphisms on *DUX4* expression, are required to further fine-tune our understanding of the genetic predisposition to FSHD.

One of the most important *cis* modifiers in FSHD is the D4Z4 copy number itself as even inheritance of a contracted allele on a disease-permissive haplotype is not 100% predictive of disease penetrance. This is a key aspect of FSHD, i.e. that the phenotypic outcome is on a continuous quantitative scale rather than categorized by simple binary qualitative groups (non-affected vs affected) as the clinical severity is often inversely correlated with D4Z4 copy-number [206–210]. Individuals with shorter alleles typically have an early onset whereas carriers of FSHD alleles in the upper size range (7 – 10 units) present with milder symptoms or even remain life-long asymptomatic [181,208]. The latter cases make prenatal diagnosis and genetic counselling challenging as it is associated with high levels of uncertainty [211]. Furthermore, differences have been observed in the D4Z4 length distribution in non-affected as well as in FSHD1 cohorts of ethnically different populations. Particularly, the median size of 4q D4Z4 repeats in unaffected Asian populations is smaller than in the Caucasian population [212] and such difference in distribution is also observed for the size of the contracted allele in FSHD1 cohorts from Asia and Europe (median of 3 – 4 units vs 5 – 6 units, respectively) [213–215]. Therefore, it seems that Asian populations are less permissive to FSHD. The factors behind this reduced permissiveness, either of environmental or genetic origin or both, remain to be elucidated but could be instrumental to our understanding of *DUX4* expression regulation. However, it should be noted that we only operate with the assumption that shorter D4Z4 alleles yield higher DUX4 levels since no larger-scale correlative studies have been conducted regarding a relationship between D4Z4 copy-number and *DUX4* expression.

Other *cis* modifiers proximal to D4Z4 have been proposed to be involved in FSHD by influencing the muscle-specific phenotype [216] and D4Z4 de-repression [217]. For example, two *DUX4* myogenic enhancers (DME1 and DME2) have been described upstream of D4Z4 which by looping to the *DUX4* promoter are supposed to enhance its expression specifically in skeletal muscle cells but not in skin fibroblasts [216]. Another D4Z4 proximal genetic element was found to give rise to a long non-coding RNA (DBE-T) which was shown to be upregulated in FSHD and to be responsible for the recruitment of the Trithorax group protein ASH1L to the D4Z4 repeat causing chromatin remodeling with subsequent *DUX4* de-repression [217]. However, the identification of individuals presenting with FSHD carrying proximally extended D4Z4 deletions encompassing aforementioned *cis* sequence elements challenged their relevance for FSHD pathology [174,218–220].

Trans modifiers in FSHD

The first indication about the existence of possible *trans* modifiers in FSHD came with the recognition that around 5% of FSHD cases do not carry a contracted D4Z4 allele and yet show DNA hypomethylation of the repeat [158,177,179,221]. These cases were classified as contraction-independent type 2 FSHD (FSHD2). In these FSHD2 individuals, both 4q and 10q D4Z4 repeats were found to be hypomethylated as opposed to only the contracted repeat in FSHD1, suggestive of the involvement of a *trans* factor affecting D4Z4 methylation [146,158]. The introduction of whole-exome sequencing (WES) into clinical genetics practice accelerated the identification of heterozygous mutations in the *SMCHD1* gene, which co-segregated with D4Z4 hypomethylation and, if combined with 4qA allele, resulted in FSHD [146]. The spectrum of *SMCHD1* mutations identified in FSHD2 include nonsense, missense, splicing-affecting mutations and even larger genomic deletions encompassing the entire *SMCHD1* locus resulting in SMCHD1 hemizygosity (detailed overview of SMCHD1 mutations is reviewed here [222]). Therefore, the current consensus is that D4Z4 hypomethylation in FSHD2 is due to reduced amounts of functional SMCHD1 protein. In addition, the nature of SMCHD1 mutations correlates with residual DNA methylation level at D4Z4. Specifically, heterozygous *SMCHD1* mutations which preserve the open reading frame (usually missense mutations) show more pronounced D4Z4 hypomethylation and thus seem to be more deleterious than heterozygous *SMCHD1* mutations which disrupt the SMCHD1 open reading frame and result in lower SMCHD1 protein levels [223]. One possible explanation for this observation is that SMCHD1 forms homodimers [224,225] and thus haploinsufficiency of SMCHD1 would reduce the number of functional WT SMCHD1 homodimers to 50% as compared to the WT situation, while the dominant negative effect of missense mutations would lead to only 25% of WT functional SMCHD1 homodimers if we assume that the mutant SMCHD1 monomer can form heterodimers with WT SMCHD1 monomer (25% WT:WT, 50% WT:MUT, 25% MUT:MUT). Furthermore, missense mutations positioned at the N-terminus of the protein were shown to have a greater effect on D4Z4 methylation than those at the C-terminus [223]. Similarly to the previously observed rough inverse correlation between the length of contracted D4Z4 repeat, its methylation and clinical severity in FSHD1 individuals, a significant correlation was also found for the length of the shortest 4qA D4Z4 allele and its DNA methylation in FSHD2

individuals [223], suggesting that also in FSHD2 cases the repeat length plays a modifying role for its epigenetic state. This is further supported by the observation that the median size of a 4qA D4Z4 repeat in European FSHD2 individuals was found to be shorter (13 units) than the median size in the control European population (23 units) [223], which would suggest that SMCHD1 mutation carriers with longer permissive alleles do not develop FSHD or its manifestation is very mild [226], creating a reservoir of *SMCHD1* mutations in the population. At that time, an enigmatic exception was a group of FSHD2 individuals with longer 4qA D4Z4 repeats (up to 70 units). However, closer genetic examination revealed that the majority of these cases have a duplication allele consisting of the earlier diagnosed longer repeat array followed by a smaller FSHD-sized repeat array duplication that is likely permissive to *DUX4* expression [227]. Mutations in SMCHD1 have been also reported to modify the disease penetrance as well as severity in FSHD1 cases [180,226,228–230]. Interestingly, only individuals with upper-sized D4Z4 repeats (7 – 10 units) in combination with an *SMCHD1* mutation were described, which prompts the question if a combination of a shorter D4Z4 repeat with *SMCHD1* mutation is under negative selection pressure and incompatible with life, or that it is the sheer rareness of this combination that has prevented its reporting thus far. Alternatively, such bias in the findings could be due to the existing FSHD diagnostics practice, when cases suspected of FSHD are first undergoing D4Z4 sizing and if a contracted 4qA allele is identified no further screening for *SMCHD1* mutations is undertaken. On the other hand, enough comparative methylation studies between FSHD1 and FSHD2 were reported [223,231,232] that would potentially reveal D4Z4 hypomethylation outliers in FSHD1 cohorts sparking the motivation for identifying possible *in trans* mutations in these individuals.

Even almost 10 years after the first description of the association of *SMCHD1* mutations with FSHD2, we still do not have a clear mechanistic explanation of how germline SMCHD1 loss-of-function relates to the observed D4Z4 hypomethylation in somatic cells. SMCHD1 is expressed in somatic cells where it binds to D4Z4 and its binding is reduced in somatic cells derived from FSHD2 individuals [146]. Furthermore, depletion of SMCHD1 either in FSHD1 or FSHD2 skeletal muscle cells leads to further *DUX4* transcriptional de-repression suggesting that it aids in D4Z4 silencing also in somatic cells with an already compromised D4Z4 chromatin state [226,233]. In addition, increasing SMCHD1 levels in FSHD1 and FSHD2 muscle cells, either by its overexpression or by its mutation correction in the case of FSHD2, was shown to result in significant *DUX4* downregulation [230,233]. Although complete transcriptional repression of *DUX4* was not achieved, low levels of DUX4 have been detected also in unaffected relatives of FSHD subjects [234], thus absolute *DUX4* somatic silencing might not be necessary to achieve clinical benefit. Such observations inspired a discussion about the possibility of modulating SMCHD1 levels as a general therapeutic strategy for FSHD.

SMCHD1 is encoded by 48 exons giving rise to a 2005 aa-long protein in humans whereas the mouse ortholog of SMCHD1 is 2 aa longer. It contains two main functional domains: an N-terminal ATPase and a C-terminal hinge domain which are connected by a flexible linker [235]. Both the hinge domain as well as the ATPase domain are required for SMCHD1 homodimerization [224,236,237]. The hinge domain was further shown to interact with nucleic

acids [236,238]. More recently, two extra domains flanking the ATPase module were characterized, namely the N-terminal ubiquitin-like (UBL) and transducer domains, which aid in ATPase dimer stabilization during ATP hydrolysis [225,237]. Mouse Smchd1 was initially identified, as mentioned before, in the dominant screen for modifiers of murine metastable epialleles and follow up studies showed that its homozygous loss-of-function results in female-specific embryonic lethality due to a failure in X inactivation [235]. In contrast to the active X (Xa), which is in its higher-order structure more similar to autosomes by being partitioned into smaller topologically associated domains (TADs), the inactive X (Xi) is folded into two megadomains with limited short-range intra-chromosomal interactions both in mouse and humans [239–241]. Smchd1 is a key factor in this folding process as its loss results in increased short-range interactions over the Xi due to enhanced CCCTC-binding factor (CTCF) binding leading to Xi decompaction [242–244]. A similar gain of Ctcf has been observed also at clustered protocadherins and Hox gene cluster in the absence of Smchd1 [238,242]. Both clusters were already known to be transcriptionally sensitive to the loss of Smchd1 [238,245,246]. In addition, Smchd1 has been shown to regulate the expression of monoallelically expressed genes such as selected genes within the *Snrpn* cluster [238,245,246]. Furthermore, Smchd1 was recently shown to act as a maternal effect gene in the mouse, when the maternal *Smchd1* allele is the only source for Smchd1 production until at least the 32-cell stage and is required for the imprinted expression of 10 genes [247]. However, whether human SMCHD1 expression is regulated similarly during human pre-implantation development remains to be elucidated. But even if so, it might be of little relevance for SMCHD1-mediated D4Z4 epigenetic regulation as both maternal and paternal transmission of an *SMCHD1* mutation has been documented in FSHD2 families with no apparent methylation or clinical differences between the sexes [146].

Nowadays, it is estimated that at least 85% of FSHD2 cases are explained by mutations in SMCHD1 [146,223]. This number is likely higher as mutations in cases suspected of FSHD2 are typically identified by WES or *SMCHD1* exon sequencing and thus potential deep intronic *SMCHD1* mutations go unnoticed. Indeed, one such FSHD2 family has been reported recently [230]. Nevertheless, further studies into other *trans* modifiers identified two families in which a heterozygous mutation in *DNMT3B* was co-segregating with D4Z4 hypomethylation and was shown to modify the disease penetrance in family members carrying a relatively short permissive D4Z4 repeat [147]. Identifying *DNMT3B* mutations was not surprising as recessive mutations in *DNMT3B* were previously shown to cause ICF1 syndrome, in which the D4Z4 repeat is also hypomethylated [248]. Interestingly, despite SMCHD1 and DNMT3B both converging at the epigenetic regulation of D4Z4, other repeats which are hypomethylated in ICF1 individuals, such as pericentromeric satellite repeat types II and III and the NBL2 macrosatellite repeat, are not hypomethylated in FSHD2 individuals with *SMCHD1* mutations [221] suggesting that these two factors do not always co-regulate the same genomic regions or alternatively, that aforementioned repeats are less sensitive to SMCHD1 than to DNMT3B dysfunction.

The epigenetic makeup of D4Z4 in somatic cells consists of high levels of DNA methylation and H3K9me3 which both ensure *DUX4* repression as treating cells either with 5-aza-2'-

deoxycytidine (Aza), a deoxycytidine analogue which cannot be methylated by DNMTs, or chaetocin, a non-specific inhibitor of histone methyltransferases, results in *DUX4* transcriptional de-repression [159,217,249,250]. Complementary experiments to reduce these marks by lowering the protein levels of both DNMT1 and DNMT3B or SUV39H1 confirmed their importance in somatic D4Z4 silencing [250,251]. Furthermore, both DNA methylation and H3K9me3 are reduced at D4Z4 in FSHD1 and FSHD2 not only in skeletal muscle cells but also in cells of other tissues derived from different germ layers [159,202,232,251,252], suggesting that either the establishment of these repressive marks was impaired before the multi-lineage differentiation or the mechanism of their maintenance is impaired in all tissues. Apart from this, additional changes in histone marks have been documented such as increase in H3K4me2/3 in both FSHD1 and FSHD2 [253,254] and a specific increase in H3K27me3 in FSHD2 individuals with SMCHD1 mutations [233]. In addition, several studies identified other factors of D4Z4 chromatin in somatic cells which contribute to the repression of this locus. First, HP1$_\gamma$ (CBX3) and the cohesin complex were found to be associated with 4q and 10q D4Z4 repeats in control somatic cells and their recruitment was shown to be dependent on H3K9me3 [159]. A follow-up study also explored the heterochromatin state of D4Z4-like sequences which are present at other chromosomes, mainly at acrocentric chromosomes [250]. Interestingly, neither DNA methylation, H3K9me3, HP1$_\gamma$ nor the cohesin complex was affected at those regions in both FSHD1 and FSHD2 cells with *SMCHD1* mutation. This prompts the question of why the function of SMCHD1 is restricted to 4q and 10q D4Z4 repeats and to what degree D4Z4-like sequences on other chromosomes are different from 4q/10q D4Z4 repeats in their chromatin regulation [250]. Limited data is available on these sequences but one noticeable difference is that different from 4q and 10q, these repeat sequences do not seem to form homogeneous tandem repeat arrays [155]. This also raises another concern regarding chromatin studies that employ PCR amplification to investigate the association of specific chromatin factors with D4Z4 as our findings and conclusions about FSHD-relevant chromatin changes are only as good as the specificity of the primers or probes we use.

Recently, an unbiased proteomic study identified 261 proteins as being enriched at D4Z4 in control myoblasts, including components of the NuRD and CAF1 complexes and interestingly also several *Momme* factors, namely PBRM1, RIF1, SMARCA4, SMARCA5, UHRF1, HDAC1, SETDB1 and TRIM28 [255]. It remains to be investigated if all of these protein factors act in a parallel or redundant fashion, if and how they contribute to disease penetrance, and if any of these repressive components can be employed for future therapeutic strategies aiming at re-repression of D4Z4 in FSHD skeletal muscle cells.

Lastly, as mentioned earlier, the epigenetic changes to D4Z4 are not specific for skeletal muscles of FSHD individuals but are also present in other somatic tissues [159,202,232,251,252] Therefore, the apparent predominant muscle phenotype in FSHD raises the question why other tissues are not affected. Either other tissues are somehow resistant to DUX4 toxicity or more likely, they do not even express *DUX4.* We know that for example cultured fibroblasts derived from skin biopsies of FSHD individuals do not express *DUX4* at all and that *DUX4*

expression can only be detected after their forced transdifferentiation into myotubes [147]. In addition, neither *DUX4* nor its transcriptional signature was detected in the RNA-seq study of whole blood from FSHD individuals [256], although EBV-transformed peripheral blood leucocytes derived from FSHD individuals do recapitulate both D4Z4 epigenetic as well as *DUX4* transcriptional changes of FSHD myoblasts [202,254,257]. Even more peculiar are the inter-muscular differences as some muscle groups seem to be more prone to *DUX4* expression than others, which could explain their differential involvement in FSHD [258,259]. Furthermore, as *DUX4* expression increases during myogenic differentiation [233], it seems that the epigenetic changes at D4Z4 only create an environment permissive for *DUX4* expression and that muscle-specific factors or intracellular changes during myogenesis or EBV-transformation are required to initiate *DUX4* transcription. Previously, it has been shown that protein levels of SMCHD1 decrease during myogenic differentiation [233] and therefore, one could hypothesize that reduced availability of some D4Z4 repressors might contribute to this muscle-restricted misexpression of *DUX4.*

Conservation of *DUX4* and consequences of its expression in skeletal muscle

DUX4 belongs to the *DUX* gene family, which includes among others also the intronless *Dux* gene present in the mouse genome [260]. Both *DUX4* and *Dux* are hypothesized to have arisen independently by a retrotransposition-related expansion of an ancestral *DUXC* gene and are organized in a tandem array, although not at a syntenic location. Furthermore, the single repeat unit of the *Dux*-forming macrosatellite is longer than that of D4Z4 (4.9 kb vs 3.3 kb) [261]. DUX4, as well as Dux, contain two highly homologous N-terminal homeodomains as well as a conserved C-terminal transcriptional transactivation domain [262]. Only recently it has been shown that they are indeed functional homologs by regulating the zygotic genome activation (ZGA), a process after fertilization during which the transcription of newly combined genetic material starts for the very first time [107,108,263]. Expression of *DUX4* mRNA was shown to peak during the 4-cell cleavage stage, whereas *Dux* mRNA expression peaks already at the 2-cell stage, both corresponding to their species-specific ZGA timepoints [107,108]. Both Dux and DUX4 activate the transcription of ZGA genes by directly binding to their promoters through their homeodomains [263]. Furthermore, both proteins bind also to a specific family of retrotransposons (MERVL in mice and HERVL in humans), which serve as alternative promoters of some cleavage-specific genes during ZGA [108,263]. However, how *Dux*/*DUX4* expression itself is so swiftly regulated during this short time window is still poorly understood. It is also not known whether a failure in *DUX4* silencing in FSHD individuals begins at this point (although it should be noted that it has not been established if the cleavage-specific *DUX4* transcripts are specifically of only 4q D4Z4 origin).

Interestingly, certain culturing conditions allow mESCs to fluctuate between pluripotent (ICM-like state of blastocyst) and totipotent state (2-cell blastomere-like cleavage stage) and at any given moment around 1% of the mESC population is in this 2-cell-like stage [106]. These 2C-like cells recapitulate many attributes of the 2-cell stage blastomeres including their transcriptome which is characteristic of the ZGA phase [106,264], chromatin accessibility

landscape [108], high core histone mobility [265] and the capacity to contribute to extra-embryonic tissues [106]. The conversion of mESCs to 2C-like cells is regulated both by a variety of chromatin factors [264,266–275]. Furthermore, induction of 2C-like cells was shown to strongly depend on Dux as ectopic expression of *Dux* forces mESCs into 2-cell like cells [108]. In line with this, *Dux* knock-out prevents mESCs from their conversion to 2-cell like cells [107]. However, follow up studies challenged the notion of Dux being an essential driver of ZGA, since *Dux* zygotic knock-out embryos can give rise to viable pups, albeit with decreased developmental potential due to delayed ZGA onset [276–278]. Therefore, Dux seems to help in synchronizing the ZGA, but probably other yet unidentified factors in addition to Dux are involved in the onset and propagation of the ZGA process *in vivo*. In contrast, efficient silencing of *Dux* past the 2-cell stage seems to be of bigger importance for proper embryonic development as its sustained expression impedes the 2-cell exit and causes embryonic arrest [272,279]. The emerging recent model suggests that *Dux* repression is achieved by tethering its genomic locus to the perinucleolar heterochromatin space by the LINE1/Nucleolin/Trim28 complex both in mESCs and early embryos [272,280,281]. Nucleoli are membrane-less nuclear organelles, whose boundaries are thought to be defined by liquid-liquid phase separation and are a place for rRNA and ribosome biogenesis (reviewed here Lafontaine et al., 2020). Both 2C-like cells and 2-cell blastomeres possess yet immature more compact nucleoli sometimes referred to as nucleolar precursor bodies (NPBs) which exhibit low rRNA transcriptional output [282]. Following fertilization, the rRNA levels sharply increase from the 2-cell stage onwards to the blastocyst stage to cope with the embryonic need for a sufficient amount of translational apparatuses [280]. This rRNA transcriptional change is associated with nucleolar maturation and with the formation of perinucleolar heterochromatin. Thus, it seems that the embryonic need for increased translational output and the termination of the ZGA phase were naturally co-opted into one regulatory mechanism during early genome spatial reorganization when activation of rRNA synthesis shuts down expression of *Dux* for cells to continue into the next cleavage stages. This also explains a prior counterintuitive observation that the LINE1/Nucleolin/Trim28 complex while positively regulating rRNA expression negatively regulates expression of *Dux* [272]. It remains to be investigated if a similar mechanism also operates in *DUX4* silencing during human embryonic development. Interestingly, other D4Z4-like sequences are present on the short arms of acrocentric chromosomes [155,250] which are responsible for the nucleolar organization (reviewed here McStay, 2016) and similarly, also the 4q D4Z4 repeat has been observed to preferentially localize either to the nuclear or nucleolar periphery in somatic cells [284–286]. Despite that, the nuclear localization of contracted 4q D4Z4 was not changed in cells from FSHD1 individuals which could otherwise explain the sporadic transcriptional activation of *DUX4*.

Overexpression of DUX4 in cultured myoblasts elicits a transcriptional response similar to what was identified during human ZGA, including upregulation of specific retroelements and cleavage-specific genes, which are also misexpressed in FSHD cultured muscle cells as well as in FSHD biopsies [287–290]. Endogenous *DUX4* expression is a rather rare event in FSHD 2D muscle cell cultures, with only around 1:200-1000 of nuclei expressing *DUX4* at any

given moment, depending on the differentiation stage, culture conditions and donor [291–294]. However, since mononuclear muscle precursor cells fuse during myogenic differentiation to form multinucleated myofibers in which they eventually share their cytoplasmic space, even one nucleus expressing *DUX4* can "infect" its neighboring nuclei with DUX4 protein upon its translation in the cytoplasm. This can be visualized by staining for DUX4 protein in differentiated muscle cells, typically creating a DUX4 staining gradient across clustered nuclei that is getting weaker with the distance of the acceptor nucleus from the donor *DUX4* expressing nucleus (Figure 2). Since DUX4 is a transcription factor, the consequence of this is that even transcriptomes of nuclei that do not express *DUX4* themselves will be rewired by DUX4, thus explaining the observed easier mRNA detection of DUX4 target genes than *DUX4* itself [289]. This was also confirmed by single-nucleus RNA-seq (snRNA-seq), when many more nuclei show expression of DUX4 target genes while *DUX4* mRNA itself is in majority of cases not detectable in them [295]. For this reason, some of the DUX4 target genes have been considered as potential biomarkers instead of direct detection of *DUX4* [296]. Interestingly, during the course of differentiation, *DUX4* expression and expression of its target genes become discordant, when nuclei can remain expressing DUX4 target genes even after the nucleus is no more DUX4 protein positive [297]. One plausible explanation for this phenomenon is that DUX4 initiates expression of, among others, a cascade of transcription factors including its gene orthologue *DUXA*, which can then contribute to their perduring expression [295,297]. In addition, DUX4 was shown to induce changes in the chromatin landscape of its target genes by at least two distinct mechanisms, which sensitize these genes for their re-activation or sustained expression. First, DUX4 was shown to recruit the p300/CBP H3K27 acetyltransferase complex to its target DNA sites via its C-terminal transactivation domain, which helps chromatin opening of these loci for transcription [298]. Indeed, treating DUX4-expressing cells with a selective p300 inhibitor was sufficient to attenuate transcriptomic changes known to be elicited by DUX4 [299]. Second, DUX4 induces expression of two histone variants, namely H3.X and H3.Y, which get incorporated into gene bodies of DUX4 target genes resulting in a more relaxed chromatin configuration [300]. As some evidence suggests that endogenous *DUX4* expression occurs in bursts [292], after initial DUX4-mediated re-setting of the chromatin, following bursts of *DUX4* expression can lead to enhanced activation of its target genes as their chromatin is already more accessible for transcription [300].

Apart from the DUX4-induced transcriptional changes, DUX4 has been linked to other disruptive processes which might contribute to its myopathic effect. High levels of DUX4 can cause apoptosis in skeletal muscle cells via distinct mechanisms including activation of caspase 3/7- [301] and p53-mediated apoptotic pathways [302], induction of hypoxia signaling [303], increasing sensitivity to oxidative stress [304,305], upregulation of the pro-apoptotic factor MYC [306] and/or activation of the double-stranded RNA (dsRNA) response pathway [306,307]. On the other hand, low expression of DUX4 in myogenic cells was shown to negatively affect their myogenic differentiation potential *in vitro* [308]. Homeodomains of DUX4 display high amino acid sequence homology to a homeodomain of the muscle specific transcription factor PAX7 [309]. PAX7 is strictly expressed in myogenic precursor satellite cells and is required

for their proliferative capacity, thus ensuring a regenerative potential of skeletal muscle tissue [310,311]. Because of this homology, it was hypothesized that DUX4 might interfere with the transcriptional program activated by PAX7 thus leading to impaired myogenesis [309]. In line with this competitive inhibition model, it was demonstrated that overexpression of Pax7 in murine C2C12 myogenic cells counteracts the DUX4-induced cytotoxic effect in a dose-dependent manner [312]. However, DUX4 and PAX7 have non-overlapping expression patterns during normal myogenic differentiation, which argues against this competitive model [313]. Despite that, a recent analysis of different gene expression studies from muscle biopsies showed that PAX7 downstream genes (so-called PAX7 target gene score) are indeed repressed in FSHD samples compared to controls [314]. Intriguingly, the PAX7 score was proposed to be a more robust discriminator of FSHD-affected muscles than the expression of *DUX4* or its target genes [315] and it was shown that this score is a good biomarker for FSHD progression over a period of at least 1 year [316], therefore offering a possibility of being utilized for monitoring of FSHD development in future clinical trials as a reliable biomarker is still missing. A more recent transcriptomic study conducted on FSHD muscle biopsies suggested that DUX4 and PAX7 expression signatures might rather mark different stages of the disease (van Den Heuvel et al., 2022).

Scope of the thesis

Research presented in this thesis focuses both on *cis* and *trans* contributors to Facioscapulohumeral muscular dystrophy. In **chapter 2**, we employ a genome editing tool termed adenine base editing to efficiently mutate the somatic polyadenylation signal of *DUX4,* an important *cis* modifier in FSHD to test this approach as a possible future FSHD gene therapy. In **chapter 3**, we describe a proband with clinical symptoms consistent with FSHD that carries a homozygous out-of-frame deletion in exon 2 of the *LRIF1* gene combined with a disease permissive D4Z4 allele of 13 units. We confirmed that the D4Z4 epigenetic profile in the proband's cells exhibits perturbations as described for FSHD2 cases and we detect also the expression of *DUX4* itself in the proband's cells, thus uncovering a novel *trans* modifier in FSHD. We further extend this finding in **chapter 4**, where we study the action of LRIF1 together with its interacting partner SMCHD1 in D4Z4 repression in human somatic cells with distinct D4Z4 chromatin contexts. And lastly, in **chapter 5**, we explore the role of all three FSHD2 genes by performing loss of function studies in mESCs and we uncover the assistance of Lrif1 in the repression of mouse *Dux*, which is a functional homologue of human *DUX4*.

References

1. VANYUSHIN, B. F., BELOZERSKY, A. N., KOKURINA, N. A. & KADIROVA, D. X. 5-Methylcytosine and 6-Methylaminopurine in Bacterial DNA. *Nat. 1968 2185146* **218**, 1066–1067 (1968).
2. Li, X. *et al.* The exploration of N6-deoxyadenosine methylation in mammalian genomes. *Protein Cell 2021* 1–13 (2021) doi:10.1007/S13238-021-00866-3.
3. Nowialis, P. *et al.* Catalytically inactive Dnmt3b rescues mouse embryonic development by accessory and repressive functions. *Nat. Commun.* **10**, (2019).
4. Doskočil, J. & Šorm, F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids. *Biochim. Biophys. Acta* **55**, 953–959 (1962).
5. Bird, A. P. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J. Mol. Biol.* **118**, 49–60 (1978).
6. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nat. 2012 4847394* **484**, 339–344 (2012).
7. Seisenberger, S. *et al.* The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Mol. Cell* **48**, 849–862 (2012).
8. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.* **10**, 2709 (1982).
9. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nat. 1986 3216067* **321**, 209–213 (1986).
10. Norris, D. P., Brockdorff, N. & Rastan, S. Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. *Mamm. Genome 1991 12* **1**, 78–83 (1991).
11. SanMiguel, J. M. & Bartolomei, M. S. DNA methylation dynamics of genomic imprinting in mouse development. *Biol. Reprod.* **99**, 252–262 (2018).
12. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nat. Genet. 2010 4212* **42**, 1093–1100 (2010).
13. Zeng, J., Nagrajan, H. K. & Yi, S. V. Fundamental diversity of human CpG islands at multiple biological levels. *http://dx.doi.org/10.4161/epi.27654* **9**, 483–491 (2014).
14. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010 (2011).
15. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. **520**, 243–247 (2015).
16. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. **543**, 72–77 (2017).
17. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 11995 (1993).
18. Shen, J. C., Rideout, W. M., 3rd & Jones, P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22**, 972 (1994).
19. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science (80-.).* **293**, 1089–1093 (2001).
20. Guibert, S., Forné, T. & Weber, M. Global profiling of DNA methylation erasure in mouse primordial germ cells. *Genome Res.* **22**, 633–641 (2012).
21. Smith, Z. D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nat. 2014 5117511* **511**, 611–615 (2014).
22. Hill, P. W. S. *et al.* Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nat. 2018 5557696* **555**, 392–396 (2018).
23. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol. 2019 2010* **20**, 590–607 (2019).
24. Bestor, T., Laudano, A., Mattaliano, R. & Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J. Mol. Biol.* **203**, 971–983 (1988).

25. Hermann, A., Goyal, R. & Jeltsch, A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites *. *J. Biol. Chem.* **279**, 48350–48359 (2004).
26. Pradhan, S. *et al.* Baculovirus-mediated expression and characterization of the full-length murine DNA methyltransferase. *Nucleic Acids Res.* **25**, 4666–4673 (1997).
27. Vilkaitis, G., Suetake, I., Klimašauskas, S. & Tajima, S. Processive Methylation of Hemimethylated CpG Sites by Mouse Dnmt1 DNA Methyltransferase. *J. Biol. Chem.* **280**, 64–72 (2005).
28. Fatemi, M., Hermann, A., Pradhan, S. & Jeltsch, A. The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. *J. Mol. Biol.* **309**, 1189–1199 (2001).
29. Li, Y. *et al.* Stella safeguards the oocyte methylome by preventing de novo methylation mediated by DNMT1. *Nat. 2018 5647734* **564**, 136–140 (2018).
30. Wang, Q. *et al.* Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. *Nat. Genet. 2020 528* **52**, 828–839 (2020).
31. Ming, X. *et al.* Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration. *Cell Res. 2020 3011* **30**, 980–996 (2020).
32. Haggerty, C. *et al.* Dnmt1 has de novo activity targeted to transposable elements. *Nat. Struct. Mol. Biol. 2021 287* **28**, 594–603 (2021).
33. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases [1]. *Nature Genetics* vol. 19 219–220 (1998).
34. Watanabe, D., Suetake, I., Tada, T. & Tajima, S. Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. *Mech. Dev.* **118**, 187–190 (2002).
35. Lei, H. *et al.* De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122**, 3195–3205 (1996).
36. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* **23**, 5594–5605 (2003).
37. Yan, X. J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat. Genet. 2011 434* **43**, 309–315 (2011).
38. Duymich, C. E., Charlet, J., Yang, X., Jones, P. A. & Liang, G. DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. *Nat. Commun.* **7**, 11453 (2016).
39. Molaro, A., Malik, H. S. & Bourc'his, D. Dynamic Evolution of De Novo DNA Methyltransferases in Rodent and Primate Genomes. *Mol. Biol. Evol.* **37**, 1882–1892 (2020).
40. Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983–1993 (2002).
41. Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H. & Tajima, S. DNMT3L Stimulates the DNA Methylation Activity of Dnmt3a and Dnmt3b through a Direct Interaction. *J. Biol. Chem.* **279**, 27816–27823 (2004).
42. Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nat. 2004 4317004* **431**, 96–99 (2004).
43. Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science (80-.).* **294**, 2536–2539 (2001).
44. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science (80-.).* **334**, 194 (2011).
45. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (80-.).* **324**, 930–935 (2009).
46. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nat. 2010 4667310* **466**, 1129–1133 (2010).
47. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (80-.).* **333**, 1300–1303 (2011).
48. Maiti, A. & Drohat, A. C. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. *J. Biol. Chem.* **286**, 35334–35338 (2011).

49. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (80-.).* **333**, 1303–1307 (2011).
50. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science (80-.).* **184**, 868–871 (1974).
51. Fukagawa, T. & Earnshaw, W. C. The Centromere: Chromatin Foundation for the Kinetochore Machinery. *Dev. Cell* **30**, 496 (2014).
52. Smith, B. C. & Denu, J. M. Chemical mechanisms of histone lysine and arginine modifications. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1789**, 45–57 (2009).
53. Heitz, E. *Das heterochromatin der moose*. (Bornträger, 1928).
54. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nat. 2004 4317010* **431**, 873–878 (2004).
55. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in polycomb-group silencing. *Science (80-.).* **298**, 1039–1043 (2002).
56. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nat. 2006 4417091* **441**, 349–353 (2006).
57. Lee, T. I. *et al.* Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. *Cell* **125**, 301–313 (2006).
58. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science (80-.).* **300**, 131–135 (2003).
59. de Napoles, M. *et al.* Polycomb Group Proteins Ring1A/B Link Ubiquitylation of Histone H2A to Heritable Gene Silencing and X Inactivation. *Dev. Cell* **7**, 663–676 (2004).
60. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nat. 2001 4106824* **410**, 120–124 (2001).
61. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nat. 2001 4106824* **410**, 116–120 (2001).
62. Peters, A. H. F. M. *et al.* Partitioning and Plasticity of Repressive Histone Methylation States in Mammalian Chromatin. *Mol. Cell* **12**, 1577–1589 (2003).
63. Peters, A. H. F. M. *et al.* Loss of the Suv39h Histone Methyltransferases Impairs Mammalian Heterochromatin and Genome Stability. *Cell* **107**, 323–337 (2001).
64. Matsui, T. *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nat. 2010 4647290* **464**, 927–931 (2010).
65. Karimi, M. M. *et al.* DNA Methylation and SETDB1/H3K9me3 Regulate Predominantly Distinct Sets of Genes, Retroelements, and Chimeric Transcripts in mESCs. *Cell Stem Cell* **8**, 676–687 (2011).
66. Keniry, A. *et al.* Setdb1-mediated H3K9 methylation is enriched on the inactive X and plays a role in its epigenetic silencing. *Epigenetics Chromatin 2016 91* **9**, 1–20 (2016).
67. Minkovsky, A. *et al.* The Mbd1-Atf7ip-Setdb1 pathway contributes to the maintenance of X chromosome inactivation. *Epigenetics Chromatin 2014 71* **7**, 1–15 (2014).
68. Feldman, N. *et al.* G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat. Cell Biol. 2006 82* **8**, 188–194 (2006).
69. Tachibana, M. *et al.* Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. *Genes Dev.* **19**, 815–826 (2005).
70. Tachibana, M. *et al.* G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. *Genes Dev.* **16**, 1779–1791 (2002).
71. Montavon, T. *et al.* Complete loss of H3K9 methylation dissolves mouse heterochromatin organization. *Nat. Commun. 2021 121* **12**, 1–16 (2021).
72. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).
73. Bernstein, B. E. *et al.* Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell* **120**, 169–181 (2005).

74. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nat. 2008 4547205* **454**, 766–770 (2008).

75. Zhang, Y. *et al.* Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Res.* **38**, 4246–4253 (2010).

76. Otani, J. *et al.* Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain. *EMBO Rep.* **10**, 1235–1241 (2009).

77. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nat. 2015 5207546* **520**, 243–247 (2015).

78. Dhayalan, A. *et al.* The Dnmt3a PWWP Domain Reads Histone 3 Lysine 36 Trimethylation and Guides DNA Methylation. *J. Biol. Chem.* **285**, 26114–26120 (2010).

79. Statham, A. L. *et al.* Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22**, 1120–1127 (2012).

80. Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138 (2012).

81. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).

82. Pan, G. *et al.* Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell* **1**, 299–312 (2007).

83. Bartke, T. *et al.* Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation. *Cell* **143**, 470–484 (2010).

84. Reddington, J. P. *et al.* Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. *Genome Biol. 2013 143* **14**, 1–17 (2013).

85. Fu, K., Bonora, G. & Pellegrini, M. Interactions between core histone marks and DNA methyltransferases predict DNA methylation patterns observed in human cells and tissues. *https://doi.org/10.1080/15592294.2019.1666649* **15**, 272–282 (2019).

86. Ren, W. *et al.* Direct readout of heterochromatic H3K9me3 regulates DNMT1-mediated maintenance DNA methylation. *Proc. Natl. Acad. Sci.* **117**, 18439–18447 (2020).

87. Liu, X. *et al.* UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nat. Commun. 2013 41* **4**, 1–13 (2013).

88. Rothbart, S. B. *et al.* Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol. 2012 1911* **19**, 1155–1160 (2012).

89. Lehnertz, B. *et al.* Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Curr. Biol.* **13**, 1192–1200 (2003).

90. Tsumura, A. *et al.* Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes to Cells* **11**, 805–814 (2006).

91. Saksouk, N. *et al.* Redundant Mechanisms to Form Silent Chromatin at Pericentromeric Regions Rely on BEND3 and DNA Methylation. *Mol. Cell* **56**, 580–594 (2014).

92. Xin, H., Yoon, H. G., Singh, P. B., Wong, J. & Qin, J. Components of a Pathway Maintaining Histone Modification and Heterochromatin Protein 1 Binding at the Pericentric Heterochromatin in Mammalian Cells. *J. Biol. Chem.* **279**, 9539–9546 (2004).

93. Hoyt, S. J. *et al.* From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *bioRxiv* 2021.07.12.451456 (2021) doi:10.1101/2021.07.12.451456.

94. Bakhtiari, M. *et al.* Variable number tandem repeats mediate the expression of proximal genes. *Nat. Commun. 2021 121* **12**, 1–12 (2021).

95. Garg, P. *et al.* Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am. J. Hum. Genet.* **108**, 809–824 (2021).

96. Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295**, 4134–4170 (2020).

97. Depienne, C. & Mandel, J. L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).

98. Trizzino, M. *et al.* Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.* **27**, 1623–1633 (2017).

99. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nat. 2001 4096822* **409**, 860–921 (2001).

100. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-.).* **291**, 1304–1351 (2001).

101. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *bioRxiv* **16**, 2021.07.12.452052 (2021).

102. Gershman, A. *et al.* Epigenetic Patterns in a Complete Human Genome. *bioRxiv* 2021.05.26.443420 (2021) doi:10.1101/2021.05.26.443420.

103. (IWGSC), T. I. W. G. S. C. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (80-.).* **361**, (2018).

104. Platt, R. N., II, Vandewege, M. W. & Ray, D. A. Mammalian transposable elements and their impacts on genome evolution. *Chromosom. Res.* **26**, 25 (2018).

105. Torres-Padilla, M.-E. On transposons and totipotency. *Philos. Trans. R. Soc. B* **375**, (2020).

106. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).

107. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).

108. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).

109. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol. 2015 171* **17**, 16–29 (2015).

110. Darrow, E. M. *et al.* Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci.* **113**, E4504–E4512 (2016).

111. Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol. 2015 161* **16**, 1–21 (2015).

112. Garrick, D., Fiering, S., Martin, D. I. K. & Whitelaw, E. Repeat-induced gene silencing in mammals. *Nat. Genet. 1998 181* **18**, 56–59 (1998).

113. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet. 2007 3910* **39**, 1256–1260 (2007).

114. Rennison, D. J., Owens, G. L. & Taylor, J. S. Opsin gene duplication and divergence in ray-finned fish. *Mol. Phylogenet. Evol.* **62**, 986–1008 (2012).

115. Fondon, J. W. & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci.* **101**, 18058–18063 (2004).

116. Deutekom, J. C. T. V. *et al.* FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum. Mol. Genet.* **2**, 2037–2042 (1993).

117. Hobbs, S. L. A., Kpodar, P. & DeLong, C. M. O. The effect of T-DNA copy number, position and methylation on reporter gene expression in tobacco transformants. *Plant Mol. Biol. 1990 156* **15**, 851–864 (1990).

118. Assaad, F. F., Tucker, K. L. & Signer, E. R. Epigenetic repeat-induced gene silencing (RIGS) inArabidopsis. *Plant Mol. Biol. 1993 226* **22**, 1067–1085 (1993).

119. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Genome Res.* **14**, 9–25 (1975).

120. Holliday, R. & Pugh, J. DNA modification mechanisms and gene activity during development. *Science (80-.).* **187**, 226–232 (1975).

121. Henikoff, S. Conspiracy of silence among repeated transgenes. *BioEs-says* **20**, 532–535 (1998).

122. Weiler, K. S. & Wakimoto, B. T. HETEROCHROMATIN AND GENE EXPRESSION IN DROSOPHILA. *https://doi.org/10.1146/annurev.ge.29.120195.003045* **29**, 577–605 (2003).

123. Muller, H. J. Types of visible variations induced by X-rays in Drosophila. *J. Genet.* **22**, 299–334 (1930).

124. Girton, J. R. & Johansen, K. M. Chapter 1 Chromatin Structure and the Regulation of Gene Expression: The Lessons of PEV in Drosophila. *Adv. Genet.* **61**, 1–43 (2008).

125. Blewitt, M. & Whitelaw, E. The Use of Mouse Models to Study Epigenetics. *Cold Spring Harb. Perspect. Biol.* **5**, a017939 (2013).

126. Blewitt, M. E. *et al.* An N-ethyl-N-nitrosourea screen for genes involved in variegation in the mouse. *Proc. Natl. Acad. Sci.* **102**, 7629–7634 (2005).

127. Daxinger, L. *et al.* An ENU mutagenesis screen identifies novel and known genes involved in epigenetic processes in the mouse. *Genome Biol. 2013 149* **14**, 1–17 (2013).

128. Preis, J. I., Downes, M., Oates, N. A., Rasko, J. E. J. & Whitelaw, E. Sensitive Flow Cytometric Analysis Reveals a Novel Type of Parent-of-Origin Effect in the Mouse Genome. *Curr. Biol.* **13**, 955–959 (2003).

129. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet.* **18**, 348–351 (2002).

130. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet. 2007 84* **8**, 272–285 (2007).

131. Duhl, D. M. J., Vrieling, H., Miller, K. A., Wolff, G. L. & Barsh, G. S. Neomorphic agouti mutations in obese yellow mice. *Nat. Genet. 1994 81* **8**, 59–65 (1994).

132. Miller, M. W. *et al.* Cloning of the mouse agouti gene predicts a secreted protein ubiquitously expressed in mice carrying the lethal yellow mutation. *Genes Dev.* **7**, 454–467 (1993).

133. Isbel, L. *et al.* Wiz binds active promoters and CTCF-binding sites and is required for normal behaviour in the mouse. *Elife* **5**, (2016).

134. Michaud, E. J. *et al.* Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. *Genes Dev.* **8**, 1463–1472 (1994).

135. Youngson, N. A. *et al.* No evidence for cumulative effects in a Dnmt3b hypomorph across multiple generations. *Mamm. Genome 2013 245* **24**, 206–217 (2013).

136. Ashe, A. *et al.* A genome-wide screen for modifiers of transgene variegation identifies genes with critical roles in development. *Genome Biol. 2008 912* **9**, 1–16 (2008).

137. Sorolla, M. A., Marqués, M., Parisi, E. & Sorolla, A. An N-ethyl-N-Nitrosourea Mutagenesis Screen in Mice Reveals a Mutation in Nuclear Respiratory Factor 1 (Nrf1) Altering the DNA Methylation State and Correct Embryonic Development. *Anim. 2021, Vol. 11, Page 2103* **11**, 2103 (2021).

138. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99**, 247–257 (1999).

139. Yehezkel, S., Segev, Y., Viegas-Péquignot, E., Skorecki, K. & Selig, S. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. *Hum. Mol. Genet.* **17**, 2776–2789 (2008).

140. Gendrel, A.-V. *et al.* Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome. *Dev. Cell* **23**, 265–279 (2012).

141. Rice, J. C. *et al.* Histone Methyltransferases Direct Different Degrees of Methylation to Define Distinct Chromatin Domains. *Mol. Cell* **12**, 1591–1598 (2003).

142. García-Cao, M., O'Sullivan, R., Peters, A. H. F. M., Jenuwein, T. & Blasco, M. A. Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases. *Nat. Genet. 2004 361* **36**, 94–99 (2003).

143. Groh, S. *et al.* Morc3 silences endogenous retroviruses by enabling Daxx-mediated histone H3.3 incorporation. *Nat. Commun. 2021 121* **12**, 1–18 (2021).

144. Matsui, T. *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nat. 2010 4647290* **464**, 927–931 (2010).

145. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nat. 2010 4637278* **463**, 237–240 (2010).

146. Lemmers, R. J. L. F. *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat. Genet.* **44**, 1370–1374 (2012).

147. van den Boogaard, M. L. *et al.* Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am. J. Hum. Genet.* **98**, 1020–1029 (2016).

148. Xu, G.-L. *et al.* Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187–191 (1999).

149. Gordon, C. T. *et al.* De novo mutations in SMCHD1 cause Bosma arhinia microphthalmia syndrome and abrogate nasal development. *Nat. Genet.* **49**, 249–255 (2017).

150. Wijmenga, C. *et al.* Mapping of facioscapulohumeral muscular dystrophy gene to chromosome 4q35-qter by multipoint linkage analysis and in situ hybridization. *Genomics* **9**, 570–575 (1991).

151. Wijmenga, C. *et al.* Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet. 1992 21* **2**, 26–30 (1992).

152. Weiffenbach, B. *et al.* Mapping the facioscapulohumeral muscular dystrophy gene is complicated by chromosome 4q35 recombination events. *Nat. Genet. 1993 42* **4**, 165–169 (1993).

153. Zampatti, S. *et al.* Facioscapulohumeral muscular dystrophy (FSHD) molecular diagnosis: from traditional technology to the NGS era. *neurogenetics 2019 202* **20**, 57–64 (2019).

154. Hewitt, J. E. *et al.* Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystropothhy. *Hum. Mol. Genet.* **3**, 1287–1295 (1994).

155. Lyle, R., Wright, T. J., Clark, L. N. & Hewitt, J. E. The FSHD-Associated Repeat, D4Z4, Is a Member of a Dispersed Family of Homeobox-Containing Repeats, Subsets of Which Are Clustered on the Short Arms of the Acrocentric Chromosomes. *Genomics* **28**, 389–397 (1995).

156. Bakker, E. *et al.* The FSHD-linked locus D4F104S1 (p13E-11) on 4q35 has a homologue on 10qter. *Muscle Nerve. Suppl.* S39-44 (1995).

157. Deidda, G. *et al.* Physical mapping evidence for a duplicated region on chromosome 10qter showing high homology with the facioscapulohumeral muscular dystrophy locus on chromosome 4qter. *Eur. J. Hum. Genet.* **3**, 155–167 (1995).

158. van Overveld, P. G. M. *et al.* Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat. Genet.* **35**, 315–317 (2003).

159. Zeng, W. *et al.* Specific Loss of Histone H3 Lysine 9 Trimethylation and HP1γ/Cohesin Binding at D4Z4 Repeats Is Associated with Facioscapulohumeral Dystrophy (FSHD). *PLoS Genet.* **5**, e1000559 (2009).

160. Gabellini, D., Green, M. R. & Tupler, R. Inappropriate Gene Activation in FSHD: A Repressor Complex Binds a Chromosomal Repeat Deleted in Dystrophic Muscle. *Cell* **110**, 339–348 (2002).

161. van Deutekom, J. C. T. *et al.* Identification of the First Gene (FRG1) from the FSHD Region on Human Chromosome 4q35. *Hum. Mol. Genet.* **5**, 581–590 (1996).

162. Rijkers, T. *et al.* FRG2, an FSHD candidate gene, is transcriptionally upregulated in differentiating primary myoblast cultures of FSHD patients. *J. Med. Genet.* **41**, 826–836 (2004).

163. Geel, M. van *et al.* Identification of a novel β-tubulin subfamily with one member (TUBB4Q) located near the telomere of chromosome region 4q35. *Cytogenet. Genome Res.* **88**, 316–321 (2000).

164. Jiang, G. *et al.* Testing the position-effect variegation hypothesis for facioscapulohumeral muscular dystrophy by analysis of histone modification and gene expression in subtelomeric 4q. *Hum. Mol. Genet.* **12**, 2909–2921 (2003).

165. Dixit, M. *et al.* DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proc. Natl. Acad. Sci.* **104**, 18157–18162 (2007).

166. Masny, P. S. *et al.* Analysis of allele-specific RNA transcription in FSHD by RNA-DNA FISH in single myonuclei. *Eur. J. Hum. Genet. 2010 184* **18**, 448–456 (2009).

167. Klooster, R. *et al.* Comprehensive expression analysis of FSHD candidate genes at the mRNA and protein level. *Eur. J. Hum. Genet. 2009 1712* **17**, 1615–1624 (2009).

168. Tupler, R. *et al.* Monosomy of distal 4q does not cause facioscapulohumeral muscular dystrophy. *J. Med. Genet.* **33**, 366–370 (1996).

169. Hewitt, J. E. *et al. Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy*. *Human Molecular Genetics* vol. 3 https://academic.oup.com/hmg/article-abstract/3/8/1287/554670 (1994).

170. Lee, J. H., Goto, K., Matsuda, C. & Arahata, K. Characterization of a tandemly repeated 3.3-kb KpnI unit in the facioscapulohumeral muscular dystrophy (FSHD) gene region on chromosome 4q35. *Muscle Nerve. Suppl.* S6-13 (1995).

171. Gabriëls, J. *et al.* Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. *Gene* **236**, 25–32 (1999).
172. Lemmers, R. J. L. F. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–3 (2010).
173. Deak, K. L. *et al.* Genotype-phenotype study in an FSHD family with a proximal deletion encompassing p13E-11 and D4Z4. *Neurology* **68**, 578–582 (2007).
174. Lemmers, R. J. L. F. *et al.* High-resolution breakpoint junction mapping of proximally extended D4Z4 deletions in FSHD1 reveals evidence for a founder effect. *Hum. Mol. Genet.* **00**, (2021).
175. Lemmers, R. J. L. F. *et al.* Chromosome 10q-linked FSHD identifies DUX4 as principal disease gene. *J. Med. Genet.* **0**, 1–9 (2021).
176. Hamanaka, K. *et al.* Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology* **94**, e2441–e2447 (2020).
177. de Greef, J. C. *et al.* Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum. Mutat.* **30**, 1449–1459 (2009).
178. Theadom, A. *et al.* Prevalence of Muscular Dystrophies: A Systematic Literature Review. *Neuroepidemiology* **43**, 259–268 (2014).
179. Greef, J. C. de *et al.* Clinical features of facioscapulohumeral muscular dystrophy 2(CME). *Neurology* **75**, 1548 (2010).
180. Sacconi, S. *et al.* FSHD1 and FSHD2 form a disease continuum. *Neurology* **92**, E2273–E2285 (2019).
181. Wohlgemuth, M. *et al.* A family-based study into penetrance in facioscapulohumeral muscular dystrophy type 1. *Neurology* **91**, e444–e454 (2018).
182. Tonini, M. M. O. *et al.* Asymptomatic carriers and gender differences in facioscapulohumeral muscular dystrophy (FSHD). *Neuromuscul. Disord.* **14**, 33–38 (2004).
183. Ricci, G. *et al.* Large scale genotype–phenotype analyses indicate that novel prognostic tools are required for families with facioscapulohumeral muscular dystrophy. *Brain* **136**, 3408–3417 (2013).
184. Zatz, M. *et al.* The Facioscapulohumeral Muscular Dystrophy (FSHD1) Gene Affects Males More Severely and More Frequently Than Females. *J. Med. Genet* **77**, 155–161 (1998).
185. Deenen, J. C. W. *et al.* Population-based incidence and prevalence of facioscapulohumeral dystrophy. *Neurology* **83**, 1056–9 (2014).
186. Padberg, G. W. A. M. *Facioscapulohumeral disease*. (1982).
187. Statland, J. M. & Tawil, R. Risk of functional impairment in Facioscapulohumeral muscular dystrophy. *Muscle Nerve* **49**, 520–527 (2014).
188. Klinge, L. *et al.* Severe phenotype in infantile facioscapulohumeral muscular dystrophy. *Neuromuscul. Disord.* **16**, 553–558 (2006).
189. Goselink, R. J. M. *et al.* Early onset facioscapulohumeral dystrophy – a systematic review using individual patient data. *Neuromuscul. Disord.* **27**, 1077–1083 (2017).
190. Padberg, G. W. *et al.* On the significance of retinal vascular disease and hearing loss in facioscapulohumeral muscular dystrophy. *Muscle Nerve* **18**, S73–S80 (1995).
191. Goselink, R. J. M. *et al.* Ophthalmological findings in facioscapulohumeral dystrophy. *Brain Commun.* **1**, (2019).
192. Lutz, K. L., Holte, L., Kliethermes, S. A., Stephan, C. & Mathews, K. D. Clinical and genetic features of hearing loss in facioscapulohumeral muscular dystrophy. *Neurology* **81**, 1374–1377 (2013).
193. Trevisan, C. P. *et al.* Facioscapulohumeral muscular dystrophy: a multicenter study on hearing function. *Audiol. Neurootol.* **13**, 1–6 (2008).
194. Brouwer, O. F., Padberg, G. W., Wijmenga, C. & Frants, R. R. Facioscapulohumeral Muscular Dystrophy in Early Childhood. *Arch. Neurol.* **51**, 387–394 (1994).
195. Funakoshi, M., Goto, K. & Arahata, K. Epilepsy and mental retardation in a subset of early onset 4q35-facioscapulohumeral muscular dystrophy. *Neurology* **50**, 1791–1794 (1998).

196. Goselink, R. J. M. *et al.* Early onset as a marker for disease severity in facioscapulohumeral muscular dystrophy. *Neurology* **92**, E378–E385 (2019).

197. Lemmers, R. J. F. L. *et al.* Contractions of D4Z4 on 4qB Subtelomeres Do Not Cause Facioscapulohumeral Muscular Dystrophy. *Am. J. Hum. Genet.* **75**, 1124–1130 (2004).

198. Lemmers, R. J. L. F. *et al.* Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *Am. J. Hum. Genet.* **81**, 884–94 (2007).

199. van Geel, M. *et al.* Genomic Analysis of Human Chromosome 10q and 4q Telomeres Suggests a Common Origin. *Genomics* **79**, 210–217 (2002).

200. Lemmers, R. J. L. F. *et al.* Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nat. Genet.* **32**, 235–236 (2002).

201. Lemmers, R. J. L. F. *et al.* Worldwide Population Analysis of the 4q and 10q Subtelomeres Identifies Only Four Discrete Interchromosomal Sequence Transfers in Human Evolution. *Am. J. Hum. Genet.* **86**, 364–377 (2010).

202. Jones, T. I., Himeda, C. L., Perez, D. P. & Jones, P. L. Large family cohorts of lymphoblastoid cells provide a new cellular model for investigating facioscapulohumeral muscular dystrophy. *Neuromuscul. Disord.* **27**, 221–238 (2017).

203. Scionti, I. *et al.* Large-Scale Population Analysis Challenges the Current Criteria for the Molecular Diagnosis of Fascioscapulohumeral Muscular Dystrophy. *Am. J. Hum. Genet.* **90**, 628–635 (2012).

204. Lemmers, R. J. *et al.* Deep characterization of a common D4Z4 variant identifies biallelic DUX4 expression as a modifier for disease penetrance in FSHD2. *Eur. J. Hum. Genet.* **26**, 94–106 (2018).

205. Snider, L. *et al.* 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 (2009).

206. Ricci, E. *et al.* Progress in the Molecular Diagnosis of Facioscapulohumeral Muscular Dystrophy and Correlation between the Number of KpnI Repeats at the 4q35 Locus and Clinical Phenotype. (1999) doi:10.1002/1531-8249.

207. Lunt, P. W. *et al.* Correlation between fragment size at D4F104S1 and age at onset or at wheelchair use, with a possible generational effect, accounts for much phenotypic variation in 4q35-facioscapulohumeral muscular dystrophy (FSHD). *Hum. Mol. Genet.* **4**, 951–958 (1995).

208. Mul, K. *et al.* Phenotype-genotype relations in facioscapulohumeral muscular dystrophy type 1. *Clin. Genet.* **94**, 521–527 (2018).

209. Wang, C.-H. *et al.* Correlation between muscle involvement, phenotype and D4Z4 fragment size in facioscapulohumeral muscular dystrophy. *Neuromuscul. Disord.* **22**, 331–8 (2012).

210. Tawil, R. *et al.* Evidence for anticipation and association of deletion size with severity in facioscapulohumerd muscular dystrophy. *Ann. Neurol.* **39**, 744–748 (1996).

211. Scionti, I. *et al.* Facioscapulohumeral muscular dystrophy: new insights from compound heterozygotes and implication for prenatal genetic counselling. *J. Med. Genet.* **49**, 171–178 (2012).

212. Schaap, M. *et al.* Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics 2013 141* **14**, 1–12 (2013).

213. Park, H. J. *et al.* Low D4Z4 copy number and gender difference in Korean patients with facioscapulohumeral muscular dystrophy type 1. *Neuromuscul. Disord.* **25**, 859–64 (2015).

214. Goto, K., Nishino, I. & Hayashi, Y. K. Rapid and accurate diagnosis of facioscapulohumeral muscular dystrophy. *Neuromuscul. Disord.* **16**, 256–261 (2006).

215. Lemmers, R. J. L. F., van der Wielen, M. J. R., Bakker, E., Frants, R. R. & van der Maarel, S. M. Rapid and accurate diagnosis of facioscapulohumeral muscular dystrophy. *Neuromuscul. Disord.* **16**, 615–7; author reply 617-8 (2006).

216. Himeda, C. L. *et al.* Myogenic Enhancers Regulate Expression of the Facioscapulohumeral Muscular Dystrophy-Associated DUX4 Gene. *Mol. Cell. Biol.* **34**, 1942–1955 (2014).

217. Cabianca, D. S. *et al.* A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* **149**, 819–31 (2012).

218. Deak, K. L. *et al.* Genotype-phenotype study in an FSHD family with a proximal deletion encompassing p13E-11 and D4Z4. *Neurology* **68**, 578–582 (2007).

219. Lemmers, R. J. L. F. *et al.* D4F104S1 deletion in facioscapulohumeral muscular dystrophy. *Neurology* **61**, 178–183 (2003).

220. Lemmers, R. J. L. F. *et al.* Inter-and Intrachromosomal Sub-Telomeric Rearrangements on 4q35: Implications for Facioscapulohumeral Muscular Dystrophy (FSHD) Aetiology and Diagnosis. *Hum. Mol. Genet.* **7**, 1207–1214 (1998).

221. De Greef, J. C. *et al.* Hypomethylation is restricted to the D4Z4 repeat array in phenotypic FSHD. *Neurology* **69**, 1018–1026 (2007).

222. Lemmers, R. J. L. F. *et al.* SMCHD1 mutation spectrum for facioscapulohumeral muscular dystrophy type 2 (FSHD2) and Bosma arhinia microphthalmia syndrome (BAMS) reveals disease-specific localisation of variants in the ATPase domain. *J. Med. Genet.* **56**, 693–700 (2019).

223. Lemmers, R. J. L. F. *et al.* Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum. Mol. Genet.* **24**, 659–669 (2015).

224. Brideau, N. J. *et al.* Independent Mechanisms Target SMCHD1 to Trimethylated Histone H3 Lysine 9-Modified Chromatin and the Inactive X Chromosome. *Mol. Cell. Biol.* **35**, 4053–68 (2015).

225. Gurzau, A. D. *et al.* SMCHD1's ubiquitin-like domain is required for N-terminal dimerization and chromatin localization. *Biochem. J.* **478**, 2555–2569 (2021).

226. Sacconi, S. *et al.* The FSHD2 Gene SMCHD1 Is a Modifier of Disease Severity in Families Affected by FSHD1. *Am. J. Hum. Genet.* **93**, 744–751 (2013).

227. Lemmers, R. J. L. F. *et al.* Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum. Mol. Genet.* **27**, 3488–3497 (2018).

228. Larsen, M. *et al.* Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. *Eur. J. Hum. Genet.* **23**, 808–816 (2015).

229. Cascella, R. *et al.* Digenic inheritance of shortened repeat units of the D4Z4 region and a loss-of-function variant in SMCHD1 in a Family with FSHD. *Front. Neurol.* **9**, 1027 (2018).

230. Goossens, R. *et al.* Intronic SMCHD1 variants in FSHD: Testing the potential for CRISPR-Cas9 genome editing. *J. Med. Genet.* **56**, 828–837 (2019).

231. Calandra, P. *et al.* Allele-specific DNA hypomethylation characterises FSHD1 and FSHD2. *J. Med. Genet.* **53**, 348–55 (2016).

232. Jones, T. I. *et al.* Identifying diagnostic DNA methylation profiles for facioscapulohumeral muscular dystrophy in blood and saliva using bisulfite sequencing. *Clin. Epigenetics* **6**, 1–16 (2014).

233. Balog, J. *et al.* Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. *Epigenetics* **10**, 1133–1142 (2015).

234. Jones, T. I. *et al.* Facioscapulohumeral muscular dystrophy family studies of DUX4 expression: evidence for disease modifiers and a quantitative model of pathogenesis. *Hum. Mol. Genet.* **21**, 4419–4430 (2012).

235. Blewitt, M. E. *et al.* SmcHD1, containing a structural-maintenance-of- chromosomes hinge domain, has a critical role in X inactivation. (2008) doi:10.1038/ng.142.

236. Chen, K., Czabotar, P. E., Blewitt, M. E. & Murphy, J. M. The hinge domain of the epigenetic repressor Smchd1 adopts an unconventional homodimeric configuration. *Biochem. J* **473**, 733–742 (2016).

237. Pedersen, L. C., Inoue, K., Kim, S., Perera, L. & Shaw, N. D. A ubiquitin-like domain is required for stabilizing the N-terminal ATPase module of human SMCHD1. *Commun. Biol.* **2**, 255 (2019).

238. Chen, K. *et al.* Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3535-44 (2015).

239. Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* **16**, 1–21 (2015).

240. Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nat. 2016 5357613* **535**, 575–579 (2016).

241. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

242. Jansz, N. *et al.* Smchd1 regulates long-range chromatin interactions on the inactive X chromosome and at Hox clusters. *Nat. Struct. Mol. Biol.* **25**, 766–777 (2018).

243. Wang, C.-Y., Jégu, T., Chu, H.-P., Oh, H. J. & Lee, J. T. SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X. *Cell* **174**, 406-421.e25 (2018).

244. Gdula, M. R. *et al.* The non-canonical SMC protein SmcHD1 antagonises TAD formation and compartmentalisation on the inactive X chromosome. *Nat. Commun.* **10**, 30 (2019).

245. Gendrel, A.-V. *et al.* Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. *Mol. Cell. Biol.* **33**, 3150–65 (2013).

246. Mould, A. W. *et al.* Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation. *Epigenetics Chromatin 2013 61* **6**, 1–16 (2013).

247. Wanigasuriya, I. *et al.* Smchd1 is a maternal effect gene required for genomic imprinting. *Elife* **9**, 1–27 (2020).

248. Kondo, T. *et al.* Whole-genome methylation scan in ICF syndrome: Hypomethylation of non-satellite DNA repeats D4Z4 and NBL2. *Hum. Mol. Genet.* **9**, 597–604 (2000).

249. Huichalaf, C., Micheloni, S., Ferri, G., Caccia, R. & Gabellini, D. DNA Methylation Analysis of the Macrosatellite Repeat Associated with FSHD Muscular Dystrophy at Single Nucleotide Level. *PLoS One* **9**, e115278 (2014).

250. Zeng, W. *et al.* Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. *Hum. Mutat.* **35**, 998–1010 (2014).

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

252. Van Overveld, P. G. M. *et al.* Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat. Genet. 2003 354* **35**, 315–317 (2003).

253. Haynes, P., Bomsztyk, K. & Miller, D. G. Sporadic DUX4 expression in FSHD myocytes is associated with incomplete repression by the PRC2 complex and gain of H3K9 acetylation on the contracted D4Z4 allele. *Epigenetics Chromatin* **11**, 47 (2018).

254. Balog, J. *et al.* Monosomy 18p is a risk factor for facioscapulohumeral dystrophy. *J. Med. Genet.* **55**, 469–478 (2018).

255. Campbell, A. E. *et al.* NuRD and CAF-1-mediated silencing of the D4Z4 array is modulated by DUX4-induced MBD3L proteins. *Elife* **7**, (2018).

256. Signorelli, M. *et al.* Evaluation of blood gene expression levels in facioscapulohumeral muscular dystrophy patients. *Sci. Reports 2020 101* **10**, 1–11 (2020).

257. Banerji, C. R. S., Panamarova, M. & Zammit, P. S. DUX4 expressing immortalized FSHD lymphoblastoid cells express genes elevated in FSHD muscle biopsies, correlating with the early stages of inflammation. *Hum. Mol. Genet.* **29**, 2285–2299 (2020).

258. Ferreboeuf, M. *et al.* DUX4 and DUX4 downstream target genes are expressed in fetal FSHD muscles. *Hum. Mol. Genet.* **23**, 171–181 (2014).

259. Williams, K. *et al.* Muscle group specific transcriptomic and DNA methylation differences related to developmental patterning in FSHD. *bioRxiv* 2021.09.28.462147 (2021) doi:10.1101/2021.09.28.462147.

260. Leidenroth, A. *et al.* Evolution of DUX gene macrosatellites in placental mammals. *Chromosoma* **121**, 489–497 (2012).

261. Clapp, J. *et al.* Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. *Am. J. Hum. Genet.* **81**, 264–279 (2007).

262. Eidahl, J. O. *et al.* Mouse Dux is myotoxic and shares partial functional homology with its human paralog DUX4. *Hum. Mol. Genet.* **25**, 4577–4589 (2016).

263. Whiddon, J. L., Langford, A. T., Wong, C. J., Zhong, J. W. & Tapscott, S. J. Conservation and innovation in the DUX4-family gene network. *Nat. Genet.* **49**, 935–940 (2017).

264. Ishiuchi, T. *et al.* Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat. Struct. Mol. Biol.* **22**, 662–671 (2015).

265. Bošković, A. *et al.* Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. *Genes Dev.* **28**, 1042–1047 (2014).

266. Huang, Z. *et al.* The chromosomal protein SMCHD1 regulates DNA methylation and the 2c-like state of embryonic stem cells by antagonizing TET proteins. *Sci. Adv.* **7**, eabb9149 (2021).
267. Grow, E. J. *et al.* p53 convergently activates Dux/DUX4 in embryonic stem cells and in facioscapulohumeral muscular dystrophy cell models. *Nat. Genet. 2021 538* **53**, 1207–1220 (2021).
268. Wu, K. *et al.* SETDB1-Mediated Cell Fate Transition between 2C-Like and Pluripotent States. *Cell Rep.* **30**, 25-36.e6 (2020).
269. Akiyama, T. *et al.* Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. *DNA Res.* **22**, 307–318 (2015).
270. Maksakova, I. A. *et al.* Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics and Chromatin* **6**, 15 (2013).
271. Genet, M. & Torres-Padilla, M.-E. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* **147**, (2020).
272. Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391-405.e19 (2018).
273. Cossec, J. C. *et al.* SUMO Safeguards Somatic and Pluripotent Cell Identities by Enforcing Distinct Chromatin States. *Cell Stem Cell* **23**, 742-757.e8 (2018).
274. Eckersley-Maslin, M. *et al.* Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* **33**, 194–208 (2019).
275. Fu, X., Wu, X., Djekidel, M. N. & Zhang, Y. Myc and Dnmt1 impede the pluripotent to totipotent state transition in embryonic stem cells. *Nat. Cell Biol.* **21**, 835–844 (2019).
276. Chen, Z. & Zhang, Y. Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat. Genet. 2019 516* **51**, 947–951 (2019).
277. De Iaco, A., Verp, S., Offner, S., Grun, D. & Trono, D. DUX is a non-essential synchronizer of zygotic genome activation. *Dev.* **147**, (2020).
278. Bosnakovski, D., Gearhart, M. D., Ho Choi, S. & Kyba, M. Dux facilitates post-implantation development, but is not essential for zygotic genome activation. *Biol. Reprod.* **104**, 83–93 (2021).
279. Guo, M. *et al.* Precise temporal regulation of Dux is important for embryo development. *Cell Res. 2019 2911* **29**, 956–959 (2019).
280. Yu, H. *et al.* rRNA biogenesis regulates mouse 2C-like state by 3D structure reorganization of peri-nucleolar heterochromatin. *Nat. Commun. 2021 121* **12**, 1–21 (2021).
281. Xie, S. Q. *et al.* Nucleolar-based Dux repression is essential for 2-cell stage exit. *bioRxiv* 2021.11.11.468235 (2021) doi:10.1101/2021.11.11.468235.
282. Lafontaine, D. L. J., Riback, J. A., Bascetin, R. & Brangwynne, C. P. The nucleolus as a multiphase liquid condensate. *Nat. Rev. Mol. Cell Biol. 2020 223* **22**, 165–182 (2020).
283. McStay, B. Nucleolar organizer regions: genomic 'dark matter' requiring illumination. *Genes Dev.* **30**, 1598–1610 (2016).
284. Masny, P. S. *et al.* Localization of 4q35.2 to the nuclear periphery: is FSHD a nuclear envelope disease? *Hum. Mol. Genet.* **13**, 1857–1871 (2004).
285. Tam, R., Smith, K. P. & Lawrence, J. B. The 4q subtelomere harboring the FSHD locus is specifically anchored with peripheral heterochromatin unlike most human telomeres. *J. Cell Biol.* **167**, 269–279 (2004).
286. Winokur, S. T., Bengtsson, U., Vargas, J. C., Wasmuth, J. J. & Altherr, M. R. The Evolutionary Distribution and Structural Organization of the Homeobox-Containing Repeat D4Z4 Indicates a Functional Role for the Ancestral Copy in the FSHD Region. *Hum. Mol. Genet.* **5**, 1567–1575 (1996).
287. Yao, Z. *et al.* DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum. Mol. Genet.* **23**, 5342–5352 (2014).
288. Jagannathan, S. *et al.* Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. *Hum. Mol. Genet.* **25**, 4419–4431 (2016).
289. Geng, L. N. *et al.* DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev. Cell* **22**, 38–51 (2012).

290. Young, J. M. *et al.* DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. *PLoS Genet.* **9**, e1003947 (2013).

291. Block, G. J. *et al.* Asymmetric Bidirectional Transcription from the FSHD-Causing D4Z4 Array Modulates DUX4 Production. *PLoS One* **7**, e35532 (2012).

292. Rickard, A. M., Petek, L. M. & Miller, D. G. Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. *Hum. Mol. Genet.* **24**, (2015).

293. Van Den Heuvel, A. *et al.* Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. *Hum. Mol. Genet.* **28**, 1064–1075 (2019).

294. Snider, L. *et al.* Facioscapulohumeral Dystrophy: Incomplete Suppression of a Retrotransposed Gene. *PLoS Genet.* **6**, e1001181 (2010).

295. Jiang, S. *et al.* Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. *PLOS Genet.* **16**, e1008754 (2020).

296. Wang, L. H. *et al.* MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD. *Hum. Mol. Genet.* **28**, 476–486 (2019).

297. Chau, J. *et al.* Relationship of DUX4 and target gene expression in FSHD myocytes. *Hum. Mutat.* **42**, 421–433 (2021).

298. Choi, S. H. *et al.* DUX4 recruits p300/CBP through its C-terminus and induces global H3K27 acetylation changes. *Nucleic Acids Res.* **44**, 5161–73 (2016).

299. Bosnakovski, D. *et al.* A novel P300 inhibitor reverses DUX4-mediated global histone H3 hyperacetylation, target gene expression, and cell death. *Sci. Adv.* **5**, 7781–7792 (2019).

300. Resnick, R. *et al.* DUX4-Induced Histone Variants H3.X and H3.Y Mark DUX4 Target Genes for Expression. *Cell Rep.* **29**, 1812-1820.e5 (2019).

301. Kowaljow, V. *et al.* The DUX4 gene at the FSHD1A locus encodes a pro-apoptotic protein. *Neuromuscul. Disord.* **17**, 611–23 (2007).

302. Wallace, L. M. *et al. DUX4* , a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. *Ann. Neurol.* **69**, 540–552 (2011).

303. Lek, A. *et al.* Applying genome-wide CRISPR-Cas9 screens for therapeutic discovery in facioscapulohumeral muscular dystrophy. *Sci. Transl. Med.* **12**, 271 (2020).

304. Bosnakovski, D. *et al.* High-throughput screening identifies inhibitors of DUX4-induced myoblast toxicity. *Skelet. Muscle* **4**, 1–11 (2014).

305. Dmitriev, P. *et al.* DUX4-induced constitutive DNA damage and oxidative stress contribute to aberrant differentiation of myoblasts from FSHD patients. *Free Radic. Biol. Med.* **99**, 244–258 (2016).

306. Shadle, S. C. *et al.* DUX4-induced dsRNA and MYC mRNA stabilization activate apoptotic pathways in human cell models of facioscapulohumeral dystrophy. *PLOS Genet.* **13**, e1006658 (2017).

307. Shadle, S. C. *et al.* DUX4-induced bidirectional HSATII satellite repeat transcripts form intranuclear double-stranded RNA foci in human cell models of FSHD. *Hum. Mol. Genet.* **28**, 3997–4011 (2019).

308. Bosnakovski, D. *et al.* Low level DUX4 expression disrupts myogenesis through deregulation of myogenic gene expression. *Sci. Rep.* **8**, 1–12 (2018).

309. Bosnakovski, D. *et al.* The DUX4 homeodomains mediate inhibition of myogenesis and are functionally exchangeable with the Pax7 homeodomain. *J. Cell Sci.* **130**, 3685–3697 (2017).

310. Von Maltzahn, J., Jones, A. E., Parks, R. J. & Rudnicki, M. A. Pax7 is critical for the normal function of satellite cells in adult skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16474–16479 (2013).

311. Seale, P. *et al.* Pax7 Is Required for the Specification of Myogenic Satellite Cells. *Cell* **102**, 777–786 (2000).

312. Bosnakovski, D. *et al.* An isogenetic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. *EMBO J.* **27**, 2766–2779 (2008).

313. Haynes, P., Kernan, K., Zhou, S. L. & Miller, D. G. Expression patterns of FSHD-causing DUX4 and myogenic transcription factors PAX3 and PAX7 are spatially distinct in differentiating human stem cell cultures. *Skelet. Muscle* **7**, 1–13 (2017).

314. Banerji, C. R. S. *et al.* PAX7 target genes are globally repressed in facioscapulohumeral muscular dystrophy skeletal muscle. *Nat. Commun. 2017 81* **8**, 1–13 (2017).

315. Banerji, C. R. S. & Zammit, P. S. PAX7 target gene repression is a superior FSHD biomarker than DUX4 target gene activation, associating with pathological severity and identifying FSHD at the single-cell level. *Hum. Mol. Genet.* **28**, 2224–2236 (2019).

316. Banerji, C. R. S. PAX7 target gene repression associates with FSHD progression and pathology over 1 year. *Hum. Mol. Genet.* **29**, 2124–2133 (2020).

317. Van Den Heuvel, A. *et al.* Facioscapulohumeral dystrophy transcriptome signatures correlate with different stages of disease and are marked by different MRI biomarkers. *Sci. Reports 2022 121* **12**, 1–18 (2022).

318. Okano, M., Xie, S. & Li, E. Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic Acids Res.* **26**, 2536 (1998).

319. Goll, M. G. *et al.* Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science (80-.).* **311**, 395–398 (2006).

320. Vogan, A. A. *et al.* The Enterprise, a massive transposon carrying Spok meiotic drive genes. *Genome Res.* **31**, 789–798 (2021).

321. Pech, M., Igo-Kemenes, T. & Zachau, H. G. Nucleotide sequence of a highly repetitive component of rat DNA. *Nucleic Acids Res.* **7**, 417–432 (1979).

322. Kazakov, V. Why did the heated discussion arise between Erb and Landouzy–Dejerine concerning the priority in describing the facio-scapulo-humeral muscular dystrophy and what is the main reason for this famous discussion? *Neuromuscul. Disord.* **11**, 421 (2001).

CHAPTER 2

Adenine base editing of the *DUX4* polyadenylation signal for targeted genetic therapy in Facioscapulohumeral muscular dystrophy

Darina Šikrová[1,#], Vlad A. Cadar[2], Yavuz Ariyurek[1,3], Jeroen F.J. Laros[1,4,5], Judit Balog[1] and Silvère M. van der Maarel[1,#]

[1]Department of Human Genetics, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
[2]Leiden University, 2300 RA Leiden, The Netherlands
[3]Leiden Genome Technology Center, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
[4]Department of Clinical Genetics, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
[5]National Institute for Public Health and the Environment (RIVM), 3721 MA Bilthoven, The Netherlands

[#]Shared corresponding authors

Šikrová et al. 2021, Molecular Therapy Nucleic Acids 1(25):342-354

Abstract

Facioscapulohumeral muscular dystrophy (FSHD) is caused by chromatin relaxation of the D4Z4 repeat resulting in misexpression of the D4Z4-encoded *DUX4* gene in skeletal muscle. One of the key genetic requirements for the stable production of full-length *DUX4* mRNA in skeletal muscle is a functional polyadenylation signal (ATTAAA) in exon three of *DUX4* that is used in somatic cells. Base editors hold great promise to treat DNA lesions underlying genetic diseases through their ability to carry out specific and rapid nucleotide mutagenesis even in postmitotic cells such as skeletal muscle. In this study, we present a simple and straightforward strategy for mutagenesis of the somatic *DUX4* polyadenylation signal by adenine base editing in immortalized myoblasts derived from independent FSHD-affected individuals. We show that mutating this critical cis regulatory element results in downregulation of *DUX4* mRNA and its direct transcriptional target genes. Our findings identify the somatic *DUX4* polyadenylation signal as a therapeutic target and represent the first step towards clinical application of the CRISPR/Cas9 base editing platform for FSHD gene therapy.

Introduction

Facioscapulohumeral muscular dystrophy (FSHD; MIM158900) is a hereditary skeletal muscle disorder that typically becomes manifest around the second decade of life and which progresses with high inter- and intra-familial variability.[1–3] It is believed that this variability in disease progression and severity can be partially explained by the underlying epigenetic mechanism of the disease, being a failure to establish and/or maintain a repressive chromatin structure of the D4Z4 macrosatellite repeat at 4q35 in somatic cells. This leads to a variegated expression of the D4Z4 repeat-encoded *DUX4* gene in muscle cells.[4] DUX4 is a pioneer transcription factor which under physiological conditions is expressed in keratinocytes,[5] testes,[4] thymus[6] and in cleavage stage embryos where it drives zygotic genome activation.[4,7–9] When mis-expressed in muscle cells, it disrupts, among others, the *bona fide* muscle transcriptome.[10,11]

The repressive chromatin environment of the D4Z4 locus in somatic cells is likely established by a repeat-mediated epigenetic silencing mechanism which partly depends on the D4Z4 repeat unit copy-number.[12] There are two genetically distinct but overlapping forms of FSHD: FSHD type 1 (FSHD1) and FSHD type 2 (FSHD2).[13,14] The more common form FSHD1 is caused by a shortening of the D4Z4 repeat to a size of 1-10 units,[15] whereas in FSHD2, the repeat size is within the lower range of healthy individuals (9-20 D4Z4 units). In the latter case, *DUX4* de-repression is caused by a malfunction of D4Z4 chromatin modifiers.[16–18] Most FSHD2 individuals can be explained by heterozygous mutations in the gene encoding for the Structural Maintenance of Chromosomes flexible Hinge Domain-Containing protein 1 (SMCHD1),[17] a protein involved in, among other pathways, epigenetic inactivation of the X chromosome in mammals.[19–23] A small number of *SMCHD1* mutation-negative FSHD2 families have been reported in which mutations in the genes encoding for the chromatin modifiers DNA Methyltransferase 3B (DNMT3B) or Ligand Dependent Nuclear Receptor Interacting Factor 1 (LRIF1) were shown to cause D4Z4 chromatin relaxation and *DUX4* expression in skeletal muscle.[16,18]

In addition to D4Z4 chromatin relaxation, the genetic background of the 4q subtelomere is critically important for FSHD manifestation. There are two equally common variants of this subtelomere, termed 4qA and 4qB,[24] however, only the 4qA variant is associated with the disease.[25,26] This is due to a sequence difference immediately distal to the distal D4Z4 unit where the 4qA allele contains an additional 260bp sequence termed pLAM which creates the third exon of *DUX4* with a functional ATTAAA polyadenylation signal (PAS) in somatic cells. Such genetic prerequisite for developing FSHD is supported by the finding that a contraction of the highly homologous D4Z4 repeat on chromosome 10 (10q26) does not lead to FSHD despite the presence of the pLAM sequence. However, this sequence contains a single nucleotide polymorphism in the corresponding *DUX4* PAS sequence (ATTAAA → ATCAAA) which renders it non-functional.[27] The critical importance of this *DUX4* PAS sequence was recently corroborated with the identification of two chromosome 10q-linked FSHD families in which the distal end of the disease-associated contracted D4Z4 repeat on

chromosome 10, including the pLAM sequence, originated from chromosome 4.[28] Likewise, 4qB chromosomes lack the pLAM sequence altogether and consequently, a D4Z4 repeat contraction on this genetic background does not lead to the development of FSHD either.[26] Previously, it has been shown by different approaches, including the application of antisense oligonucleotides, DNA nucleases and U7 snRNA, that interference with the usage of the endogenous 4qA *DUX4* PAS in myogenic cells derived from FSHD patients results in transcriptional downregulation of *DUX4* and its target genes,[29–33] further emphasizing the necessity of the annotated 4qA *DUX4* PAS for proper 3' end processing of *DUX4* pre-mRNA and suggesting that interfering with its usage is sufficient to alleviate the FSHD expression signature in myogenic cells.

Currently, there is no cure for FSHD and because of the underlying genetic character of the disease, CRISPR/Cas9 genome editing could be a promising tool to treat FSHD. Unfortunately, due to repetitive nature of the *DUX4* gene (every D4Z4 units contains one copy of the *DUX4* ORF), a straightforward Cas9 nuclease-mediated knock-out strategy might lead to multiple breaks, trigger genomic instability and result in cell death as has been shown for targeting multicopy genomic regions.[34] Therefore, a different approach is required. The novel RNA-programmable base editing system, which consists of a wild-type tRNA adenosine deaminase (TadA) and an artificially evolved version of TadA (TadA*) fused as a dimer to the D10A nicking version of *Streptococcus pyogenes* Cas9 (nSpCas9), hereafter referred to as nSpABE, enables robust adenine to guanine substitution without reliance on homology-directed repair (HDR) or introduction of double-stranded DNA breaks.[35] Such editing system has already been shown to faithfully edit the desired nucleotides also in postmitotic cells such as neurons[36] or skeletal muscle cells.[37,38] In this study, we aimed to take advantage of this system by demonstrating that with this approach the 4qA *DUX4* PAS can be efficiently disrupted resulting in downregulation of *DUX4* transcript levels in FSHD myogenic cells.

Results

Validation of sgRNA targeting *DUX4* polyadenylation signal in HAP1 cells

In myonuclei, the FSHD disease gene *DUX4* is transcribed from the distal unit of the D4Z4 repeat on the 4qA subtelomere where its transcripts are stabilized by a PAS in exon 3. The adjacent SpCas9 PAM site (TGG) downstream of this PAS allows for the design of an sgRNA that places the last three adenines of the *DUX4* PAS (ATTAAA) in the activity window of nSpABE (Figure 1A). To test whether this sgRNA can effectively direct the Cas9 machinery to the locus of interest, we first performed a T7E1 assay on HAP1 cells transfected with the sgRNA and a human codon-optimized SpCas9 nuclease. Despite having a repeat of 25 D4Z4 units on chromosome 4 which is most probably compacted into a dense chromatin structure perhaps hindering the interaction of the DNA with CRISPR/Cas9, we could clearly detect cleavage of the intended locus (Figure 1B). To evaluate A→G base editing of the *DUX4* PAS, we used a one-vector system for delivery of all adenine base editing components. HAP1 cells were individually transfected with two variants of the all-in-one vector in which

the CAG promoter drives expression of the SpCas9 nickase fused either to the ABE7.10 or the ABEmax version of the adenine base editor, hereafter referred to as nSpABE7.10 and nSpABEmax, respectively (Figure 1C). After puromycin selection, cells were examined for base editing at the *DUX4* PAS site by Sanger sequencing. In nSpABE7.10-transfected cells, we could detect on average 11,2 ± 3,6% of A→G conversion for the adenine at position 4 of the protospacer (A_4) as assessed by Sanger sequencing. We did not detect editing of adenines at positions 5 to 7 ($A_{5\text{-}7}$) despite these adenines still fitting into the reported activity window of nSpABE7.10.[35]

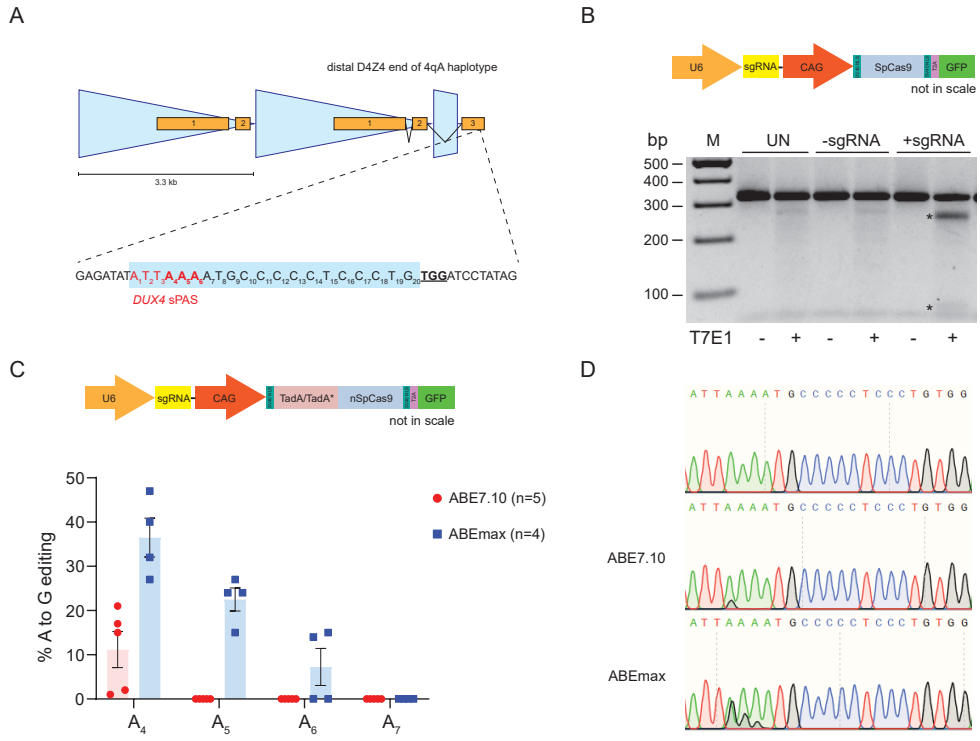


Figure 1. Adenine base editors can edit the *DUX4* PAS. A) Schematic representation of the distal end of the 4qA-derived D4Z4 macrosatellite repeat (each blue triangle represents one D4Z4 repeat unit) including the adjacent downstream sequence containing the polyadenylation signal of *DUX4* in exon 3 (*DUX4* exons are indicated by orange boxes) and zoom in on the sequence to be targeted by the adenine base editor. The sgRNA protospacer is outlined in the blue box, the PAM site for SpCas9 is underlined in bold and the *DUX4* PAS sequence (ATTAAA) is in red font with adenines that can be targeted by the adenine base editor in bold. **B)** Schematic map of the pX458 vector for simultaneous sgRNA and SpCas9 nuclease expression (top). Result of the T7E1 assay performed on HAP1 cells which were transfected with a pX458 vector expressing the sgRNA targeting the *DUX4* PAS together with a plasmid encoding for puromycine resistance to select for transfected HAP1 cells (bottom). Untransfected cells (UN) or cells transfected with no sgRNA containing vector (-sgRNA) served as negative control. Asterisks mark the T7E1 cleavage products. **C)** Schematic map of the modified all-in-one pX458 vector coding for the adenine base editor (top). Editing efficiency was assessed in HAP1 cells for the ABE7.10 and ABEmax version of the adenine base editor. The A→G editing efficiency was calculated from Sanger sequencing tracks with EditR[79] for each adenine in the editing window. Graph shows mean ±SEM of at least 4 independent biological replicates (dots). **D)** Representative Sanger sequencing tracks for ABE7.10- or ABEmax-mediated editing of the *DUX4* PAS used for quantification.

In nSpABEmax-transfected cells, we achieved more efficient adenine base editing at A_4 (36,5 ± 3,8%) as well as at downstream adenines A_5 (22,5 ± 2,25%) and A_6 (7,3 ± 3,6%), which is in agreement with a previous report that nSpABEmax is superior to nSpABE7.10 in terms of editing efficiency and processivity.[39]

Next, we assessed adenine editing of the *DUX4* PAS using the ABEmax in combination with two other Cas9 orthologues, SaCas9 and CjCas9, since their cognate PAM sites, NNGRRT and NNNVRYM, respectively, are in the vicinity of the *DUX4* PAS such that adenines on the forward or reverse strand in the *DUX4* PAS could be amenable to adenine base editing (Suppl. Figure 1A). We used the same all-in-one vector architecture as was used for nSpABEmax, including the same linkers length and the new constructs are hereafter referred to as nSaABEmax and nCjABEmax (Suppl. Figure 1B). Surprisingly, both constructs failed to exert adenine base editing activity at the *DUX4* PAS in HAP1 cells based on evaluation by Sanger sequencing as was done for SpABE7.10 and SpABEmax (data not shown).

Base editing of *DUX4* PAS in patient-derived immortalized FSHD1 and FSHD2 myoblasts

To explore the effect of the mutated PAS on *DUX4* steady-state transcript levels, we carried out base editing in FSHD patient-derived immortalized myoblasts since HAP1 cells do not express *DUX4*. We used three different FSHD myogenic cell lines with different genetic characteristics, D4Z4 methylation status and *DUX4* expression levels (Suppl. Figure 2A and 2B). We selected one FSHD2 cell line which has a heterozygous missense mutation in SMCHD1 (K204E) combined with an 11 units long D4Z4 repeat on 4qA and two FSHD1 cell lines, one with a 3 units long D4Z4 repeat ($FSHD1^{3U}$) and one with an 8 units-long 4qA repeat ($FSHD1^{8U}$). Shorter D4Z4 repeats are generally correlating with lower D4Z4 methylation levels,[40] a more severe FSHD phenotype and a worse prognosis,[2] whereas repeats in the upper size limit of FSHD1 typically show a higher incidence of familial non-penetrance and a milder disease presentation.[3,41] Furthermore, we chose cell lines heterozygous for 4qA and 4qB to facilitate unequivocal assignment of successful editing of the FSHD allele, except for $FSHD1^{8U}$ which carries two variant alleles of 4qA (with the healthy allele being of the 4qA161L variant and the FSHD allele of the 4qA161S variant).[42] However, these two allelic variants of 4qA161 can be distinguished by the presence of a SNP (Suppl. Figure 3A). Clonal cell cultures from all three cell lines were genotyped for the *DUX4* PAS after transfection with nSpABEmax and single cell sorting of GFP^+ cells. Untransfected cells underwent the same sorting procedure to obtain clones with a WT PAS sequence to ensure the same experimental conditions and population doublings between compared groups. Successfully edited clones showed a plethora of A→G editing outcomes (Suppl. Figure 3A). We also obtained one clone from the $FSHD1^{3U}$ and one clone from the FSHD2 cell line in which the editing attempt resulted in small deletions fully or partially involving the *DUX4* PAS (Suppl. Figure 3A). *DUX4* steady-state mRNA levels were measured as well as that of four well-established DUX4 target genes (*ZSCAN4*, *KHDC1L*, *TRIM43* and *MBD3L2*)[11,43] serving as an indirect readout for DUX4 transcription factor activity. The steady-state mRNA levels of *DUX4* and its target genes were

reduced in all three cell lines upon editing of the *DUX4* PAS under proliferating (Suppl. Figure 3B) and differentiating conditions (Figure 2A). Since it has been shown that *DUX4* expression increases during myogenic differentiation,[44] we analysed the expression of early (*MYOG*) as well as late (*MYH3*) myogenic markers by RT-qPCR to rule out the possibility that lower *DUX4* levels were due to reduced differentiation potential of edited clones (Figure 2B). On the contrary, it seemed that edited clones show equal if not slightly increased myogenic differentiation which is in agreement with previous findings that DUX4 inhibits myogenic differentiation.[10] However, unedited clones showed a high variability in *DUX4* expression levels and that of its target genes ranging from one order of magnitude in the FSHD1^{3U} and FSHD2 line up to 3 orders of magnitude in clones derived from the FSHD1^{8U} line. Such high expression variability makes it difficult to confidently determine the effect of *DUX4* downregulation conferred by base editing.

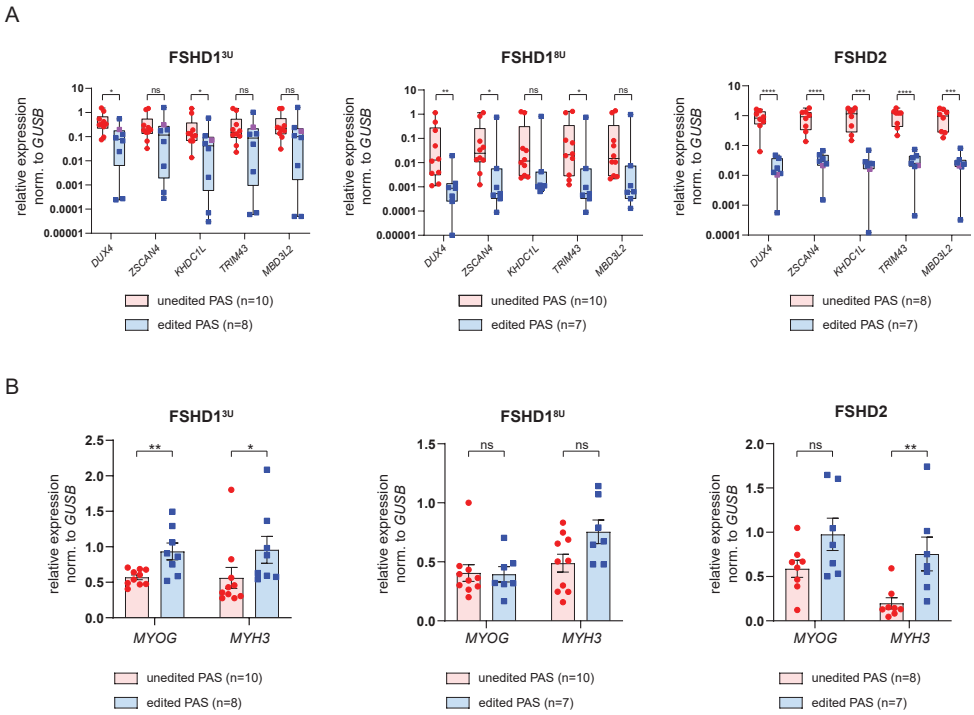


Figure 2. Adenine base editing of the *DUX4* PAS reduces expression of *DUX4* and its target genes in FSHD myogenic cells derived from polyclonal cultures. A) mRNA levels as assessed by RT-qPCR of *DUX4* and four DUX4 target genes (*MBD3L2*, *ZSCAN4*, *TRIM43* and *KHDC1L*) in PAS unedited vs edited clones derived from two FSHD1 and one FSHD2 cell lines differentiated into myotubes. Statistical significance was calculated with unpaired two-tailed t-test (ns: non-significant, *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001) on log2 transformed expression values to correct for skewed distribution. Expression values normalized to *GUSB* as house-keeping gene are plotted. Line represents mean and whiskers represent min and max value. Individual dots represent individual clones, the two violet clones carry a deletion affecting the *DUX4* PAS. **B)** mRNA levels of two myogenic markers (*MYOG* and *MYH3*) for all unedited and edited clones of all three FSHD cell lines are plotted. Statistical significance was calculated with unpaired two-tailed t-test (ns: non-significant, *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001). Bars represent mean ±SEM with individual clones expression values plotted as individual dots.

Reducing the clonal variability in *DUX4* expression

Since D4Z4 displays highly variable transcriptional activity between individuals[45] and across cells from the same individual (this study), a behaviour which is also described for genomic loci known as metastable epialleles[46] of which their epigenetic profile is stochastically established in early embryogenesis, we hypothesized that starting the editing from a monoclonal cell culture rather than a polyclonal culture may resolve large part of inter-clonal variability in *DUX4* expression. This would facilitate a better comparison of *DUX4* levels between *DUX4* PAS pre-editing and post-editing clones in the absence of large expression variability at WT baseline. We therefore first tested the "mitotic stability" of *DUX4* expression by deriving new daughter clones from two clones showing different levels of *DUX4* expression (referred to as DUX4^{high} and DUX4^{low}) originating from the FSHD1^{8U} line as it showed the highest *DUX4* expression variability. Indeed, after resorting, new single-cell derived cultures exhibited more homogeneous *DUX4* and DUX4 target genes (*ZSCAN4* and *MBD3L2*) expression levels comparable to the parental clone as measured by RT-qPCR (Suppl. Figure 4).

We selected one unedited DUX4^{high} clone derived from either the FSHD1^{3U} or the FSHD1^{8U} cell line and repeated the editing procedure to obtain new *DUX4* PAS unedited and edited clones. As expected, deriving new unedited clones from a monoclonal culture resulted in lower *DUX4* expression variability between clones with clones carrying an edited *DUX4* PAS showing significantly reduced *DUX4* steady-state mRNA levels as well as DUX4 target gene levels (Figure 3A). Again, the reduced *DUX4* expression levels could not be attributed to a difference in myogenic differentiation as shown by comparable expression of the two myogenic differentiation markers between edited and unedited clones (Figure 3B). Interestingly, editing the *DUX4* PAS seems to have a more negative impact on *DUX4* mRNA levels in FSHD1^{8U} (approximately 1000-fold downregulation) than in cells from FSHD1^{3U} line (approximately 10-fold downregulation).

Editing of the *DUX4* PAS induces alternative pre-mRNA cleavage and polyadenylation

Previously, it was shown that hindering the *DUX4* PAS with phosphorodiamidate morpholino oligomers (PMOs) causes a redirection of the *DUX4* pre-mRNA cleavage site (CS) ~40 nt upstream of its canonical CS despite the absence of a recognizable alternative PAS motif in the upstream sequence.[31] Since base editing of the *DUX4* PAS does not completely abolish *DUX4* expression, we tested if the mutated PAS is still being used for *DUX4* transcript termination, albeit less efficiently, or if alternative PAS/CS are being used. Using a semi-quantitative 3'RACE to identify 3' UTR sequences of *DUX4* mRNAs from unedited and edited clones derived from all three FSHD immortalized cell lines from Figure 2A, we detected three different CSs 16-24 nt downstream of *DUX4* PAS in close proximity to each other in unedited cells (Figure 4A), as was previously described.[31] In edited clones, however, two different shifts in the CS occur, either proximally or distally to the canonical CS. Interestingly, the FSHD2 edited clones strictly used the proximal CS, the same one as reported by Marsollier et al.[31] after using PMOs against the *DUX4* PAS region, whereas the distal CS switch is

predominant in the FSHD1 clones independent of their 4qA permissive allele size (Figure 4B). Moreover, opposite to the single proximal CS being used after PAS editing, the distal CS is not as deterministic, since we observed multiple different 3' ends in FSHD1 edited clones.

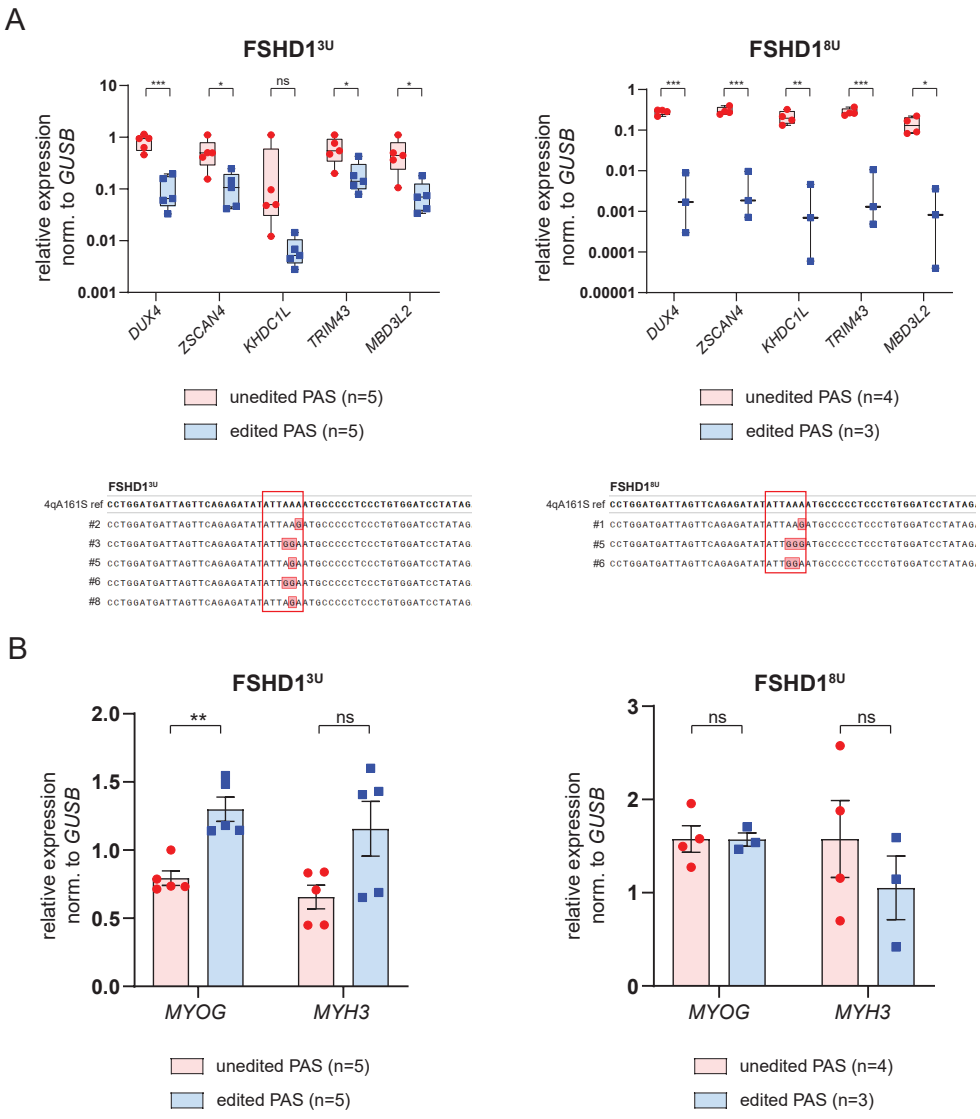


Figure 3. Adenine base editing of the *DUX4* PAS reduces expression of *DUX4* and its target genes in FSHD myogenic cells derived from monoclonal cultures. A) mRNA levels as assessed by RT-qPCR of *DUX4* and four DUX4 target genes (*MBD3L2*, *ZSCAN4*, *TRIM43* and *KHDC1L*) in *DUX4* PAS unedited vs edited subclones derived from two clones with different FSHD1 cell line origins (top). Genotypes of edited clones aligned to the reference WT sequence with the *DUX4* PAS highlighted in a red rectangle and red colored bases denote mismatches (bottom). **B)** mRNA levels of two myogenic markers (*MYOG* and *MYH3*) for unedited and edited clones from A). Statistical significance was calculated with unpaired two-tailed t-test (ns: non-significant, *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001). Bars represent mean ±SEM with individual clones expression values plotted as individual dots.

Of note, the small proportion of *DUX4* mRNAs using the canonical CS position in FSHD2 clones is coming from the single clone which carries a partial deletion of *DUX4* PAS. Despite the clear shift in the CS upon *DUX4* PAS editing, we could not detect a nearby PAS-like sequences (±100 nt from original PAS) which could explain the CS shifts. Overall this data show that *DUX4* PAS base editing prevents proper 3' end formation of the *DUX4* transcript.

Off-target analysis by targeted next generation sequencing

To explore potential off-target effects, we used the CRISPOR prediction tool[47] to identify genomic sites that have a sequence homology to the sgRNA used for targeting the *DUX4* PAS. This resulted in the identification of 227 potential off-target (OT) sites, of which none are predicted to target polyadenylation signals of other genes. Only 3 are predicted to target coding sequences, however, with low off-target scores due to the number and position of individual mismatches (Suppl. Table 4). We further filtered predicted off-target sites by the following criteria: i) having up to 4 mismatches outside of the PAM region and the seed region of the sgRNA, ii) containing at least one adenine in the editing window of nSpABEmax, and iii) representing a single copy locus. Based on these criteria, we performed targeted next generation sequencing on 10 selected potential off-target sites in DNA samples obtained from HAP1 cells that were transfected with nSpABEmax with or without sgRNA targeting the *DUX4* PAS from Figure 1C and D (Figure 5A). At 7 out of 10 examined sites, deep sequencing did not reveal any appreciable increase in A→G transitions within or near the editing window as compared to the control samples (Figure 5B). However, the nucleotide sequences of OT1 and OT10 contained a SNP in the HAP1 genome, producing an extra mismatch in the sgRNA protospacer (Figure 5A). Therefore, their off-target potential might be higher in genomes that do not contain this mismatch. At three sites, OT2 (chr6: 13,331,126-13,331,148), OT5 (chr12: 2,444,719-2,444,741) and OT6 (chr2: 218,831,310-218,831,332), we detected editing efficiencies of 0.17 %, 1,72 % and 0.43 % of adenosines within the editing window, respectively (Figure 5B, C). None of the three affected OT sites reside in coding regions. OT2 is in an intergenic region approximately 2 kb upstream of the *TBC1D7* gene, while OT5 and OT6 map to intron 3 of *CACNA1C* and intron 2 of *PRKAG3*, respectively. Both genes, *CACNA1C* and *PRKAG3*, are expressed in skeletal muscle according to the Human Protein Atlas[48] but neither edits are predicted to affect the splicing of these genes when modelled with the Alamut software. In summary, these results show that sgRNA-dependent off-target DNA editing is likely rare.

Discussion

So far, therapeutic attempts for FSHD have been mainly focused on oligonucleotide- or small molecule-based transient modulation of *DUX4* levels.[49,50] Three recent studies focused on gene therapy approaches that inhibit the production of full-length *DUX4* mRNA.[32,33,51] Two of these studies used CRISPR/Cas9 strategies, either employing a standard Cas9 nuclease to introduce deletions affecting the *DUX4* PAS by homology-directed repair with a provided template[32] or using Cas9 coupled to a transcriptional inhibitor domain to repress *DUX4* expression[51].

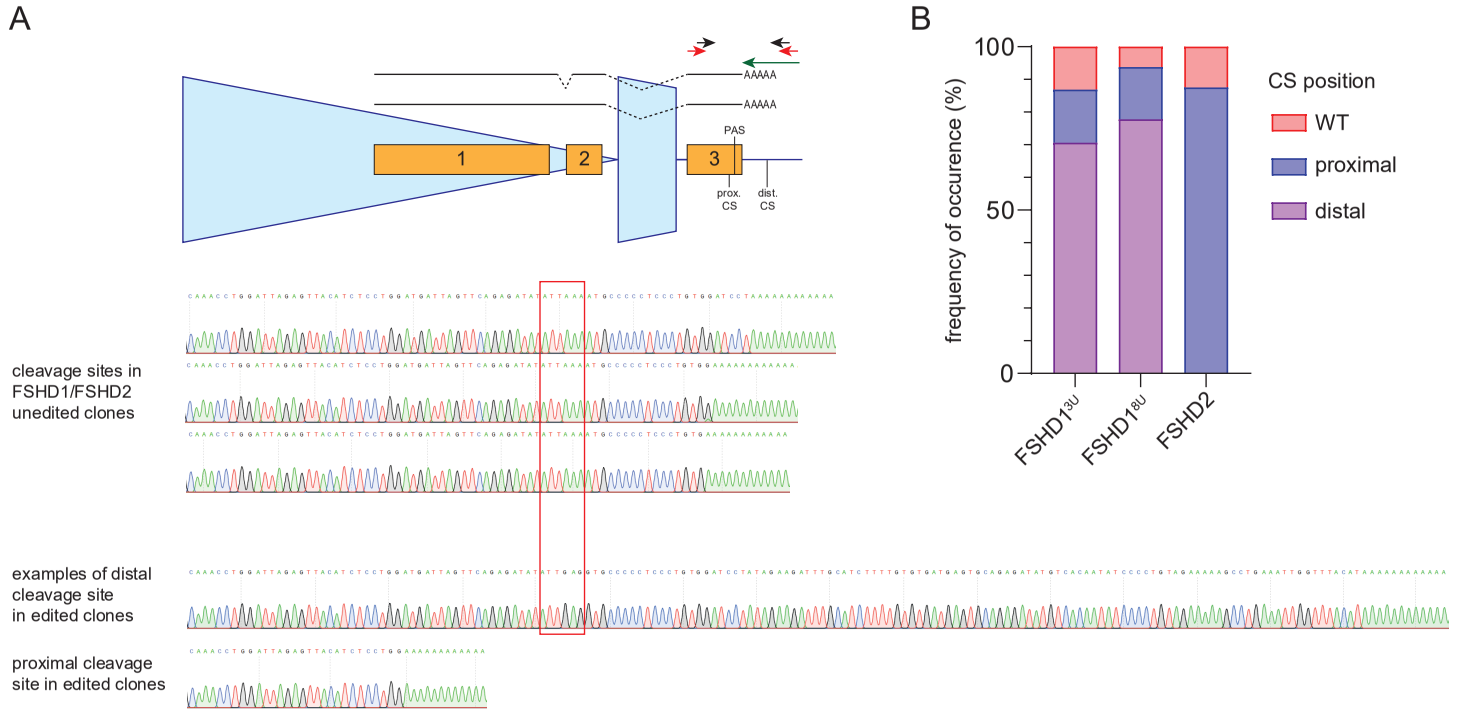


Figure 4. Editing of the *DUX4* PAS induces alternative pre-mRNA cleavage site A) Schematic of the terminal D4Z4 repeat unit with short ending (4A161S haplotype) showing the design of 3'RACE experiment to determine the cleavage and polyadenylation site of *DUX4* mRNA in the edited clones. Two known *DUX4* mRNA isoforms are depicted with splicing or retention of intron 1 (top). Arrows represent primers used for oligo-dT reverse transcription (green), first PCR (red) and second nested PCR (black). The identified proximal and distal cleavage sites, for which Sanger sequencing traces are provided, are marked. Sanger sequencing tracks (bottom) show representative examples of 3' ends of *DUX4* mRNA in *DUX4* PAS unedited and edited FSHD1/FSHD2 clones. The red rectangle outlines the *DUX4* PAS sequence. Three different CSs were identified in unedited clones (as reported previously[31]), while different shifts in CSs were identified in edited clones. One representative Sanger sequencing track is shown for each CS choice. **B)** Barplots representing the frequency of occurrence of different CSs identified in *DUX4* PAS edited clones with respect to WT CSs from ≥4 clones for each condition.

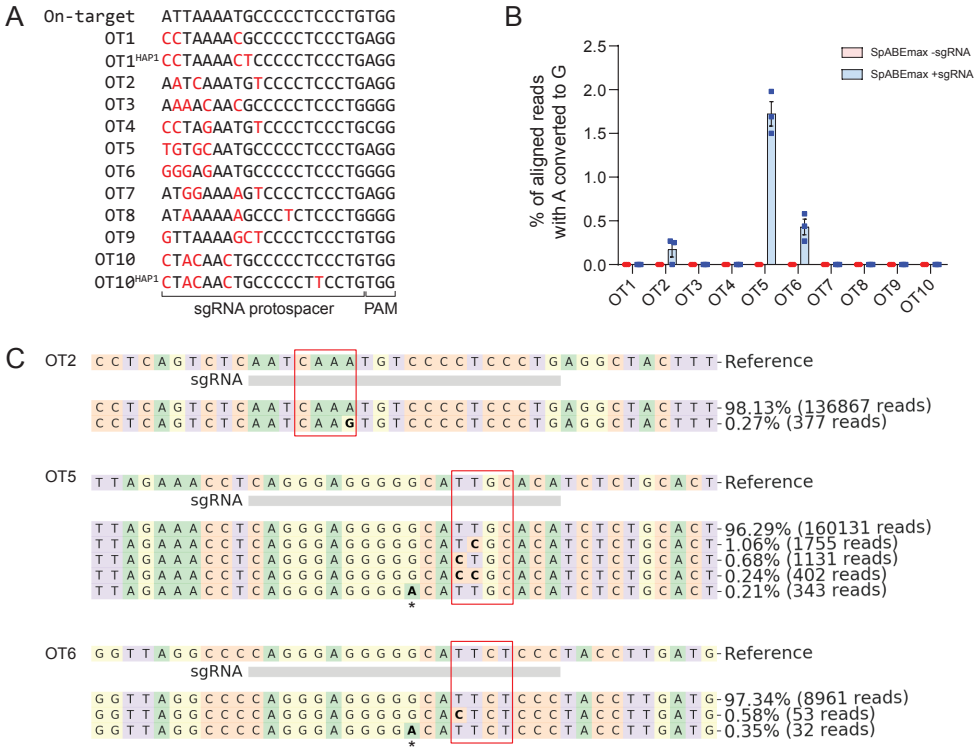


Figure 5. sgRNA-dependent off-target analysis in HAP1 cells. A) DNA sequences of 10 predicted off-target sites identified by CRISPOR.[47] Nucleotide mismatches compared to the *DUX4* PAS target sequence are highlighted with red font. Two off-target sites (OT1 and OT10) carried an extra mismatch in HAP1 cells as compared to the reference sequence obtained from GRCh38. **B)** Editing frequencies at predicted off-target sites were assessed in HAP1 cells which were transfected with nSpABEmax and either with or without *DUX4* PAS targeting sgRNA. The A->G editing efficiency was assessed by amplicon next generation sequencing and analysed with CRISPResso2.[80] Graph shows mean ±SEM of 3 independent biological replicates. **C)** Representative allele frequencies of three off-target sites (OT2, OT5 and OT6) with the highest editing outcome are shown. OT5 and OT6 sequences are shown in forward orientation, while sgRNA targets the reverse complement strand. The editing windows are highlighted in the red box. Only allele frequencies of at least 0.1% were considered. The mutation rate in the G homopolymer (marked by asterisk) preceding the editing window was not included in the editing frequency calculation plotted in **B)** since it occurred also in the control samples and was more likely introduced either during PCR steps or Illumina sequencing itself rather than in an sgRNA-dependent fashion.

The third study used custom U7 nuclear RNAs (snRNA) to mask important regulatory features of *DUX4* mRNA maturation such as splice sites and the *DUX4* PAS.[33] In this study, we demonstrate the use of a CRISPR/Cas9-based genome editing application to directly modify the *DUX4* locus while avoiding DNA double-strand breaks. We show that by using an adenine base editor we can target and disable one of the important genetic prerequisites for FSHD manifestation – the *DUX4* somatic polyadenylation signal. We were able to successfully edit the *DUX4* PAS with SpCas9-based base editors nSpABE7.10 and nSpABEmax, with the latter showing higher editing efficiency which is in agreement with previous reports.[52] Fusing ABEmax to two other Cas9 orthologues, namely SaCas9 and CjCas9, has previously been

shown to also result in adenine editing activity.[52–54] However, we did not observe adenine to guanine conversion at the *DUX4* PAS when using such fusion proteins in HAP1 cells as determined by Sanger sequencing. The T7E1 assay nevertheless did show evidence for recruitment of the SaCas9 nuclease to the *DUX4* PAS site (Suppl. Figure 1C) suggesting that the complex can be recruited to the *DUX4* PAS but that the nSaABEmax fusion protein is likely not efficient at this site. Previously, a lower editing efficiency has been reported for nSaABEmax as compared to nSpABEmax[52] which could explain our findings. Recently, a new version of the adenine base editor, termed ABE8e, was described.[53] When paired with a variety of Cas effectors, including SaCas9, it demonstrated a further enhanced editing efficiency. Therefore, coupling ABE8e to SaCas9 might result in successful adenine base editing of the *DUX4* PAS. In addition, such fusion construct would be more favourable compared to SpCas9 construct due to its smaller size, which could facilitate the use of the AAV system for its *in vivo* delivery and testing. Alternatively, an AAV split system could be used for *in vivo* delivery of SpABEmax or SpABE8e. Indeed, such approach has been already tested for delivering base editors to a range of tissues[37,55] reaching 20% editing efficiency in skeletal muscle tissue.[55] Since published strategies were aiming at whole body delivery and were not optimized for skeletal muscle targeting nor expression, further optimization by using a tissue-specific promoter and a muscle-trophic AAV serotype might increase the editing efficiencies in the skeletal muscle. On the other hand, the failure to detect editing of the *DUX4* PAS with nCjABEmax might be attributed to a suboptimal nearby PAM sequence (5'-AATCATC-3') that was predicted for the targeting. We identified this PAM site based on the PAM consensus sequence (5'-NNNVRYM-3') reported by Yamada M. et al..[56] Another study by Kim E. et al.[57] reported a slightly different PAM consensus (5'-NNNNRYAC-3') for CjCas9 targeting which is more refined and differs from the sequence that we used for deriving our sgRNA. Moreover, such fusion construct has not been characterized in depth yet, therefore there is no knowledge about its precise editing window nor its efficiency.

As anticipated, editing of the *DUX4* PAS in immortalized myogenic lines obtained from different FSHD-affected individuals resulted in lower *DUX4* mRNA levels and lower DUX4 transcription factor activity as indirectly measured by the steady-state mRNA levels of its target genes. We could not determine if editing more adenines at once or if editing an adenine at a particular position in the *DUX4* PAS motif results in a more profound *DUX4* downregulation, since multiple clones with the same editing outcome would be required to confidently assess this. Nevertheless, we show that even a single adenine substitution is sufficient to negatively impact proper 3' end processing of the *DUX4* transcript. To our surprise, mutating the *DUX4* PAS in this manner does not completely abolish the production of polyadenylated *DUX4* transcripts as opposed to the situation on chromosome 10, which might suggest the presence of other cis modifiers acting as regulators of *DUX4* expression than just the previously recognized SNP in 4q/10q *DUX4* PAS motif. These cis factors are likely in linkage disequilibrium with the *DUX4* PAS considering the exclusive linkage of FSHD with the presence of a *DUX4* PAS. Interestingly, in two independent FSHD clonal cell lines we observed different steady-state *DUX4* mRNA levels reduction upon editing (Figure 3A). Since we cannot correlate this outcome to either the initial *DUX4* expression levels, to the nucleotide edit at the *DUX4* PAS, or to the methylation

levels at the targeted region, this outcome may be a reflection of its suspected role as metastable epiallele, as the chromatin environment has also been suggested to influence PAS usage efficiency.[58,59] Such individualistic response will require further studies to elucidate its mechanism and to be able to predict the benefit of this approach for FSHD patients.

In addition, the study by Joubert et al. reported the use of either paired transcription activator-like effector nucleases (TALENs) or paired CRISPR/Cas9 nucleases to excise the *DUX4* PAS sequence with the aim to incorporate a mir-1 sequence by homology-directed repair in immortalized myoblasts.[32] This approach yielded only two successfully edited clones out of 227 (0,8%). In contrast, with our approach we achieved 30/163 successfully edited immortalized myoblast clones (nearly 20%) across five different experiments including three different FSHD cell lines (Suppl. Table 5). Nevertheless, despite the limited number of successfully edited clones in the Joubert study, they also observed reduced, but not abolished, *DUX4* and DUX4 target gene levels and a switch in the *DUX4* mRNA cleavage and polyadenylation site which corroborates our findings. The increased editing efficiency in our study could be explained by the fact that adenine base editors act independently of the homology-directed repair pathway, a pathway that is only available in S and G2 phase of the cell cycle. This cell cycle-independent feature of the ABE system makes it a viable candidate for its future *in vivo* translatability. The main bottleneck for adenine editing efficiency may therefore very well be the optimal delivery of editing components to skeletal muscle tissue.

One of the main concerns for the use of genome editing platforms is their potential off-target effect. Adenine base editors have been shown to suffer from sgRNA-dependent off-target DNA editing, albeit to a lesser extent than cytidine base editors.[60] In this study, we detected at least three sites that were edited in an sgRNA-dependent fashion but to a much lesser extent than the intended site. We observed approximately 23-fold more efficient editing at the A_4 position of the on-target site, i.e. 40% as assessed by Sanger sequencing in gDNA samples which were used also for the inspection of off-target editing in HAP1 cells, as compared to the most efficiently edited off-target site (OT5, 1,7%) as assessed by Illumina short read sequencing. Additionally, off-target editing of cellular RNAs by adenine base editors has been reported.[61] However, we have not explored this particular side effect of nSpABEmax. In any case, both DNA and RNA off-target activity of adenine base editors can be minimized by making use of further engineered adenine deaminases[53,62,63] linked to higher fidelity Cas9 versions,[64–66] modified sgRNAs,[67] and by reducing exposure time and/or effector molecule concentrations by employing different delivery strategies such as in the form ribonucleoprotein particles.[53,68] The specificity of the adenine base editing approach for *DUX4* PAS targeting should be therefore carefully evaluated to ensure the safety in case of its therapeutic application.

Base editors have been already used to achieve efficient gene silencing by targeting cis regulatory elements important for proper gene expression by either introducing in-frame stop codons,[69,70] mutating a start codon[71] or by disrupting splice sites.[72,73] Since deviations from the canonical PAS hexamers generally reduce their cleavage and polyadenylation efficiency,[74] we explored how many polyadenylation signals genome-wide would be amenable for such

editing approach. We focused on the two most widely used hexameric motifs, namely AATAAA and ATTAAA, as they constitute around 80% of all identified polyadenylation signals (Suppl. Figure 5A). These PAS motifs can be disrupted with adenine base editors either by modifying any of the adenines of the last three nucleotide positions of the PAS motif on the coding strand or alternatively by targeting the adenine on the non-coding strand which pairs with middle thymine on the coding strand leading to its substitution with a cytidine (Suppl. Figure 5B). Based on these criteria, we established that approximately 25% of all PASs with either AATAAA or ATTAAA motifs are editable with nSpABEmax (Suppl. Figure 5C). However, it should be pointed out that weakening the core PAS motif might not always lead to the expected transcriptional downregulation since other cis auxiliary elements are known to influence the efficiency of PAS usage.[75] Moreover, alternative polyadenylation is widespread for genes which contain multiple functional PASs,[76] therefore invalidating only one of them might not be sufficient to achieve an overall desired level of silencing. Rather, since alternative polyadenylation is tissue-specific and globally regulated, PAS editing might represent a more refined tool for gene editing in some conditions. Therefore, the utility of this approach requires locus-specific validation. Nevertheless, due to challenging gene structure, *DUX4* represents an excellent candidate for adenine base editing-mediated mutagenesis of its PAS as a mean for its expression interference.

Materials and Methods

Cloning

To create the all-in-one base editing vector pX458-ABE7.10, overlapping PCR products of the TadA dimer from pCMV-ABE7.10 (Addgene #102919), nCas9-SV40 NLS from pX335 (Addgene #42335) and T2A-GFP from pX458 were cloned in pX458 using the AgeI and EcoRI restriction sites. The pX458-ABEmax vector was created by cutting out the TadA dimer together with N-terminal domain of Cas9 from pX458-ABE7.10 using the AgeI and ApaI sites and replacing it with the PCR amplified TadA dimer missing the N-terminal domain of Cas9 from the pCMV-ABEmax-GFP vector (Addgene #112101). The pX601-SaABEmax vector was cloned by first creating a new insert consisting of the TadA dimer linked to the N-terminal domain of SaCas9. This was achieved by overlapping PCR amplifications on pCMV-ABEmax (for the TadA dimer) and pX601 (for the SaCas9 domain) during which a D10A mutation was introduced into SaCas9. The resulting PCR product was cloned in pX601 using the XbaI and HindIII sites. The pX601-CjABEmax was created by first mutating the KpnI site upstream of the CAG promoter in the pX601-SaABEmax vector by replacing it with the same PCR fragment containing a KpnI mutation and cloned using XbaI and AgeI. Next, the SaABEmax-T2A-GFP-bGH insert was replaced by CjABEmax-T2A-GFP-bGH, which was produced by overlapping PCRs on pX601-SaABEmax for TadA dimer, pX404 (Addgene #68338) for CjCas9 (D8A mutation was introduced during this PCR step) and on pX601-SaABEmax for T2A-GFP-bGH PAS. The final insert was cloned into pX601-SaABEmax via the AgeI and KpnI sites. Further, the SaCas9 sgRNA expression cassette was replaced with an CjCas9 sgRNA expression cassette. The CjCas9 sgRNA expression cassette was assembled by overlapping PCRs on pX601 to amplify the U6 promoter sequence and on the pU6-Cj-sgRNA plasmid (Addgene #89753) to amplify the sgRNA scaffold. The resulting insert was cloned into the pX601-CjABEmax plasmid created in the previous step via the KpnI and NotI sites. All sgRNAs were cloned into their target vector according to the Zhang's lab protocol.[77] For the pX458 vector (Addgene #48138) and its adenine base editor derivatives (SpABE7.10 and SpABEmax), the BbsI sites were used and for pX601 vector's derivatives (SaABEmax and CjABEmax) the BsaI sites were used. For optimal transcription from the U6 promoter, an extra G nucleotide was added to the 5' end of the sgRNA in case the sequence did not start with one already. All constructs were verified by Sanger sequencing. All used primers are listed in Suppl. Table 1. The following restriction enzymes were used for cloning: AgeI-HF (New England Biolabs, ##R3552), EcoRI (Thermo Fisher Scientific, #ER0271), ApaI (New England Biolabs, #R0114), HindIII (New England Biolabs, #R0104), KpnI-HF (New England Biolabs, #R3142), NotI-HF (New England Biolabs, #R3189), BbsI (New England Biolabs, #R3539), BsaI (Thermo Fisher Scientific, #ER0291).

Cell culture and transfection

The HAP1 cell line was maintained in IMDM – GlutaMAX (Thermo Fisher Scientific, #31980) supplemented with 10 % foetal bovine serum (FBS) (Gibco, #10270106) and 1 % (v/v) penicillin/streptomycin (Gibco, #15140). Immortalized myoblast cell lines 073iMB (FSHD1^{8U}) and 200iMB (FSHD2) were a kind gift from Prof. S. Tapscott, Fred Hutchinson Cancer Research Center. The 2402iMB line (FSHD1^{3U}) was obtained by immortalizing primary myoblasts, which were a kind gift of Prof. R. Tawil from University of Rochester, by stable integration of hTERT and CDK4 retroviruses as described previously.[78] All myogenic lines were maintained in Ham's F-10 Nutrient Mix (Gibco, #31550) supplemented with 20 % (v/v) FBS, 1 % (v/v) penicillin/streptomycin, 10 ng/ml FGF-b (Promokine, #C-60240) and 1 µM dexamethasone (Sigma-Aldrich, #D2915). Myogenic differentiation was achieved by switching myoblasts at 100% confluency to DMEM (Gibco, #31966021) supplemented with 2 % (v/v) KnockOut™ serum replacement (Gibco, #10828028). All cell lines were maintained at 37°C and 5 % CO_2 and were tested for *Mycoplasma* contamination with the MycoAlert™ Mycoplasma detection kit (Lonza, #LT07-318) according to the vendor's instructions. One day prior to transfection, 2 x10^5 HAP1 cells were seeded in a 12-well plate. Transfection was performed with 1.5 µg of the base editing vector and 0.5 µg of a vector containing puromycin resistance cassette (AA19_pLKO.1-puro. U6.sgRNA.BveI-stuffer plasmid, a kind gift from Prof. M.A.F.V. Gonçalves, Leiden University Medical Center) using Lipofectamine 3000 (Thermo Fisher Scientific, #L3000008) according to the manufacturer's instructions. The next day, the media was replaced with media containing 0.5 µg/ml of puromycin and cells were selected for 48h after which the media was replaced again with non-puromycin media and cells were grown for an additional 72h after which they were harvested for subsequent analysis. For myoblasts experiments, 3 x10^5 myoblasts were seeded in a 6-well plate and the following day cells were transfected with 2 µg of plasmid DNA using Lipofectamine 3000 (Thermo Fisher Scientific, #L3000008) according to the manufacturer's instructions. Media was changed the next day and cells were harvested for further analysis 72h after transfection.

T7E1 cleavage assay

CRISPR/Cas9 induced indels at the targeted locus were examined with the T7E1 cleavage assay. Three days after transfection, cells were harvested in lysis buffer for genomic DNA (100 mM Tris-HCl pH 8.0, 50 mM EDTA pH 8.0, 2 % (w/v) SDS) and DNA was extracted by protein precipitation by adding saturated salt to the solution and subsequent isopropanol precipitation. The target locus was amplified by PCR using DreamTaq (Thermo Fisher Scientific, #EP0701) with the following cycling conditions: 95°C for 5 min followed by 35 cycles of 95°C for 25 sec, 67°C for 25 sec and 72°C for 20 sec with a final extension step at 72°C for 5 min. Resulting PCR products were subjected to re-annealing in a thermal cycler using the following conditions: 95°C for 5 min followed by cooling down from 95°C to 85°C at 2°C/sec and from 85°C to 25°C at 0.1°C/sec. After reannealing, 10 µl of PCR product was incubated with T7E1 enzyme (New England Biolabs, #E3321) according to the manufacturer's instructions. Resulting products were resolved on a 2 % TBE agarose gel with ethidium bromide.

Fluorescence-activated cell sorting

Cells were trypsinized, collected in their respective culturing media, spun down and the cell pellet was resuspended in FACS buffer (10 % v/v FBS in PBS). Cells were sorted using a BD FACS Aria™ III cell sorter according to GFP fluorescence and collected cells were used for further analysis or expansion.

***DUX4* PAS genotyping and quantification of base editing efficiency**

Exon 3 of *DUX4* containing the PAS was amplified from genomic DNA by PCR as described in the T7E1 cleavage assay. The product's purity was first assessed by an electrophoretic separation on a 2 % TBE agarose gel and then extracted from gel using the NucleoSpin Gel and PCR Clean-up kit (Bioké, #740609) and submitted for Sanger sequencing with the forward primer used in the PCR. Base editing efficiency in the initial test in HAP1 cells was assessed by Sanger sequencing and estimated with Edit-R[79] (online tool available at http://baseeditr.com/).

RNA isolation, cDNA synthesis and qPCR

Cells were harvested in QIAzol lysis reagent (Qiagen, #79306) and RNA was isolated with the RNeasy mini kit (Qiagen, #74101) with DNase I treatment according to the manufacturer's protocol. Oligo-dT primed cDNA was synthesized from 2 µg of input RNA using the Minus First Strand cDNA synthesis kit (Thermo Fisher Scientific, #K1621). Gene expression was measured with the CFX384 system (BioRad) in technical triplicates using iQ™ SYBR® Green Supermix (Biorad, #1708887). qPCR primers are listed in Suppl. Table 1. *GUSB* was used as a housekeeping gene.

3'RACE

The 3'RACE was carried out as reported previously[31] with minor modifications. The cDNA synthesis was carried out with the Minus First Strand cDNA synthesis kit with modified oligo-dT primer: 5'-GCTGTCAACGATACGCTACGTAACGGCATGACAGTGTTTTTTTTTTTTTTTTTTTTTTTT-3'. The first PCR was performed using 2 µl cDNA as template in a final volume of 20 µl using AccuPrime' *Taq* high fidelity DNA polymerase (Thermo Fisher Scientific, #1236086) with previously published forward and reverse primers and according to established PCR cycling conditions.[31] Nested PCR was performed using 2 µl of primary PCR product using AccuPrime' *Taq* high fidelity DNA polymerase with previously published forward and reverse primers and according to established PCR cycling conditions.[31] Final PCR products were purified from 2 % TBE agarose gel and subcloned into the TOPO-TA vector (Thermo Fisher Scientific, #450641). At least 6-8 individual bacterial colonies were screened to determine the *DUX4* mRNA 3' ends.

Methylation analysis of *DUX4* exon 3 (FasPAS region) by bisulfite PCR followed by TOPO-TA subcloning

500 ng of genomic DNA was converted using the EZ DNA Methylation-Lightning kit (Zymo Research, #D5030) according to the manufacturer's protocol. The FasPAS region was amplified from converted DNA with previously published primers (Suppl. Table 1) using high fidelity Accuprime™ Taq DNA polymerase (Thermo Fisher Scientific, #12346086) with the following PCR program: 95°C for 4 min followed by 35 cycles of 95°C for 4 min, 58°C for 20 sec and 72°C for 40 sec, followed by a final extension step at 72°C for 5 min. PCR products were purified by electrophoresis and isolated from gel using the NucleoSpin Gel & PCR Clean-up kit (Bioke, #740609) followed by subcloning into the TOPO-TA vector. Plasmid DNA from individual bacterial colonies was sent for Sanger sequencing using the M13R primer and methylation levels were assessed with BiQ Analyzer software. Methylation lollipop plots were produced with the online QUMA tool (http://quma.cdb.riken.jp/top/index.html).

sgRNA-dependent off-target analysis using targeted next generation sequencing

Potential off-target sites were predicted by CRISPOR (http://crispor.tefor.net/crispor.py).[47] Ten predicted off-target sites were chosen based on the MIT specificity score and uniqueness of the region for specific amplification. Genomic regions of interest were amplified with specific primers containing appropriate Illumina forward and reverse adaptor sequences (Suppl. Table 1). For the first PCR, 100 ng of genomic DNA was used as starting material in a 25 µl reaction further containing 0.4 µM of forward and reverse primer and 12.5 µl of 2x KAPA HiFi HotStart ReadyMix (Kapa Biosystems, #KK2601). PCR reactions were carried out as follows: 95°C for 3 min followed by 27 cycles of 98°C for 20 sec, 64°C for 15 sec and 72°C for 15 sec with a final extension step at 72°C for 3 min. This first PCR product was purified with AMPure beads (Beckman Coulter, #A63881) with a 0.8 PCR:beads ratio according to the manufacturer's instructions and DNA was eluted in 10 µl of EB buffer. A subsequent barcoding PCR was performed in a total volume of 25 µl using 3 µl of purified first PCR product, 2 µl of Illumina barcoding primer mix and 12.5 µl of 2x KAPA HiFi HotStart ReadyMix. The barcoding PCR was carried out as follows: 95°C for 3 min followed by 7 cycles of 98°C for 20 sec, 60°C for 20 sec and 72°C for 20 sec with a final extension step at 72°C for 3 min. PCR products were purified with AMPure beads in a 0.8 PCR:beads ratio according to manufacturer's instructions and DNA was eluted to 10 µl of EB buffer. The concentration of the final purified amplicons was measured with Qubit and all amplicons were pooled in equimolar ratio and sequenced on an Illumina MiSeq instrument. Paired-end reads were evaluated for mutations by alignment to the provided predicted off-target sequence using CRISPResso2[80] (CRISPRessoBatch --batch_settings 'my_tab_separated_batchfile' --amplicon_seq 'my_reference_sequence' --base_edit -g 'my_sgrna_sequence' -wc -10 -w 20). The effect of intronic mutations on gene splicing was predicted using the Alamut Visual software (Interactive Biosoftware, Rouen, France, version 2.15).

Genome-wide detection of editable polyadenylation signals

In order to find all editable polyadenylation signals in the genome with an AATAAA or ATTAAA motif, we constructed a regular expression that combines the polyadenylation signal motif sequence with a PAM site for SpCas9 (5'-NGG-3') at appropriate distance from the targeted base so that it falls into the reported activity window of nSpABEmax.[52] This regular expression was used to find all matching patterns in the human reference genome GRCh38 (https://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/). A similar approach was used to find all occurrences on the reverse complement strand. The results of this search were intersected with a list of known polyadenylation signals (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/gencode.v35.polyAs.gff3.gz) to obtain the final list of editable polyadenylation signals. We used the `famotif2bed` subcommand of the Fastools (https://fastools.readthedocs.io/en/latest/) package (version 1.0.2) for finding patterns in a reference sequence using regular expressions. All genome arithmetic was done using bedtools (https://bedtools.readthedocs.io/en/

latest/) (version 2.27.1). The full procedure is available online (https://github.com/jfjlaros/motif-edit) under the MIT Open Source license.

Statistical methods
A GraphPad Prism software v.8.4.2 was used for calculation of statistics. Sample sizes were not pre-determined prior experiments and a concrete statistical test is stated in the respective Figure legend.

Data Availability
The sequencing data generated for the off-target editing evaluation are available via SRA database under BioProject ID PRJNA732823.

Author Contributions

D.Š. designed and performed experiments, analysed results and wrote the manuscript. V.A.C. cloned and tested SpABEmax, SaABEmax and CjABEmax in HAP1 cells. A.Y. and D.Š. performed and analysed deep sequencing for off-target regions. J.F.J.L. and D.Š. performed the genome-scale analysis of editable polyadenylation signals. J.B. immortalized 2402 primary myoblasts and edited the manuscript. S.M.M. provided feedback and co-wrote the manuscript. All authors contributed to manuscript review.

Conflict of Interest

Authors declare no conflict of interest.

Acknowledgments

We thank members of van der Maarel group for all their helpful suggestions. This research was supported by funds from Friends of FSH Research, the Prinses Beatrix Spierfonds (W.OP14-01 and W.OR21-04) and from Spieren voor Spieren. D.Š. J.B and S.M.M. are members of the European Reference Network for Rare Neuromuscular Diseases [ERN EURO-NMD].

References

1. Wohlgemuth, M. *et al.* A family-based study into penetrance in facioscapulohumeral muscular dystrophy type 1. *Neurology* **91**, e444–e454 (2018).
2. Lunt, P. W. *et al.* Correlation between fragment size at D4F104S1 and age at onset or at wheelchair use, with a possible generational effect, accounts for much phenotypic variation in 4q35-facioscapulohumeral muscular dystrophy (FSHD). *Hum. Mol. Genet.* **4**, 951–958 (1995).
3. Statland, J. M. *et al.* Milder phenotype in facioscapulohumeral dystrophy with 7-10 residual D4Z4 repeats. *Neurology* **85**, 2147–2150 (2015).
4. Snider, L. *et al.* Facioscapulohumeral Dystrophy: Incomplete Suppression of a Retrotransposed Gene. *PLoS Genet.* **6**, e1001181 (2010).
5. Gannon, O. M., Merida de Long, L. & Saunders, N. A. DUX4 Is Derepressed in Late-Differentiating Keratinocytes in Conjunction with Loss of H3K9me3 Epigenetic Repression. *J. Invest. Dermatol.* **136**, 1299–1302 (2016).
6. Das, S. & Chadwick, B. P. Influence of Repressive Histone and DNA Methylation upon D4Z4 Transcription in Non- Myogenic Cells. doi:10.1371/journal.pone.0160022.
7. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
8. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
9. Young, J. M. *et al.* DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. *PLoS Genet.* **9**, e1003947 (2013).
10. Bosnakovski, D. *et al.* Low level DUX4 expression disrupts myogenesis through deregulation of myogenic gene expression. *Sci. Rep.* **8**, 1–12 (2018).
11. Geng, L. N. *et al.* DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev. Cell* **22**, 38–51 (2012).
12. Tawil, R., van der Maarel, S. M. & Tapscott, S. J. Facioscapulohumeral dystrophy: The path to consensus on pathophysiology. *Skeletal Muscle* vol. 4 (2014).
13. Sacconi, S. *et al.* FSHD1 and FSHD2 form a disease continuum. *Neurology* **92**, E2273–E2285 (2019).
14. Sacconi, S. *et al.* The FSHD2 Gene SMCHD1 Is a Modifier of Disease Severity in Families Affected by FSHD1. *Am. J. Hum. Genet.* **93**, 744–751 (2013).
15. Deutekom, J. C. T. V. *et al.* FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum. Mol. Genet.* **2**, 2037–2042 (1993).
16. van den Boogaard, M. L. *et al.* Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am. J. Hum. Genet.* **98**, 1020–1029 (2016).
17. Lemmers, R. J. L. F. *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat. Genet.* **44**, 1370–1374 (2012).
18. Hamanaka, K. *et al.* Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology* **94**, e2441–e2447 (2020).
19. Blewitt, M. E. *et al.* SmcHD1, containing a structural-maintenance-of- chromosomes hinge domain, has a critical role in X inactivation. (2008) doi:10.1038/ng.142.
20. Gendrel, A.-V. *et al.* Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome. *Dev. Cell* **23**, 265–279 (2012).
21. Gendrel, A.-V. *et al.* Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. *Mol. Cell. Biol.* **33**, 3150–65 (2013).
22. Nozawa, R.-S. *et al.* Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* **20**, 566–573 (2013).
23. Brideau, N. J. *et al.* Independent Mechanisms Target SMCHD1 to Trimethylated Histone H3 Lysine 9-Modified Chromatin and the Inactive X Chromosome. *Mol. Cell. Biol.* **35**, 4053–68 (2015).
24. van Geel, M. *et al.* Genomic Analysis of Human Chromosome 10q and 4q Telomeres Suggests a Common Origin. *Genomics* **79**, 210–217 (2002).

25. Lemmers, R. J. L. F. *et al.* Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *Am. J. Hum. Genet.* **81**, 884–94 (2007).
26. Lemmers, R. J. L. F. *et al.* Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nat. Genet.* **32**, 235–236 (2002).
27. Lemmers, R. J. L. F. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–3 (2010).
28. Lemmers, R. J. L. F. *et al.* Chromosome 10q-linked FSHD identifies *DUX4* as principal disease gene. *J. Med. Genet.* jmedgenet-2020-107041 (2021) doi:10.1136/jmedgenet-2020-107041.
29. Chen, J. C. *et al.* Morpholino-mediated Knockdown of DUX4 Toward Facioscapulohumeral Muscular Dystrophy Therapeutics. *Mol. Ther.* **24**, 1405–1411 (2016).
30. Ansseau, E. *et al.* Antisense Oligonucleotides Used to Target the DUX4 mRNA as Therapeutic Approaches in FaciosScapuloHumeral Muscular Dystrophy (FSHD). *Genes (Basel).* **8**, (2017).
31. Marsollier, A.-C. *et al.* Antisense targeting of 3- end elements involved in *DUX4* mRNA processing is an efficient therapeutic strategy for facioscapulohumeral dystrophy: a new gene-silencing approach. *Hum. Mol. Genet.* **25**, 1468–1478 (2016).
32. Joubert, R., Mariot, V., Charpentier, M., Concordet, J. P. & Dumonceaux, J. Gene Editing Targeting the DUX4 Polyadenylation Signal: A Therapy for FSHD? *J. Pers. Med.* **11**, 7 (2020).
33. Rashnonejad, A., Amini-Chermahini, G., Taylor, N. K., Wein, N. & Harper, S. Q. Designed U7 snRNAs inhibit DUX4 expression and improve FSHD-associated outcomes in DUX4 overexpressing cells and FSHD patient myotubes. *Mol. Ther. - Nucleic Acids* **23**, 476–486 (2021).
34. Aguirre, A. J. *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6**, 914–929 (2016).
35. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
36. Lim, C. K. W. *et al.* Treatment of a Mouse Model of ALS by In Vivo Base Editing. *Mol. Ther.* **28**, 1177–1189 (2020).
37. Levy, J. M. *et al.* Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nat. Biomed. Eng.* **4**, 97–110 (2020).
38. Ryu, S.-M. *et al.* Adenine base editing in mouse embryos and an adult mouse model of Duchenne muscular dystrophy. *Nat. Biotechnol. 2018 366* **36**, 536 (2018).
39. Koblan, L. W. *et al.* Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
40. Lemmers, R. J. L. F. *et al.* Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum. Mol. Genet.* **24**, 659–669 (2015).
41. Ricci, G. *et al.* Large genotype–phenotype study in carriers of D4Z4 borderline alleles provides guidance for facioscapulohumeral muscular dystrophy diagnosis. *Sci. Rep.* **10**, (2020).
42. Lemmers, R. J. *et al.* Deep characterization of a common D4Z4 variant identifies biallelic DUX4 expression as a modifier for disease penetrance in FSHD2. *Eur. J. Hum. Genet.* **26**, 94–106 (2018).
43. Yao, Z. *et al.* DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum. Mol. Genet.* **23**, 5342–5352 (2014).
44. Balog, J. *et al.* Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. *Epigenetics* **10**, 1133–1142 (2015).
45. Tawil, R. *et al.* Individual epigenetic status of the pathogenic D4Z4 macrosatellite correlates with disease in facioscapulohumeral muscular dystrophy. *Neurotherapeutics* **5**, 601–606 (2008).
46. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends in Genetics* vol. 18 348–351 (2002).
47. Concordet, J. P. & Haeussler, M. CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
48. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, (2015).

49. Cohen, J., DeSimone, A., Lek, M. & Lek, A. Therapeutic Approaches in Facioscapulohumeral Muscular Dystrophy. *Trends in Molecular Medicine* (2020) doi:10.1016/j.molmed.2020.09.008.
50. Bouwman, L. F., van der Maarel, S. M. & de Greef, J. C. The prospects of targeting DUX4 in facioscapulohumeral muscular dystrophy. *Curr. Opin. Neurol.* **33**, 635–640 (2020).
51. Himeda, C. L., Jones, T. I. & Jones, P. L. Targeted epigenetic repression by CRISPR/dSaCas9 suppresses pathogenic DUX4-fl expression in FSHD. *Mol. Ther. - Methods Clin. Dev.* **20**, 298–311 (2021).
52. Huang, T. P. *et al.* Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat. Biotechnol.* **37**, 626–631 (2019).
53. Richter, M. F. *et al.* Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* **38**, 883–891 (2020).
54. Li, X. *et al.* Programmable base editing of mutated TERT promoter inhibits brain tumour growth. *Nature Cell Biology* vol. 22 282–288 (2020).
55. Koblan, L. W. *et al.* In vivo base editing rescues Hutchinson–Gilford progeria syndrome in mice. *Nature* **589**, 608–614 (2021).
56. Yamada, M. *et al.* Crystal Structure of the Minimal Cas9 from Campylobacter jejuni Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. *Mol. Cell* **65**, 1109-1121.e3 (2017).
57. Kim, E. *et al.* In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat. Commun.* **8**, 1–12 (2017).
58. Zhang, J., Zhang, Y. Z., Jiang, J. & Duan, C. G. The Crosstalk Between Epigenetic Mechanisms and Alternative RNA Processing Regulation. *Frontiers in Genetics* vol. 11 998 (2020).
59. Nanavaty, V. *et al.* DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex. *Mol. Cell* **78**, 752-764.e6 (2020).
60. Xin, H., Wan, T. & Ping, Y. Off-Targeting of Base Editors: BE3 but not ABE induces substantial off-target single nucleotide variants. *Signal Transduct. Target. Ther. 2019 41* **4**, 1–2 (2019).
61. Grünewald, J. *et al.* Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
62. Grünewald, J. *et al. CRISPR DNA base editors with reduced RNA off-target and self-editing activities*. *Nature Biotechnology* vol. 37 1041–1048 (Nature Publishing Group, 2019).
63. Zhou, C. *et al.* Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* (2019) doi:10.1038/s41586-019-1314-0.
64. Lee, J. K. *et al.* Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat. Commun.* **9**, (2018).
65. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
66. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science (80-.).* **351**, 84–88 (2016).
67. Ryan, D. E. *et al.* Improving CRISPR-Cas specificity with chemical modifications in single-guide RNAs. *Nucleic Acids Res.* **46**, 792–803 (2018).
68. Rees, H. A. *et al.* Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery. *Nat. Commun.* **8**, 15790 (2017).
69. Billon, P. *et al.* CRISPR-Mediated Base Editing Enables Efficient Disruption of Eukaryotic Genes through Induction of STOP Codons. *Mol. Cell* **67**, 1068-1079.e4 (2017).
70. Kuscu, C. *et al.* CRISPR-STOP: Gene silencing through base-editing-induced nonsense mutations. *Nat. Methods* **14**, 710–712 (2017).
71. Wang, X. *et al.* Efficient Gene Silencing by Adenine Base Editor-Mediated Start Codon Mutation. *Mol. Ther.* **28**, 431–440 (2020).
72. Li, Z., Xiong, X., Wang, F., Liang, J. & Li, J. Gene disruption through base editing-induced <scp>messenger RNA</scp> missplicing in plants. *New Phytol.* **222**, 1139–1148 (2019).
73. Gapinske, M. *et al.* CRISPR-SKIP: Programmable gene splicing with single base editors. *Genome Biol.* **19**, 107 (2018).

74. Sheets, M. D., Ogg, S. C. & Wickens, M. P. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* **18**, 5799–805 (1990).

75. Nunes, N. M., Li, W., Tian, B. & Furger, A. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J.* **29**, 1523–1536 (2010).

76. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001–1010 (2000).

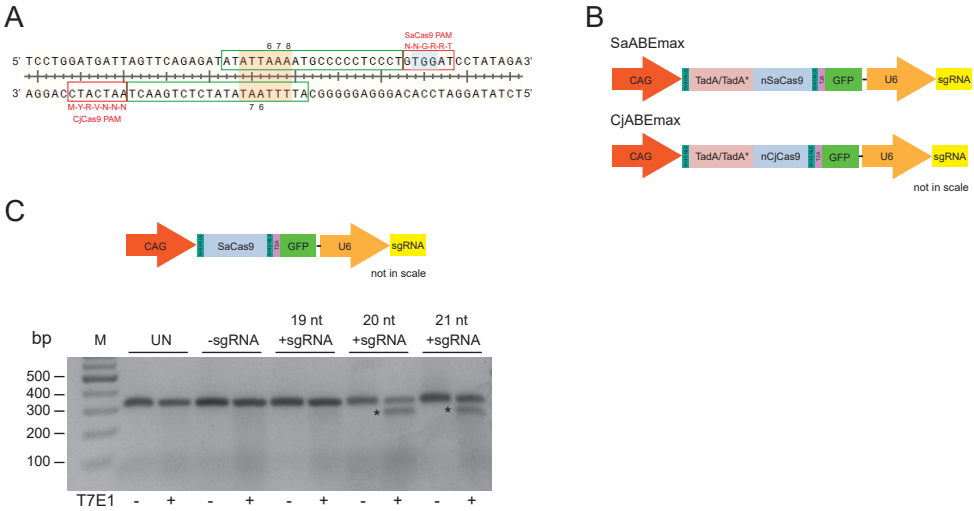
77. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

78. Zhu, C. H. *et al.* Cellular senescence in human myoblasts is overcome by human telomerase reverse transcriptase and cyclin-dependent kinase 4: Consequences in aging muscle and therapeutic strategies for muscular dystrophies. *Aging Cell* **6**, 515–523 (2007).

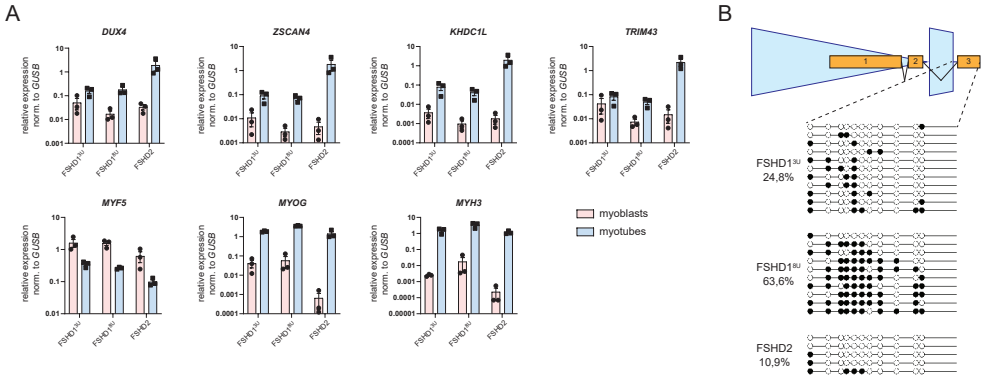
79. Kluesner, M. G. *et al.* EditR: A Method to Quantify Base Editing from Sanger Sequencing. *Cris. J.* **1**, 239–250 (2018).

80. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nature Biotechnology* vol. 37 224–226 (2019).

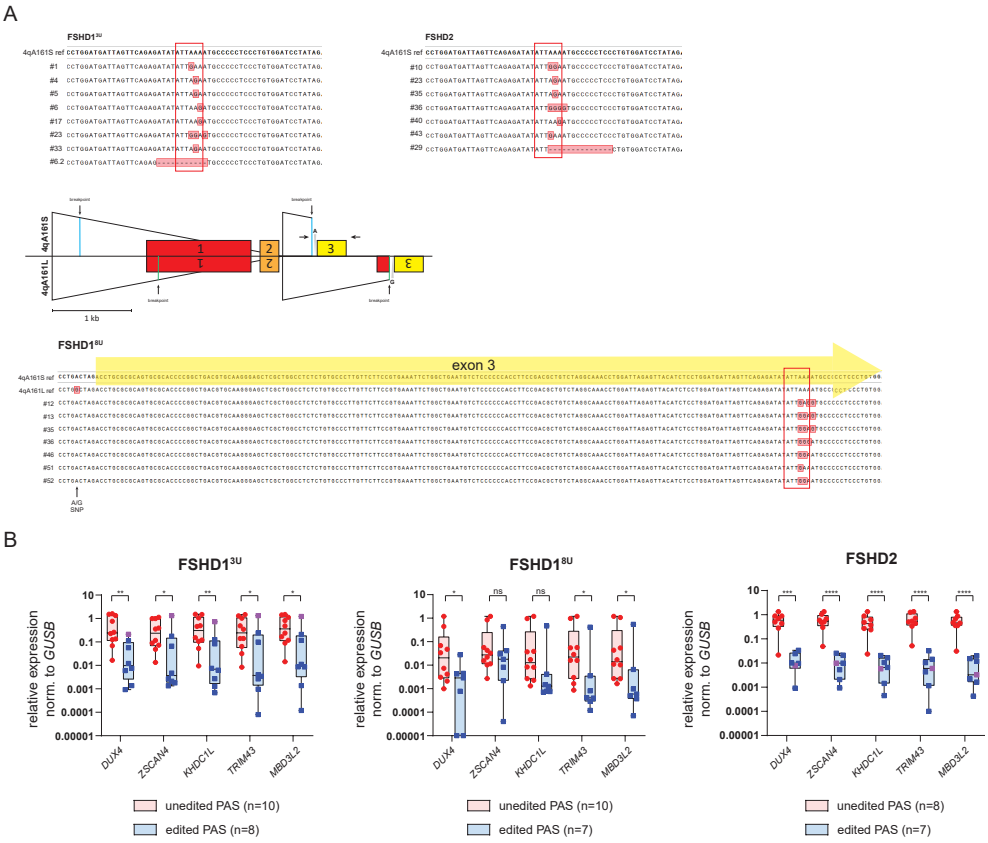
Supplementary Information



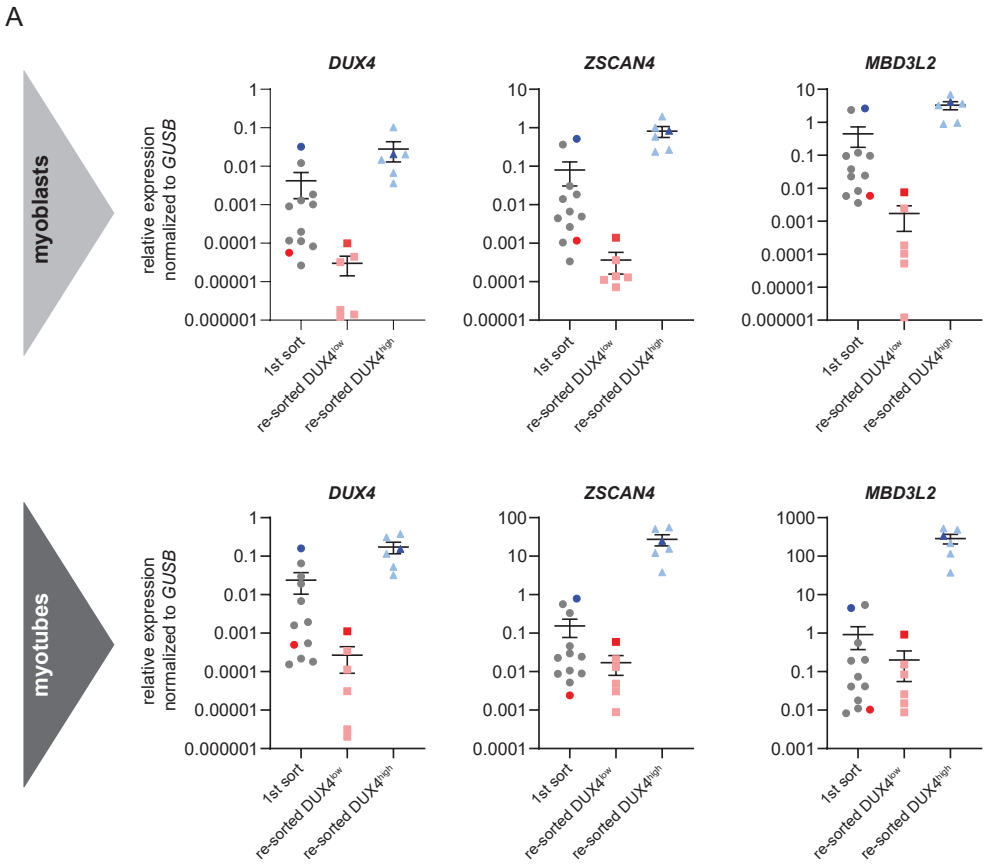
Suppl. Figure 1. *DUX4* PAS is not editable either by nSaABEmax or by nCjABEmax. A) DNA sequence surrounding of *DUX4* PAS (highlighted in the orange box). Cognate PAM sites for CjCas9 and SaCas9 are outlined in red rectangles, while sgRNA protospacer regions are outlined in green rectangles. The PAM site for SpCas9 is highlighted in the blue box. Adenines within the *DUX4* PAS amenable for base editing are numbered from the beginning of the sgRNA protospacer. **B)** Schematic maps of modified all-in-one vectors coding for nSaABEmax (top) and nCjABEmax (bottom). **C)** Schematic map of the pX601 vector for simultaneous sgRNA and SaCas9 nuclease expression (top). Result of the T7E1 assay performed on HAP1 cells which were transfected with a pX601 vector expressing the SaCas9 nuclease and sgRNAs of different length (19 nt-, 20 nt- or 21 nt-long) targeting the *DUX4* PAS (bottom). Untransfected cells (UN) or cells transfected with no sgRNA containing vector (-sgRNA) served as negative control. Asterisks mark the T7E1 cleavage products.



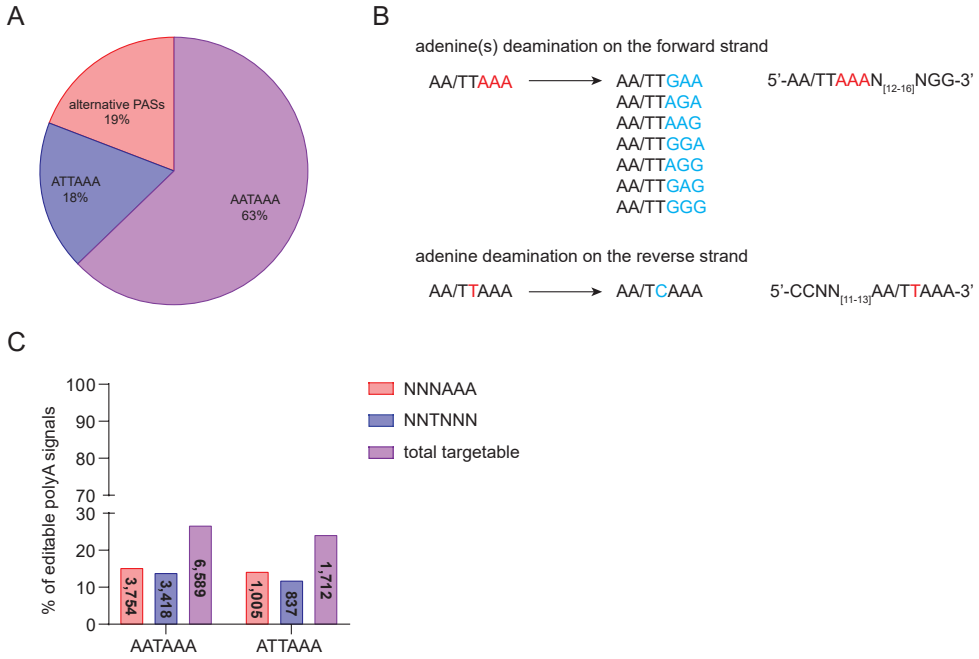
Suppl. Figure 2. Expression and methylation profiles of FSHD immortalized myoblasts used for editing. A) mRNA expression levels of *DUX4* and four DUX4 target genes in 3 model FSHD cell lines used for base editing experiments at myoblast and myotubes stage. Expression of myogenic markers (*MYOG* and *MYH3*) is provided to show successful myogenic differentiation. *GUSB* was used as a housekeeping gene. Bars represent mean ±SEM. Cells were grown three independent times and analysed for their gene expression. **B)** CpG methylation level of the FasPAS region encompassing exon 3 of *DUX4* in the three parental FSHD immortalized myoblast lines used for base editing. Individual rows represent a single molecule, empty circles denote unmethylated cytosines in a CpG context, while full circles denote methylated cytosines in a CpG context. Average methylation of the region (in %) is provided below the name of each sample. Note, that in case of FSHD1^{8U}, both alleles (contracted and non-contracted allele) are amplified in bisulfite PCR.



Suppl. Figure 3. Genotypes of successfully edited clones and their expression data in proliferating myoblasts. **A)** Genotypes of successfully *DUX4* PAS edited clones obtained from three independent FSHD lines (top left: FSHD1^{3U}, top right: FSHD2 and bottom: FSHD1^{8U}) aligned to the WT reference sequence. The *DUX4* PAS sequence is highlighted in a red rectangle and red colored bases denote mismatches. Mirror schematic of the 4qA161S and 4qA161L D4Z4 haplotype termini is provided to show the genotyping approach for the FSHD1^{8u} cell line. Red box represents exon 1, orange box represents exon 2 and yellow box represents exon 3 which corresponds to the yellow arrow highlighting the exon 3 sequence in the genotyping tracks. In the 4qA161L haplotype, a small 5' part of exon 1 (red box) precedes exon 3 due to a different breakpoint. A specific forward primer was used to selectively amplify the 4qA161S allele, which was confirmed by the presence of the SNP (A instead of G) in the Sanger sequencing tracks (marked by arrow) for all the genotyped clones. Reference sequence for both, 4qA161S and 4qA161L allele is provided. B**)** mRNA levels as assessed by RT-qPCR of *DUX4* and four DUX4 target genes (*MBD3L2*, *ZSCAN4*, *TRIM43* and *KHDC1L*) in PAS unedited vs edited clones derived from two FSHD1 and one FSHD2 cell lines during proliferation. Statistical significance was calculated with unpaired two-tailed t-test (ns: non-significant, *: <0.05, **: <0.01, ***: <0.001, ****: <0.0001) on log2 transformed expression values to correct for skewed distribution. Expression values normalized to *GUSB* as house-keeping gene are plotted. Line represent mean and whiskers represent min and max value. Individual data points represent individual clones, two violet clones carry a deletion over *DUX4* PAS.



Suppl. Figure 4. DUX4 expression signature is clonally stable. A) mRNA levels as assessed by RT-qPCR of *DUX4* and its two target genes (*ZSCAN4* and *MBD3L2*) were measured in clonal lines established from FSHD1^{8U} immortalized myoblasts and 5 new daughter clones derived from a parental clone with either high *DUX4* expression (dark blue colour) or low *DUX4* expression (dark red colour). Daughter clones are marked by light blue or light red colour. Expression data for both myoblasts (top) and myotubes (bottom) are provided.



Suppl. Figure 5. Identification of polyadenylation signals targetable by nSpABEmax in the human genome. **A)** Genome-wide prevalence of most common polyadenylation signal hexamers based on all annotated polyadenylation signals in Gencode. **B)** Representation of two different approaches of targeting polyadenylation signals by nSpABEmax either on the coding or non-coding strand and their possible outcomes. Targeted positions are in red and expected modified bases are in blue. **C)** Percentage of annotated polyadenylation signals in the GRCh38 human genome with the most prevalent motifs (AATAAA and ATTAAA) whose adenines are targetable by nSpABEmax either on the coding (red) or non-coding (blue) strand. The number within each bar represents the actual number of targetable polyadenylation signals.

Suppl. Table 1. Oligos used in the study.

Primer Name	Sequence
genotyping	
34bp-gct-161 F	CAGCTGCCAGCGCGGAGCT
12591F	CCCGCCCGGGCCCCTGCA
R-GOTO #4	CAGGGGATATTGTGACATATCTCTGCACTCATC
sgRNA cloning	
SpABEmax/ABE7.10 PAS gRNA F	CACC**GATTAAAATGCCCCCTCCCTG**
SpABEmax/ABE7.10 PAS gRNA R	AAAC**CAGGGAGGGGGCATTTTAATC**
SaABEmax PAS gRNA F	CACC**GATATTAAAATGCCCCCTCCCT**
SaABEmax PAS gRNA R	AAAC**AGGGAGGGGGCATTTTAATATC**
CjABEmax PAS gRNA F	CACC**GATTTTAATATATCTCTGAACT**
CjABEmax PAS gRNA R	AAAC**AGTTCAGAGATATATTAAAATC**
SpABEmax AAVS1-TS2 gRNA F	CACC**GAGTAGAGGCGGCCACGACC**
SpABEmax AAVS1-TS2 gRNA R	AAAC**GGTCGTGGCCGCCTCTACTC**
CjABEmax AAVS1-TS2 gRNA F	CACC**GAGTAGAGGCGGCCACGACCTG**
CjABEmax AAVS1-TS2 gRNA R	AAAC**CAGGTCGTGGCCGCCTCTACTC**
pX458-SpABE7.10 cloning primers	
TadA AgeI F	TGGACCGGTGAGAGCCGCCACCATGTC
ABE7.10 D10A R	GAGTTGGTGCCGATGGCCAGACCAATAGAATACTTTTTATC
pX335 Cas9 D10A F	GATAAAAAGTATTCTATTGGTCTGGCCATCGGCACCAACTC
pX335 SV40 R	CTGCCCTCTCCACTGCCGAACACCTTTCTCTTCTTCTTGGGGCTGT
SV40-T2A F	ACAGCCCCAAGAAGAAGAGAAAGGTGTTCGGCAGTGGAGAGGGCAG
EGFP EcoRI R	GTTAGAATTCCTTGTACAGCTCGTCC
pX458-SpABEmax cloning primers	
ABEmax AgeI F	TTGGACCGGTGGCCGCTAATACGACTCACTATAGG
ABEmax ApaI R	CAGAGGGCCCACGTAGTAGGG
pX601-SaABEmax cloning primers	
pX458 CAG F	TTCTGCAGACAAATGGCTCTAGAGGTACCCG
pX458 ABEmax 32aa R	AGTTCCGCTTTGACCCCCCGCTGCTGCC
pX601 SaCas9 D10A F	CGGGGGGTCAAAGCGGAACTACATCCTGGGCCTGGCCATCG
pX601 SaCas9 HindIII R	GCTCTGGATGAAGCTTCTCTTCACGACGG
pX404-CjABEmax cloning primers	
CMV enh XbaI F	GCGGCCTCTAGAGATACCCGTTACATAAC

Primer Name	Sequence
pX601-ABEmax CAG R	TGACTCGAACTCGCTTCCGTC
ABEmax AgeI F	TTGGACCGGTGGCCGCTAATACGACTCACTATAGG
ABEmax-CjCas9 R	GGATGCGGGCTGACCCCCCGCTGCTGCCCC
pX404 CjCas9 D8A F	CGGGGGGTCAGCCCGCATCCTCGCTTTCGCCATCG
pX404 CjCas9 SV40 R	ATCCTCTGCCCTCCACCTTTCTCTTCTTCTTGGGG
CjCas9-T2A F	GAAGAGAAAGGTGGAGGGCAGAGGATCCCTGCTAACATGTGG
bGH PAS Bsu36I R	TAGGCCCTCAGGTACCTCCCCAGCATG
U6 KpnI F	GGGAGGTACCTGAGGGCC
U6 BsaI R	AGGGACTAAAACTGAGACCTGCCGT
Cj sgRNA BbsI F2	AGGTCTCAGTTTTAGTCCCTGAAAAGGG
Cj sgRNA NotI R3	GGTTCCTGCGGCCGCAAAAAAAGCGGTTTTAGGGG
qPCR primers	
GUSB qPCR F	CTCATTTGGAATTTTGCCGATT
GUSB qPCR R	CCGAGTGAAGATCCCCTTTTTA
DUX4 4qA-S qPCR F	CCCAGGTACCAGCAGACC
DUX4 4qA-S qPCR R	TCCAGGAGATGTAACTCTAATCCA
MYOG qPCR F	GCCAGACTATCCCCTTCCTC
MYOG qPCR R	GGGGATGCCCTCTCCTCTAA
MYF5 qPCR F	TTCTCCCCATCCCTCTCGCT
MYF5 qPCR R	AGCCTGGTTGACCTTCTTCAG
MYH3 qPCR F	TGATCGTGAAAACCAGTCCATTCT
MYH3 qPCR R	TTGGCCAGGTCCCCAGTAGCT
ZSCAN4 qPCR F	TGGAAATCAAGTGGCAAAAA
ZSCAN4 qPCR R	CTGCATGTGGACGTGGAC
TRIM43 qPCR F	ACCCATCACTGGACTGGTGT
TRIM43 qPCR R	CACATCCTCAAAGAGCCTGA
KHDC1L qPCR F	TGAATCAGGTGGGAGCACAG
KHDC1L qPCR R	CAATGCAGCGAAGGTACGTG
MBD3L2 qPCR F	GCGTTCACCTCTTTTCCAAG
MBD3L2 qPCR R	GCCATGTGGATTTCTCGTTT
3'RACE primers	
3'RACE cDNA primer	GCTGTCAACGATACGCTACGTAACGGCATGACAGTGTTTTTTTTTTTTTTTTTTTTTTT
DUX4 3RACE ex3 F	CGCACCCCGGCTGACGTGCAAG
3RACE R	GCTGTCAACGATACGCTACGTAACG
DUX4 3RACE ex3 nest F	CGCTGGCCTCTCTGTGCCCTTG

Primer Name	Sequence
3RACE nested R	CGCTACGTAACGGCATGACAGTG
NGS primers for predicted off-target sites	
off site #1 P5	GATGTGTATAAGAGACAGTTCTGACCATGCTCTCTCCAG
off site #1 P7	CGTGTGCTCTTCCGATCTCCCTCAGCTCTCACCTATGG
off site #2 P5	GATGTGTATAAGAGACAGTGCTCACAACAGCTTACAACAC
off site #2 P7	CGTGTGCTCTTCCGATCTTGGGATCATTGCTGTGAAAG
off site #3 P5	GATGTGTATAAGAGACAGTCCAATGTAAAATGGCAGCTC
off site #3 P7	CGTGTGCTCTTCCGATCTTGTTCCTTGCCTTGAAGGTT
off site #4 P5	GATGTGTATAAGAGACAGGGCTCCTCACTGCTGAAAAGC
off site #4 P7	CGTGTGCTCTTCCGATCTCAGAAGAGGAAGGCAGGATG
off site #5 P5	GATGTGTATAAGAGACAGGCGGGAGCCAAAGTAGAGAT
off site #5 P7	CGTGTGCTCTTCCGATCTCCCCACCTCTGTAGGTCTGA
off site #6 P5	GATGTGTATAAGAGACAGAGGGGTTTGGGAATGTAAGG
off site #6 P7	CGTGTGCTCTTCCGATCTAGGTGAGGGAAACACACCTG
off site #7 P5	GATGTGTATAAGAGACAGGTCACACCCAGGAAGGGATA
off site #7 P7	CGTGTGCTCTTCCGATCTGACCCTGTCAGCCTCATCTC
off site #8 P5	GATGTGTATAAGAGACAGGTTGGGAGAGTCTTGCTTGC
off site #8 P7	CGTGTGCTCTTCCGATCTTGTGAAGGCCGGAACTCTAT
off site #9 P5	GATGTGTATAAGAGACAGAAACCTGCCCATACCAAGTG
off site #9 P7	CGTGTGCTCTTCCGATCTATGCAGCAGGGAACTTGTTT
off site #10 P5	GATGTGTATAAGAGACAGTTTGCTATTGCCCAGTTTCC
off site #10 P7	CGTGTGCTCTTCCGATCTAGAAGAGGCTTCACCACCAA
FasPAS bisulfite primers	
FasPAS_Tag	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATAGGGAGGGGGTATTTTA
RevAS_Tag	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACRATCAAAAACATACCTC-TATCTA

Suppl. Table 2. Detailed information about cell lines used in the study.

Cell Line (ID)	Cell Type	Clinical Status and *SMCHD1* Mutation Status	Sex	4q allele #1	4q allele #2
HAP1	derived from KBM-7	NA	M	25U 4qA161S	NA
2402 ($FSHD^{3U}$)	immortalized myoblasts	FSHD1	M	3U 4qA161S	16U 4qB163
073 ($FSHD^{8U}$)	immortalized myoblasts	FSHD1	M	7U 4qA161S	36U 4qA161L
200 (FSHD2)	immortalized myoblasts	FSHD2 (SMCHD1 p.Lys204Glu)	M	11U 4qA161S	39U 4qB168

Suppl. Table 3. Results of 3' RACE experiment.

		FSHD1 (3U)					FSHD1 (8U)							FSHD2					
% of specified CS detected in clones	CS type/clone #	#1	#5	#6	#23	all clones	#12	#13	#35	#46	#51	#52	all clones	#10	#23	#29	#35	#36	all clones
	WT CS	0	20	20	12,5	13,125	0	0	0	0	37,5	0	6,25	0	0	62,5	0	0	12,5
	proximal CS	0	20	20	25	16,25	0	0	62,5	0	0	33,33333	15,97222	100	100	37,5	100	100	87,5
	distal CS	100	60	60	62,5	70,625	100	100	37,5	100	62,5	66,66667	77,77778	0	0	0	0	0	0

Suppl. Table 4. Table of predicted off-target sites by nSpABEmax. Off-target sites in red font were examined with NGS.

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATCAAAATTC-CCCCTCCCTGTGG	..*.....*.....	2	4,030448592	0,228571429	chr10	133685702	133685724	+	intergenic:AL845259.2-AL732375.3
ATCAAAATTC-CCCCTCCCTGTGG	..*.....*.....	2	4,030448592	0,228571429	chr10	133761891	133761913	+	intergenic:AL732375.6-Gap
ATCAAAATGC-CCGCTCCCTGCGG	..*.....*.....	2	3,295480909	0,058441558	chrY	11323586	11323608	+	intergenic:AC134882.2-AC134882.3
CCTAAAACGC-CCCCTCCCTGAGG	**.....*.....	3	2,543507363	0,734693877	chr9	1267479	1267501	-	intergenic:DMRT2-RNA5SP279
AATCAAATGTC-CCCTCCCTGAGG	.*.*.....*.....	3	2,461181435	0,281075326	chr6	13331126	13331148	+	intergenic:TBC1D7-GFOD1
GTCACAATGC-CCCCTCCCTGCGG	*.*.*.....	3	2,392592593	0,214285715	chrY	11334358	11334380	+	intergenic:AC134882.3-RN7SL702P
GTCACAATGC-CCCCTCCCTGTGG	*.*.*.....	3	2,392592593	0,214285715	chr14_GL000225v1_random	135024	135046	-	intergenic:AL671532.7-AL671532.4
GTCACAATGC-CCCCTCCCTGCGG	*.*.*.....	3	2,392592593	0,214285715	chrUn_GL000216v2	170471	170493	-	intergenic:BX546479.2-BX546479.1
GTCACAATGC-CCCCTCCCTGCGG	*.*.*.....	3	2,392592593	0,214285715	chr14_GL000225v1_random	62891	62913	-	intergenic:AL671532.6-AL671532.5
GTCACAATGC-CCCCTCCCTGCGG	*.*.*.....	3	2,392592593	0,214285715	chr14_GL000225v1_random	105546	105568	+	intergenic:AL671532.7-AL671532.4
GTCACAATGC-CCCCTCCCTGCGG	*.*.*.....	3	2,392592593	0,214285715	chr14_GL000225v1_random	603	625	-	intergenic:Gap-AL671532.1
GTCACAATGC-CCCCTCCCTGTGG	*.*.*.....	3	2,392592593	0,214285715	chr18	111595	111617	-	intergenic:RP11-683L23.1-MIR8078
GTCACAATGC-CCCCTCCCTGCGG	*.*.*.....	3	2,392592593	0,214285715	chr14_GL000225v1_random	71181	71203	-	intergenic:AL671532.6-AL671532.5
AAAACAACGC-CCCCTCCCTGGGG	.**.*..*.....	4	1,345833333	0,302197802	chr5	54850658	54850680	+	intergenic:CTD-2591A1.1-AC112198.2
CCTAGAATGTC-CCCTCCCTGCGG	**..*....*.....	4	1,317695783	0,497863145	chr20	32515379	32515401	-	intergenic:RP5-1184F4.7/NOL4L-RP5-1184F4.7

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
TGTGCAATGC-CCCCTCCCTGAGG	**.**.....	4	1,304945055	0,2625	chr12	2444719	2444741	-	intergenic:CACNA1C-IT3-CACNA1C
GTCACAATGTC-CCCTCCCTGCGG	*.*.*....*.....	4	1,299248042	0,201680673	chr14_GL000225v1_random	32085	32107	+	intergenic:Gap-AL671532.1
GTCACAATGTC-CCCTCCCTGTGG	*.*.*....*.....	4	1,299248042	0,201680673	chrUn_GL000216v2	145243	145265	+	intergenic:BX546479.2-BX546479.1
GTCACAATGAC-CCCTCCCTGTGG	*.*.*....*.....	4	1,299248042	0,185714286	chr14_GL000225v1_random	60117	60139	+	intergenic:AL671532.6-AL671532.5
GGGAGAATGC-CCCCTCCCTGGGG	***.*.....	4	1,286675824	0,3024	chr2	218831310	218831332	-	intron:PRKAG3
AGAAAAAGGAC-CCCTCCCTGGGG	.**....*.*.....	4	1,2395125	0,381333333	chr10	62692200	62692222	-	intergenic:ZNF365-RP11-436D10.3
ATGGAAAAGTC-CCCTCCCTGAGG	..**...*.*.....	4	1,2395125	0,235294118	chr21	42351395	42351417	+	intergenic:TFF2-TFF1
ATAAAAAAGC-CCTCTCCCTGGGG	..*....*....*.....	3	1,074082667	0,4	chrX	109927881	109927903	-	intergenic:Y_RNA-TMEM164
CTCAAATAGC-CCCCTCCCTGGGG	*.*...**.....	4	0,919204167	0,128571429	chr13	98865880	98865902	-	intron:DOCK9
CTACAACTGC-CCCCTCCCTGTGG	*.**..*.....	4	0,919204167	0,094537815	chr17	47882617	47882639	+	intergenic:RP11-6N17.2-RP11-6N17.4
ACTTCATTGC-CCCCTCCCTGAGG	.*.**.*.....	4	0,891277473	0,119318182	chr11	66897638	66897660	+	intergenic:AP003176.1-PC
ATTTAAGAGGC-CCCTCCCTGTGG	...*..**.*.....	4	0,858607543	0,139750446	chr14	61570002	61570024	-	intergenic:RP11-47I22.2-RP11-47I22.2/FLJ22447
CTTCAGAAGC-CCCCTCCCTGTGG	*..*.*.*.....	4	0,82579023	0,172869147	chr17	44991828	44991850	+	intergenic:CTD-2534I21.9-DCAKD
GTCACAATTC-CCCCTCCCTGCGG	*.*.*...*.....	4	0,822304167	0,114285714	chr14_GL000225v1_random	88470	88492	-	intergenic:AL671532.5-AL671532.3
ATTATAGGGGC-CCCTCCCTGTGG	....*.**.*.....	4	0,820866552	0,073202614	chr1	228876405	228876427	+	intergenic:RHOU-RP5-1065P14.2
GTTAAAAGCTC-CCCTCCCTGTGG	*.....***.....	4	0,805112123	0,371644042	chr13	105165594	105165616	+	intergenic:LINC01309-DAOA-AS1/DAOA-AS1_1
GTCACAATGCAC-CCTCCCTGCGG	*.*.*.....*.....	4	0,782934488	0,160714286	chrUn_GL000216v2	175985	176007	+	intergenic:BX546479.1-Gap

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
GTCACAATGCAC-CCTCCCTGCGG	*.*.*.....*.....	4	0,782934488	0,160714286	chr14_KI270724v1_random	918	940	-	intergenic:Gap-AC242209.2
GTCACAATGCAC-CCTCCCTGCGG	*.*.*.....*.....	4	0,782934488	0,160714286	chr14_GL000225v1_random	100028	100050	-	intergenic:AL671532.3-AL671532.7
TTGATAATGCAC-CCTCCCTGGGG	*.*.*.....*.....	4	0,782934488	0,136363637	chr7	54004331	54004353	-	intergenic:RP11-806J6.1-HPVC1
ATTGAAAATTC-CCCTCCCTGGGG	...*...***.....	4	0,768095474	0,250980392	chr5	142319282	142319304	-	intergenic:SPRY4-IT1/SPRY4-IT1_1/SPRY4-IT1_2-SPRY4
ATTAATATGTAC-CCTCCCTGGGG	.....*...**.....	3	0,750412098	0,504201681	chr15	87913594	87913616	+	intron:NTRK3
ACTGAAATGTTC-CCTCCCTGAGG	.*.*.....**.....	4	0,73132116	0,155138979	chr9	99464423	99464445	-	intergenic:NAMA-RP11-547C13.1
ACCAGAATGC-CACCTCCCTGAGG	.**.*.....*.....	4	0,694060843	0,188921283	chr6	26457319	26457341	+	intergenic:BTN3A3-BTN2A1
GTTTAAATGC-CCCCTCCATGGGG	*..*.....*..	3	0,656790123	0,342657342	chr3	110983651	110983673	-	intron:PVRL3-AS1
ACTGAAATGT-CACCTCCCTGTGG	.*.*.....*.*.....	4	0,648306325	0,360144058	chr13	51148475	51148497	-	intergenic:LINC00371-RP11-457D13.4
AATAAAAGGTC-CCCTCCCTCTGG	.*.....*.*.....*	4	0,642348856	0,250290886	chr9	42599886	42599908	-	intergenic:RP11-475I24.8-RP11-175I6.5
AATAAAAGGTC-CCCTCCCTCTGG	.*.....*.*.....*	4	0,642348856	0,250290886	chr9	62501815	62501837	+	intergenic:RP11-111F5.8-LINC01189
AATAAAAGGTC-CCCTCCCTCTGG	.*.....*.*.....*	4	0,642348856	0,250290886	chr9	42694426	42694448	+	intergenic:RP11-475I24.8-RP11-175I6.5
AATAAAAGGTC-CCCTCCCTCTGG	.*.....*.*.....*	4	0,642348856	0,250290886	chr9	39843342	39843364	+	intergenic:GLIDR-Metazoa_SRP
AATAAAAGGTC-CCCTCCCTCTGG	.*.....*.*.....*	4	0,642348856	0,250290886	chr9	62407357	62407379	-	intergenic:RP11-111F5.8-LINC01189
ATTTAAATTC-CCCCTCCCCGTGG	...*....*.....*.	3	0,606440299	0,093625914	chr4	104550531	104550553	-	intergenic:CXXC4-AC004063.1
TTCCAAATGC-CCACTCCCTGGGG	*.**.....*.....	4	0,573580538	0,058177117	chr1	18818355	18818377	+	intergenic:PAX7-TAS1R2
CTCAAAATGC-CCCCTTCCTGTGG	*.*.....*....	3	0,534369486	0,113029828	chr20	58463715	58463737	-	intron:APCDD1L

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
TTGAAAAAGC-CCCCTCCCGGGGG	*.*....*.....*.	4	0,519472711	0,1	chr17	4520446	4520468	-	intron:SPNS2
CTTTAAATCCAC-CCTCCCTGGGG	*..*....*.*.....	4	0,485165286	0,22027972	chr5	176680579	176680601	-	intergenic:TSPAN17-RP11-375B1.1
GTTATAATACTC-CCTCCCTGGGG	*...*...*.*.....	4	0,485165286	0,071928072	chr6	37177220	37177242	+	intergenic:PIM1-TMEM217
GTCACAATGC-CCCCTCCCTGCAG	*.*.*..........	3	0,478518519	0,055555556	chr14_GL000225v1_random	44827	44849	+	intergenic:AL671532.1-AL671532.2
AAGAAATTG-CCTCCTCCCTGGGG	.**...*....*.....	4	0,474043556	0,09966716	chr14	24767009	24767031	-	intergenic:RP11-104E19.1-STXBP6
ATGATAGTGC-CACCTCCCTGGGG	..*.*.*....*.....	4	0,474043556	0,091673033	chr22	26061580	26061602	-	exon:CTA-796E4.5
CTAAATATGCTC-CCTCCCTGGGG	*.*..*....*.....	4	0,473675365	0,134559318	chr11	32539555	32539577	+	intergenic:AC087653.1-EIF3M
ATTATAAATCTC-CCTCCCTGGGG	....*..**.*.....	4	0,462858836	0,047738928	chr10	31873953	31873975	+	intron:ARHGAP12
ATTAAAAAGC-CAGCTCCCTGTGG	.......*...**.....	3	0,462027586	0,077922078	chr11	34223093	34223115	+	intron:ABTB2
ATGGACATGCTC-CCTCCCTGAGG	..**.*....*.....	4	0,451897188	0,043706294	chr3	61040894	61040916	-	intergenic:U3-FHIT
AATTAATTGC-CCCCTCACTGAGG	.*.*..*.....*...	4	0,416345417	0,109935897	chr2	213186386	213186408	+	intergenic:RP11-105N14.3-RP11-105N14.2
AGGAAAAGGC-CCCCACCCTGGGG	.**....*.....*.....	4	0,397208228	0,178315789	chr1	65033120	65033142	+	intergenic:RNU6-1176P-MIR3671
ATCAGGATGC-CCCCTCCCTATGG	..*.**.....*	4	0,393858933	0,206632653	chr10	107986748	107986770	+	intergenic:RNA5SP326-LINC01435
CCTAAAATGCAG-CCTCCCTGCGG	**.....**.....	4	0,390673193	0,244897959	chr7	131029646	131029668	+	intron:LINC-PINT
TTTAAAAAGCTC-CCTCCCTTTGG	*.....*..*.....*	4	0,387083187	0,172307692	chr9	85570807	85570829	+	intron:AGTPBP1
ATTCAAATGC-CCCCACCCTCAGG	...*.....*....*	3	0,374491005	0,087571871	chr3	159951352	159951374	+	intergenic:IL12A-AS1-IL12A/IL12A-AS1
GTTTAAATAC-CCTCTCCCTGGGG	*..*....*...*.....	4	0,355433782	0,286363636	chr4	86472877	86472899	-	intergenic:MAPK10-MIR4452
GATAAACTGC-CCCCTCCCTGGAG	**....*.....	3	0,347443106	0,095975783	chr13	25971732	25971754	-	intron:ATP8A2

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATCATAATAC-CCACTCCCTGTGG	..*.*...*...*.....	4	0,333568181	0,038532896	chr2	121396548	121396570	-	intron:CLASP1
AGGAAGATGC-CCTCTCCCTGTGG	.**..*.....*.....	4	0,330292552	0,21	chr4	141412433	141412455	+	intergenic:RP11-362F19.1-RP11-362F19.3
AAGAAAGTGC-CCCCTCCCTGAGA	.**...*.....	3	0,326828148	0,020739065	chr1	82754062	82754084	-	intergenic:AL357973.1-LINC01362
ATTACTAGGC-CCTCTCCCTGGGG	....**.*....*.....	4	0,319580819	0,183333333	chr17	50112688	50112710	+	exon:PDK2/SAMD14
GATAAAATGCTC-CCTCCCTGAGA	**.....*.....	3	0,312444444	0,01808021	chr9	3187345	3187367	-	intron:LINC01231
ATGAAGCTGCAC-CCTCCCTGGGG	..*..**...*.....	4	0,308645779	0,1171875	chr7	5227133	5227155	-	intron:WIPI2
AAAAAAATTC-CCCCTCCCTGCAG	.**.....*.....	3	0,306465167	0,08357075	chr2	229670981	229671003	-	intergenic:RNU7-9P-DNER
ATTATATTGTC-CCCTCCCTGAAG	....*.*..*.....	3	0,305282682	0,038819568	chr18	10313075	10313097	+	intergenic:RP11-419J16.1-RP11-138E9.2
ATTAAAATGCAG-CCTCCCTTTGG	.....**.....*	3	0,304283316	0,233333333	chr8	35626392	35626414	+	intron:UNC5D
ATTAAATATCTC-CCTCCCTGGGG	.....***.*.....	4	0,302236647	0,057435897	chr3	27892373	27892395	-	intergenic:AC098973.2-AC092415.1
GTTCAAATGC-CCCCTCCCTCAGA	*..*.....*	3	0,298418079	0,010504202	chr8	140299073	140299095	-	intron:TRAPPC9
AAAAACATGC-CCCCTCCCTGCAG	.**..*.....	3	0,289503704	0,071225071	chr8	103144201	103144223	-	intergenic:BAALC-AS2-MIR3151
TTTAAAAAGC-CAGCTCCCTGTGG	*.....*...**.....	4	0,286208544	0,077922078	chrY	18018460	18018482	-	intergenic:CDY2B-CDY2A
TTTAAAAAGC-CAGCTCCCTGTGG	*.....*...**.....	4	0,286208544	0,077922078	chrY	17887521	17887543	+	intergenic:CDY2B-CDY2A
ATAAAAATCAC-CCCTCCCTGTGA	..*.....**.....	3	0,282254419	0,023148148	chr20	24958136	24958158	-	intron:CST7
GTTGAACTGC-CCCCCCCCTGAGG	*..*..*.....*.....	4	0,275145253	0,074573864	chr10	131360895	131360917	-	intergenic:TCERG1L-LINC01164
ATTGAAAAGCCT-GCTCCCTGGGG	...*...*...**.....	4	0,272415361	0,036713287	chrX	151825372	151825394	-	intergenic:CNGA2-RP11-366F6.2
ATTATTATGGC-CCCTCCCTGAAG	....**...*.....	3	0,270418774	0,026187804	chr1	30748432	30748454	+	intron:LAPTM5

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
TTTACAATGCCCTCTCCCTATGG	*...*.....*.....*	4	0,26991206	0,328125	chr3	20933827	20933849	-	intergenic:RNU6-815P-AC104441.1
AGAAAAATGCCAGCTCCCTGGGG	.**.....**.....	4	0,268601546	0,058441558	chr6	110285887	110285909	+	intergenic:CDC40-METTL24
ATGCAAATGCCAGCTCCCTGAGG	..**.....**.....	4	0,268601546	0,017188694	chr3	12942141	12942163	+	intron:IQSEC1
ACAAAAATGCCCCCTCACTAGGG	.**.....*..*	4	0,264757925	0,267857143	chr12	49277489	49277511	-	intergenic:TUBA1C/RP11-161H23.5-RP11-161H23.9/PRPH
ATTTAAATGTCACCTCTCTGTGG	...*.....*.*....*...	4	0,262235805	0,199643494	chr2	179564324	179564346	-	intron:ZNF385B
ATTATATTGCAGCCTCCCTGAGG	....*.*...**.....	4	0,254561754	0,053030303	chr8	20417863	20417885	-	intergenic:RP11-563N12.2-RP11-108E14.1
ATTCAAATGCCACCTGCCTGGGG	...*.....*...*....	3	0,251620657	0,038784745	chr16	88649592	88649614	-	intron:CYBA
AGTCTAATGTCCCCTCCCTGGAG	.*.**....*.....	4	0,251422414	0,026305966	chr9	135145814	135145836	+	intergenic:OLFM1-RP11-399H11.3
AATAAAATTCTCCCTCCCTCAGG	.*.....*.*.....*	4	0,236507827	0,059509721	chr3	86323028	86323050	-	intergenic:RP11-789F5.1-RN7SKP284
ATTCAACTTCCCGCTCCCTGAGG	...*..*.*...*.....	4	0,231061934	0,011229946	chr11	11377537	11377559	+	intron:GALNT18
ATTACAAGGACCCCTGCCTGAGG	....*..*.*.....*....	4	0,226643675	0,048888889	chr17	78233315	78233337	+	intron:TMEM235
ATTCAAATTCTCCCTCCCTTGGG	...*....*.*.....*	4	0,223894076	0,040542986	chr21	39189548	39189570	-	intergenic:BRWD1/AF129408.17-BRWD1
GTTCAAATGCTCCCACCCTGGGG	*..*.....*...*.....	4	0,223580696	0,062872112	chr1	143787016	143787038	-	intergenic:RP11-403I13.5-RP11-403I13.7
GTTCAAATGCTCCCACCCTGGGG	*..*.....*...*.....	4	0,223580696	0,062872112	chr1	145174118	145174140	+	intergenic:U1-AC241585.2
GTTCAAATGCTCCCACCCTGGGG	*..*.....*...*.....	4	0,223580696	0,062872112	chr1	149695676	149695698	-	intergenic:LINC00869-U1
GTTCAAATGCTCCCACCCTGGGG	*..*.....*...*.....	4	0,223580696	0,062872112	chr1	121003860	121003882	-	intergenic:CH17-118O6.3-CH17-118O6.2
ATGAAAATGGCACCTCCCAGGGG	..*.....*.*.....*.	4	0,222835591	0,03968254	chr4	139843926	139843948	-	intergenic:RN7SKP253-MAML3
ATAATAATGCACCCCCCTGCGG	..*.*.....*...*.....	4	0,220450566	0,05312869	chr2	152437459	152437481	+	intron:FMNL2

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATTAAAATGCTCACTCACTGGGG	.....*.*...*...	3	0,210328283	0,055226825	chr16	69217212	69217234	+	intergenic:Y_RNA-RP11-70O5.2
TTTAAAATGTCCCCTCCCAGAGA	*.....*.....*.	3	0,207615254	0,018674136	chr19	36794830	36794852	+	intergenic:CTD-2162K18.3/CTD-2162K18.4-ZNF790-AS1
ATTTGAATGCCCTCTCCCTGGGA	...**.....*.....	3	0,206835443	0,022272727	chr13	104168330	104168352	+	intergenic:LINC01309-DAOA-AS1/DAOA-AS1_1
ACTCAAATGTCCCTTCCCTGAGG	.*.*.....*...*.....	4	0,206277769	0,208798813	chr15	101360808	101360830	+	intron:PCSK6
AATAAAATGTGCCCACCCTGGGG	.*.....**...*.....	4	0,205917821	0,115265539	chr20	56830272	56830294	-	intergenic:RNU6-929P-AL117380.2
GTAAAAATGTCCCATCCCTGGGG	*.*.....*...*.....	4	0,20338988	0,235294118	chr15	67715908	67715930	+	intron:MAP2K5
ATTAAAATTATTCCTCCCTGTGG	.....****.....	4	0,200516979	0,076581197	chr2	240669485	240669507	-	intergenic:GPR35-AQP12B
ATTTAAATGCCCTCTACCTGGGG	...*.....*..*....	3	0,197921127	0,445454545	chr17	48947160	48947182	+	intergenic:SNF8-RNU1-42P
ATTGAGATGCCACCTCCCTAGGG	...*.*.....*.....*	4	0,196530015	0,298947704	chrX	151892032	151892054	+	intergenic:CNGA2-RP11-366F6.2
ATTAAACTTCCCCCTACCTGAGG	.....*.*.....*....	3	0,191811376	0,233333333	chr3	2582841	2582863	+	intron:CNTN4
AATAAAATGCAGCCTCCCTCTGG	.*.....**.....*	4	0,190444928	0,120879121	chr11	2375159	2375181	+	intergenic:Y_RNA-CD81/CD81-AS1
AATATAATTCCCCCTCCTTGGGG	.*..*...*.....*..	4	0,189613667	0,105494506	chr2	217213796	217213818	+	intergenic:AC010887.1-RN7SKP43
ATTAAAATTAACCCTCCCTTAGG	.....***.....*	4	0,186331124	0,242666667	chr4	144229772	144229794	+	intergenic:GYPA-RP11-361D14.2
ATTAAGATGCCACCTCCTTGGGG	.....*.....*.....*..	3	0,173471962	0,327988338	chr4	151449698	151449720	-	intron:FAM160A1
TTTAAAAGGCACCCTCCTTGTGG	*.....*..*.....*..	4	0,172235	0,353571428	chr5	56205692	56205714	+	intron:ANKRD55
TTTGAAATGCACCCTCCTTGTGG	*..*.....*.....*..	4	0,172235	0,301339286	chr16	62288401	62288423	-	intergenic:RNU6-21P-RP11-75D3.1
GTTCAAATGCTCCCTCCATGGGG	*..*.....*.....*..	4	0,172235	0,058475461	chrX	112704564	112704586	+	intergenic:LHFPL1-AMOT
ACTAAAATACCATCTCCCTGAGG	.*.....*..**.....	4	0,166445786	0,275510204	chr6	159352417	159352439	-	intergenic:FNDC1-RP11-125D12.1

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
AATAAAATAC-CCTCTCCCTAGGG	.*.....*...*.....*	4	0,164916269	0,356971154	chr2	216106764	216106786	+	intergenic:TMEM169-XRCC5
AGTAAAATC-CCCCCTCTCTTTGG	.*.....*.....*..*	4	0,164063988	0,147753846	chr12	67751315	67751337	-	intergenic:RP11-43N5.1-RP11-43N5.2
GATAAAATGC-CCCCACTCTGCGG	**.....*.*...	4	0,163368333	0,228609986	chr4	184175529	184175551	-	intron:ENPP6
ATTTAAATGC-CCTATCCCTGTGG	...*.....**.....	3	0,162310667	0,155909091	chr3	131709446	131709468	+	intron:CPNE4
ATAACAATC-CCCCCTACCTGTGG	..*.*...*.....*....	4	0,155758982	0,192307692	chr4	106022813	106022835	-	intergenic:RP11-710F7.3-TBCK
ATCCAAATTC-CCCCTTCCTGGGG	..**....*.....*....	4	0,155758982	0,024822237	chr10	24308353	24308375	-	intergenic:MIR603-KIAA1217
ATTAAAATGC-CACCTCACAGAGG	.....*....*.*.	3	0,151764217	0,095238095	chr19	3706404	3706426	-	intergenic:PIP5K1C-TJP3
AACAAACTGC-CCCTTCCCTGTGG	.**...*.....*.....	4	0,150830932	0,116346154	chr21	29807256	29807278	+	intergenic:GRIK1-AS1-GRIK1
ATTAAAATGGTG-CCTCCCTAAGG	.....***.....*	4	0,150040774	0,04985755	chr14	103110553	103110575	+	intergenic:EXOC3L4-LINC00677
ATTAGAATG-CAAGCTCCCTGAGG	....*.....***.....	4	0,144239237	0,052597402	chr10	119381344	119381366	+	intron:GRK5
AGTGAAATGCTC-CCTGCCTGGGG	.*.*.....*....*....	4	0,143492089	0,024852071	chr1	225446085	225446107	-	intergenic:LBR-RP11-496N12.6
ATTAGCATGGC-CCCTACCTGAGG	....**...*.....*....	4	0,137119423	0,127272727	chr15	84939652	84939674	-	exon:RNU6-796P
TTTAAAAC-CCCCCATCCCT-GAGG	*.....**....*.....	4	0,136846598	0,188461538	chr2	10569853	10569875	+	intergenic:AC007249.3-NOL10
ATTACAAGGCTC-CCTACCTGGGG	....*..*..*....*....	4	0,136576807	0,112820513	chr13	59189785	59189807	+	intergenic:AL359262.1-RNU7-88P
AGTAATAAGC-CCCATCCCTGGGG	.*...*.*.....*.....	4	0,135502769	0,168	chr3	52353729	52353751	+	intron:DNAH1
AGTAAAAGGC-CCCCACCCCGAGG	.*.....*.....*...*.	4	0,133665	0,098381125	chr6	155696307	155696329	+	intergenic:RNU7-152P-MIR1202
ATTCAAATGT-CACCTCCTTGGGG	...*.....*.*.....*..	4	0,133501865	0,152531601	chr12	85807118	85807140	-	intron:RASSF9
ATTAGAATGC-CCCCCTCCTGTGG	....*.....**....	3	0,129751704	0,060419581	chr13	60480975	60480997	-	intron:TDRD3

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATCCAAATTC-CCCTTCCCTGCGG	..**....*....*.....	4	0,128428059	0,059159664	chr7	87575614	87575636	-	intron:ABCB1
ATTAAAATG-GAAGCTCCCTGAGG	.........****.....	4	0,127005026	0,028409091	chr8	95537416	95537438	-	intergenic:RP11-90D11.1-RNU6-690P
ACTAAAATTTC-CCTTCCCTGTGG	.*.....**...*.....	4	0,126035717	0,315518207	chr2	116096075	116096097	-	intergenic:DPP10-AC012365.1
GTTAAAATGC-CCCACCCCTGCGG	*...........**.....	3	0,125822222	0,095454546	chr11	123525424	123525446	-	intergenic:LINC01059-GRAMD1B
ATAAGAATGC-CACCTTCCTGTGG	..*.*.....*...*....	4	0,125422944	0,113029827	chr20	24303219	24303241	-	intron:RP4-564O4.1
ATTAGAATG-CCACCTCTCTCTGG	....*.....*....*..*	4	0,125064555	0,102857143	chr3	166004855	166004877	-	intergenic:BCHE-LINC01326
AAGAAAATGC-CCACTCCGTGGGG	.**.........*....*..	4	0,118417614	0,021696252	chr11	77183934	77183956	+	intron:MYO7A
AACAAAATGC-CCCCTGCCTTAGG	.**...........*...*	4	0,118281463	0,039053255	chr10	83873958	83873980	-	intergenic:RNU6-129P-RP11-219F10.1
ATGAAAATTC-CCTCTCCCCGTGG	..*.....*...*.....*.	4	0,116281868	0,051494253	chr9	127519917	127519939	-	intron:FAM129B
TTTAAAATGC-CAACTCACTGGGG	*.........**...*...	4	0,116067105	0,128205128	chr12	42366682	42366704	-	intergenic:RP11-351C21.2-PPHLN1
ATTGTAATGC-CCTCTCCGTGTGG	...**.......*....*..	4	0,114018038	0,021212121	chr3	169240924	169240946	+	intron:MECOM
ACTAAACTGC-CCCCACGCTGTGG	.*....*.......*.*...	4	0,111580572	0,012770898	chr22	46880688	46880710	-	intron:TBC1D22A
ATTACAATGACG-CCTTCCTGGGG	....*....*.*...*....	4	0,111508688	0,059259259	chr19	31063210	31063232	-	intergenic:CTC-400I9.3-CTC-439O9.3
GTTTAAATGC-CACTTCCCTGAGG	*..*.......*.*.....	4	0,110193987	0,333333333	chr15	42285220	42285242	-	intron:GANC
ATCAAATTC-CCCCCTTCCTGAGG	..*...*.*.....*....	4	0,106383384	0,031065089	chrX	39083305	39083327	-	intergenic:RNU6-591P-RP11-265P11.1
GTTCAAATGC-CCCCTGTCTGGGG	*..*.........**...	4	0,104848333	0,025339367	chr16	81609661	81609683	-	intron:CMIP
AGTGAAATGC-CCCCTGGCTGTGG	.*.*.........**...	4	0,104848333	0,004751131	chr2	219857836	219857858	+	intergenic:AC009502.4-AC008281.1
GTTAAAATGCAC-CCACCCTCAGG	*.........*...*....*	4	0,103738294	0,186090226	chr19	301466	301488	+	intergenic:PPAP2C-MIER2/CTD-3113P16.5

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATTCAAATGC-CACCTCCCCTGGG	...*.....*.....**	4	0,102325545	0,048681542	chr10	33214197	33214219	-	intergenic:RP11-342D11.2/NRP1-NRP1
CTTAAAATGTC-CCATCCCTATGG	*.....*...*.....*	4	0,095709979	0,264705882	chrX	79720304	79720326	+	intergenic:ITM2A-TBX22
ATTAAAATGTGC-CCTCTCAGTGG	.....**.....*.*.	4	0,088690974	0,031372549	chrY	14474354	14474376	+	intergenic:AC010723.1-NLGN4Y
ATTTATTTGC-CCCTTCCCTGTGG	...*.**.....*.....	4	0,088088228	0,145833333	chr1	91456576	91456598	+	intergenic:HFM1-CDC7
ATTAGAATGTC-CACTTCCTGAGG	....*....*..*..*....	4	0,087711102	0,080194918	chr14	89059653	89059675	+	intergenic:TTC8-FOXN3/RP11-681H18.2
ATAAGAATGC-CCCCACCATGAGG	..*.*.....*..*..	4	0,082004963	0,160323886	chr3	189891944	189891966	+	intron:TP63
CTTAAAATGC-CCCGTGCCTGTGG	*.....*.*....	3	0,080751575	0	chr18	7458422	7458444	-	intergenic:SNORA48-PTPRM
ATTAAAAAGTC-CCTTCTCTGAGG	.....*.*...*..*...	4	0,075589619	0,257673203	chr6	100510662	100510684	+	intron:ASCC3
TTTAAGATGC-CCCCTCGTTGAGG	*....*.....**..	4	0,072284392	0,027010804	chr20	44884872	44884894	-	intergenic:RIMS4-YWHAB
ATTTAAATGCTC-CCTCCATAGGG	...*.....*.....*.*	4	0,071821995	0,098843464	chr12	82618638	82618660	-	intergenic:RP11-263K4.3-TMTC2
ATGAAAATGCTC-CCTCCATTGGG	..*.....*.....*.*	4	0,070816487	0,057988166	chr22	22683663	22683685	-	intergenic:IGLV3-27-IGLV3-25
ATTCAATTGC-CCTCTTCCTGAGG	...*..*.....*..*....	4	0,068338689	0,033257919	chr16	24956441	24956463	-	intron:ARHGAP17
AAAAAAATGC-CCCCATCCTGCGG	.**.....**....	4	0,068319815	0,107665614	chr4	316444	316466	+	intergenic:ZNF732-ZNF141
CTTAAAATG-CCTCCTCCATCAGG	*.....*.....*.*	4	0,067256227	0,106508876	chr3	11605568	11605590	+	intergenic:VGLL4-RP11-169K17.4
ACTAAAATGCA-CACTCCATGAGG	.*.....*.*....*..	4	0,066654945	0,133136095	chr13	20730175	20730197	+	intron:N6AMT2
ATTTAAATGCTC-CCTCTGTGGGG	...*.....*.....**..	4	0,062952983	0,012183372	chr2	35759526	35759548	+	intergenic:RNU6-1117P-RP11-490M8.1
ACTAAAATGC-CAGCTCCATGAGG	.*.....**....*..	4	0,059088708	0,044955045	chr6	111983754	111983776	-	intergenic:AL158035.1-WISP3
ATCAATATGC-CCCATCCCTCTGG	..*..*.....*.....*	4	0,058684981	0,045918367	chr9	107226654	107226676	+	intergenic:RP11-196I18.4-RAD23B

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATTAGAATGCCAGCTCCATGAGG	....*.....**....*..	4	0,056096875	0,037762238	chr22	38210500	38210522	-	intron:MAFF
ATTAAAATGCTGGCTCTCTGAGG	.........***...*...	4	0,055532106	0,008702409	chrX	23302290	23302312	+	intergenic:AC004673.1-PTCHD1
ATAAAAATGTCCCAGCCCTGAGG	..*.....*...**.....	4	0,054508488	0,011764706	chr12	106113067	106113089	-	intron:NUAK1
TTTAAAAAGCCCCCTTCTTGGGG	*.....*.......*.*..	4	0,053377333	0,158241758	chr14	98091814	98091836	-	intergenic:RP11-61O1.1-RP11-61O1.2
AATAAAATGCCTCTTCCCTATGG	.*.........*.*.....*	4	0,051128458	0,313239645	chr2	21791868	21791890	+	intergenic:AC018742.1-RN7SL117P
ATGAAAATCCCCAATCCCTGGGG	..*.....*...**.....	4	0,049701659	0,036242604	chr3	123474189	123474211	+	intergenic:ADCY5-PTPLB
ATTCAAATGCCACCTAACTGAGG	...*.......*...**...	4	0,048973462	0,117647059	chr4	76766793	76766815	+	intron:SHROOM3
ATAAAAATGCTCCCTACCAGAGG	..*.......*....*..*.	4	0,046944126	0,062794349	chr14	94960962	94960984	+	intron:RP11-991C1.1
ATTAAAATGTCAGCTCCATGAGG	.........*.**....*..	4	0,046914397	0,049362402	chr15	61974857	61974879	+	intron:VPS13C
ATTAGAATGCCCTCTTCCTCAGG	....*.......*..*...*	4	0,043948881	0,066461539	chr22	24201091	24201113	+	intergenic:SUSD2-GGT5
ATTTAAATGCCCTCTCTTTGCGG	...*........*...**..	4	0,043896945	0,133636364	chr7	33858904	33858926	-	intergenic:RP11-89N17.4-RP11-89N17.3
AGTAAAATGCCATATCCCTGGGG	.*.........***.....	4	0,042645073	0,147	chr20	1948835	1948857	-	intron:RP4-684O24.5
CTTAAAATGCCAGGTCCCTGTGG	*.........***.....	4	0,042645073	0	chr15	33757643	33757665	+	intron:RYR3
ATAAAAATGACCCATCCTTGAGG	..*.....*...*...*..	4	0,041990519	0,139285714	chr7	44434059	44434081	+	intron:NUDCD3
CTTAAACTGCCCCTGCCCTGGGG	*.....*.....**.....	4	0,040996643	0,01375	chr22	50435821	50435843	-	intron:PPP6R2
TTTAAAATGCCCTTTCCCTCTGG	*..........**.....*	4	0,040216897	0,22	chr8	104148957	104148979	-	intron:RIMS2
CTCAAAATGCCCCTTTCCTGAGG	*.*.........*.*....	4	0,040009579	0,08288854	chr15	64904000	64904022	-	intergenic:PLEKHO2-ANKDD1A
ATTTAACTGCCCCTGCCCTGGGG	...*..*.....**.....	4	0,039020901	0,010208333	chr8	6746615	6746637	+	intergenic:MIR4659B-AGPAT5

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATAAAAATGC-CCACTAACTGAGG	..*.....*..**...	4	0,037982503	0,128205128	chr2	163399506	163399528	+	intergenic:RNU6-627P-FIGN
ATTAACATC-CCACTTCCCTGGGG	.....*..*..*.*.....	4	0,036988125	0,128205128	chr20	17717839	17717861	+	intron:BANF2
ATAAAAATGC-CCCCCCCTTATGG	..*.....*..*.*	4	0,034196069	0,117404917	chr7	39485703	39485725	-	intergenic:POU6F2-AC011290.4
ATGAAAATGC-CCACCCCTTGTGG	..*.....*.*..*..	4	0,031735921	0,033716284	chr20	53287377	53287399	-	intergenic:RP4-678D15.1-RP4-669H2.1
ATTAAAATG-CAAACTCCATGAGG	.....***....*..	4	0,02827089	0,110946746	chr3	34778044	34778066	-	intergenic:AC018359.1-RNU6-243P
ATTAAAATG-CAGACTCCTTGAGG	.....***....*..	4	0,02827089	0,082417582	chr17	5931462	5931484	+	intergenic:NLRP1-WSCD1
ATTAAAATGCCCT-GTCCATGAGG	.....**...*..	3	0,027424906	0	chr18	65316878	65316900	+	intergenic:AC007631.1-RP11-453M23.1
ATTAAGAAGC-CCCTTCCTTGAGG	.....*.*.....*...*..	4	0,026558543	0,269387755	chr2	19696306	19696328	-	intergenic:AC010096.2-AC019055.1
ATGAAAATG-CCTCCTTCATGGGG	..*.....*...*.*..	4	0,025893985	0,044606281	chr3	8275672	8275694	-	intergenic:LMCD1-AS1-RNU4ATA-C17P
ATTAAAATGTC-CCCAACCTCAGG	.....*....**...*	4	0,025328791	0,233524989	chr7	16989821	16989843	+	intergenic:AGR3-Metazoa_SRP
ATTAAAATGCAA-CATCCCTAAGG	.....**.*.....*	4	0,024273697	0,17578125	chr11	61949267	61949289	+	intergenic:RNU6-1243P-BEST1
ATTAAAATGC-CCCTTTGCTGAGG	.....*.**...	3	0,022889966	0,013273001	chr18	39063697	39063719	+	intergenic:RNU6-706P-RP11-244M2.1
ATTCAAATGCCT-CATCCATGAGG	...*.....*.*...*..	4	0,021598022	0,03581622	chr9	84558819	84558841	-	intergenic:SLC28A3-NTRK2
ATTAAAATGCCT-GCTCCTTTGGG	.....**....*.*	4	0,021241372	0,033041958	chr2	38776617	38776639	-	intron:GEMIN6
TTTAAAATGC-CCCATCCTTATGG	*.....*...*.*	4	0,020368248	0,2109375	chr2	47230479	47230501	-	intron:AC106869.2
ATTTAAATGC-CCTCTACTTGGGG	...*.....*..*.*..	4	0,019611103	0,286363636	chr3	58691795	58691817	-	intergenic:FAM3D-C3orf67
ATTAAATTGC-CCCCATGCTGGGG	.....*.....***...	4	0,017342041	0,004584425	chr5	15047642	15047664	-	intergenic:AC016575.1-U8
ATTAAAATGC-CACCTCCTCCAGG	.....*.....***	4	0,017289489	0,054287725	chr19	6501050	6501072	+	intron:TUBB4A

offtargetSeq	mismatchPos	mismatchCount	mitOfftargetScore	cfdOfftargetScore	chrom	start	end	strand	locusDesc
ATTAAAATGCCCTTCTTGCTGTGG	..........**..**...	4	0,016453469	0,006822137	chr1	184561183	184561205	-	intron:C1orf21
ATTAAAATGCCAGCTCCAGGAGG	..........**....**.	4	0,016045641	0,013111888	chr5	176602385	176602407	+	intergenic:CDHR2/GPRIN1-GPRIN1
TTTAAAATGCCCCTTAACTGTGG	*...........*.**...	4	0,015622402	0,342222222	chr4	188228028	188228050	-	intergenic:AC138781.1-LINC01060
ATTAAAATTCCCCCCTGCTGAGG	........*.....***...	4	0,014800608	0,002632661	chr2	237245644	237245666	+	intergenic:AC107079.1-AC112715.2
ATTAGAATGCCCTTTGCCTGAGG	....*.......**.*....	4	0,014189961	0,056861538	chr13	39483540	39483562	-	intron:LHFP
ATTAAAATGCTACCTTCATGAGG	..........**...*.*..	4	0,01256484	0,036413291	chr16	47786225	47786247	+	intergenic:PHKB-AC140542.2
ATTAAAATCCTCCTTTCCTGAGG	........*.*..*.*....	4	0,011862146	0,037384312	chr2	125887008	125887030	+	intergenic:AC097499.1-AC114813.1
ATTAAAATGCCACCCCCTAGAGG	...........*..*..**.	4	0,01111171	0,035780546	chr13	107359696	107359718	-	intergenic:SNORD31-RP11-381O6.1
ATTAAAATGCAACTTTCCTGGGG	..........**.*.*....	4	0,009131982	0,120879121	chr1	48068026	48068048	+	intergenic:RP4-683M8.2-RP11-193P11.3
ATTAAAATCCCCCTTTGCTGAGG	........*....*.**...	4	0,008228696	0,007147001	chr11	106115878	106115900	-	intron:RP11-677I18.4
ATAAAAATGCCCCGTTCTTGAGG	..*..........*.*.*..	4	0,007841878	0	chr5	127925448	127925470	-	intron:LINC01183
ATTAAAATGCCCACTACGTAGGG	............*..*.*.*	4	0,007425845	0,048076923	chr6	28924981	28925003	+	intergenic:TRIM27-LINC01556
ATTAAAATACCCCATGCCAGAGG	........*....*.*..*.	4	0,007057031	0,00989011	chr1	44523887	44523909	-	intergenic:RNF220-MIR5584
ATTAAAATGCCCCAGCACAGTGG	............**.*.*.	4	0,006319527	0,002333333	chr19	58024426	58024448	+	intergenic:CTD-2368P22.1-ZSCAN1
ATTAAAATGCTCCGGACCTGAGG	..........*..***....	4	0,004974331	0	chr6	143906151	143906173	-	intron:ZC2HC1B/RP3-468K18.7
ATTAAAATGCCCCACCCATTGGG	............**..*.*	4	0,004454793	0,035979021	chr6	49636959	49636981	+	intergenic:RHAG-CRISP2

Suppl. Table 5. Summary of efficiencies of deriving edited clones in immortalized myoblasts.

	# of post-FACS grown-out clones	# of edited clones	% efficiency (100*edited/all grown-out)
FSHD1^{3U} poly	40	8	20
FSHD1^{8U} poly	52	7	13,46153846
FSHD2 poly	46	7	15,2173913
FSHD1^{3U} mono	8	5	62,5
FSHD1^{8U} mono	17	3	17,64705882
total	163	30	18,40490798

CHAPTER 3

A homozygous nonsense variant in *LRIF1* associated with facioscapulohumeral muscular dystrophy

Kohei Hamanaka, MD, PhD[1,2,*], Darina Šikrová, MSc[3,*], Satomi Mitsuhashi, MD, PhD[1,4], Hiroki Masuda, MD, PhD[7], Yukari Sekiguchi, MD, PhD[7], Atsuhiko Sugiyama, MD, PhD[7], Kazumoto Shibuya, MD, PhD[7], Richard J.L.F. Lemmers, PhD[3], Remko Goossens, MSc[3], Megumu Ogawa, MSc[1], Koji Nagao, PhD[8], Chikashi Obuse, PhD[8], Satoru Noguchi, PhD[1], Yukiko K Hayashi, MD, PhD[9], Satoshi Kuwabara, MD, PhD[7], Judit Balog, PhD[3], Ichizo Nishino, MD, PhD[1,4,#], Silvère M. van der Maarel, PhD[3,#]

[1]Department of Neuromuscular Research, National Institute of Neuroscience, National Center of Neurology and Psychiatry, Tokyo, Japan.
[2]Department of Neurology, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[3]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
[4]Department of Clinical Development, Medical Genome Center, National Center of Neurology and Psychiatry, Tokyo, Japan
[5]Department of Neurology, National Center Hospital, National Center of Neurology and Psychiatry, Tokyo, Japan
[6]Department of Neurology, Fujita Health University School of Medicine, Aichi, Japan.
[7]Department of Neurology, Graduate School of Medicine, Chiba University, Chiba, Japan.
[8] Department of Biological Sciences, Graduate School of Science, Osaka University, Osaka, Japan
[9]Department of Pathophysiology, Tokyo Medical University, Tokyo, Japan

[*]Shared first authors
[#]Shared senior authors

Hamanaka et al. 2020, Neurology 94(23):e2441-e2447

Abstract

Objective: Facioscapulohumeral muscular dystrophy (FSHD) is a heterogenetic disorder predominantly characterized by progressive facial and scapular muscle weakness. FSHD patients either have a contraction of the D4Z4 repeat on chromosome 4q35 or mutations in D4Z4 chromatin modifiers SMCHD1 and DNMT3B, both causing D4Z4 chromatin relaxation and inappropriate expression of the D4Z4-encoded *DUX4* gene in skeletal muscle. In this study we tested the hypothesis if LRIF1, a known SMCHD1 protein interactor, is a disease gene for idiopathic FSHD2.

Methods: Clinical examination of an idiopathic FSHD2 patient was combined with pathological muscle biopsy examination and with genetic, epigenetic and molecular studies.

Results: A homozygous *LRIF1* mutation was identified in a patient with a clinical phenotype consistent with FSHD. This mutation resulted in the absence of the long isoform of LRIF1 protein, D4Z4 chromatin relaxation, and *DUX4* and DUX4 target gene expression in myonuclei, all molecular and epigenetic hallmarks of FSHD. In concordance, LRIF1 was shown to bind to the D4Z4 repeat and knock down of the LRIF1 long isoform in muscle cells results in *DUX4* and DUX4 target gene expression.

Conclusions: *LRIF1* is a bona fide disease gene for FSHD2. This study further reinforces the unifying genetic mechanism which postulates that FSHD is caused by D4Z4 chromatin relaxation resulting in inappropriate *DUX4* expression in skeletal muscle.

Introduction

Facioscapulohumeral muscular dystrophy (FSHD; MIM: 158900) is an inherited myopathy in which patients typically suffer from asymmetric weakness of facial, scapular-girdle and upper arm muscles. With disease progression, other muscles may become involved.[1] Most patients (FSHD1) have a contraction of the D4Z4 macrosatellite repeat to a size of 1-10 D4Z4 units on one of their chromosomes 4, while European control individuals have 8 to ~100 units.[1,2] D4Z4 repeat contractions are associated with partial D4Z4 chromatin relaxation in somatic cells evidenced by, amongst others, DNA hypomethylation and distinct changes in histone modifications.[3] These epigenetic changes result in expression of the D4Z4-encoded *DUX4* (MIM: 606009) retrogene in skeletal muscle.[2,3] *DUX4* lacks a polyadenylation sequence (PAS) in the D4Z4 unit and requires a distally located PAS that is present only in the 4qA haplotype, but not the 4qB haplotype, nor on chromosome 10, which contains a highly homologous repeat. Hence, FSHD1 patients have a contracted D4Z4 repeat on the 4qA haplotype.[2] *DUX4* encodes for a germline and cleavage stage double homeobox transcription factor and is toxic when expressed in myogenic cells *in vitro* and *in vivo*.[1,4]

In <5% of FSHD patients (FSHD2; MIM: 158901), D4Z4 chromatin relaxation occurs in the absence of D4Z4 repeat contraction. While in FSHD1 cases D4Z4 hypomethylation only occurs on the contracted allele, in FSHD2 pan-D4Z4 hypomethylation is observed on chromosomes 4 and 10.[5] This suggests that trans-acting factors essential for epigenetic repression of the D4Z4 repeat array are defective. Indeed in many FSHD2 patients, heterozygous mutations in the structural maintenance of chromosomes flexible hinge domain containing 1 gene (*SMCHD1;* MIM: 614982), which has a role in epigenetic silencing, are responsible for this pan-D4Z4 hypomethylation.[6] SMCHD1 was reported to compact the inactive X chromosome (Xi) through interaction with the ligand-dependent nuclear receptor-interacting factor 1 (LRIF1 aka HBiX1; MIM: 615354). Together with heterochromatin protein 1 (HP1), LRIF1 and SMCHD1 bridge the H3K9me3 and XIST-H3K27me3 domains to organize the Xi chromatin structure.[7] SMCHD1 was shown to bind to the D4Z4 repeat with reduced binding in FSHD2 patients. Depletion of SMCHD1 in healthy control myotube cultures de-represses *DUX4*. Conversely, *DUX4* de-repression can be partially reversed by increasing SMCHD1 levels in muscle cells.[6,8]

Recently, heterozygous mutations in the DNA methyltransferase 3B (*DNMT3B*) gene were identified in some FSHD2 patients who were negative for a *SMCHD1* mutation. Like *SMCHD1* mutation carriers, these individuals have pan-D4Z4 hypomethylation accompanied by *DUX4* expression in myogenic cells.[9] Since the genetic cause underlying some FSHD2 patients remains unresolved, we speculated that SMCHD1 and LRIF1 together ensure a repressed state of D4Z4 repeat in somatic cells, rendering *LRIF1* a candidate gene for idiopathic FSHD2.

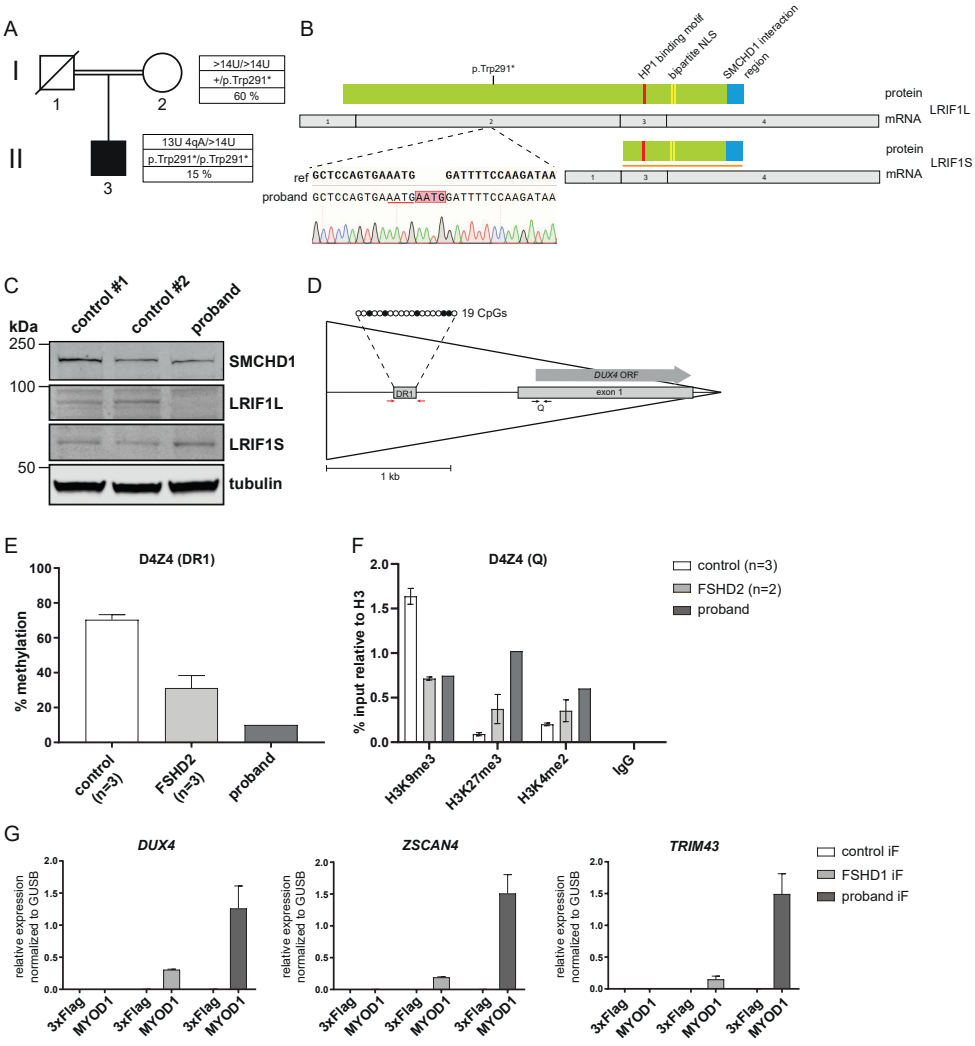


Figure 1. Proband exhibits (epi)genetic and molecular characteristics of FSHD. A) Pedigree of the family. Number of D4Z4 units (U) on chromosome 4, *LRIF1* variant status and D4Z4 DNA methylation as assessed by pyrosequencing from blood DNA are indicated. Alleles with more than 14 D4Z4 units on chromosome 4, which could not be confirmed by Southern blot is described as >14. **B)** Schematic representation of two *LRIF1* mRNA and protein isoforms with zoom in of Sanger sequencing trace showing homozygous 4 nt duplication (underlined AATG) in exon 2 found in proband which leads to a premature stop codon (p.Trp291*). Previously described regions in LRIF1 are also depicted. NLS – nuclear localization signal. Both αLRIF1 antibodies used in this study (western blot and ChIP) recognize amino acid sequence corresponding to the C-terminus of LRIF1 as indicated by the orange line. **C)** Western blot analysis of LRIF1 and SMCHD1 in immortalized fibroblasts from the proband and two independent control individuals showing loss of LRIF1L in the proband's sample, whereas the short isoform (LRIF1S) is still present. SMCHD1 protein levels in the proband were comparable to those in control fibroblasts. Tubulin was used as a loading control. **D)** Schematic representation of a D4Z4 repeat unit as a triangle. Analyzed regions for methylation (DR1 site consisting of 19 CpGs which each can be either methylated – full circle or unmethylated – empty circle) and ChIP-qPCRs (Q) are indicated. Exon 1 of *DUX4* is shown as a bar with *DUX4* open reading frame depicted with a thick arrow over exon 1. Arrows represent the position of primers used for respective analyses. **E)** DNA methylation levels at the D4Z4 DR1 site as assessed by bisulfite PCR followed by TOPO-TA subcloning from

three control, three FSHD2 and proband's immortalized fibroblasts. At least 10 individual colonies were sequenced from each fibroblast line. Bars represent mean methylation ± SEM of all CpGs present in the amplicon from all analyzed colonies. **F)** ChIP for histone modifications (H3K9me3, H3K27me3 and H3K4me2) was performed in three control, two FSHD2 and proband's immortalized fibroblasts followed by qPCR with primers specific for the Q region in D4Z4. Normal anti-rabbit IgG was used as a negative control. Data represent the ChIP enrichment relative to input and normalized to H3 enrichment. The error bars represent mean ± SEM. **G)** Expression analysis of *DUX4* and its two transcription target genes (*ZSCAN4* and *TRIM43*) expression in myotubes using RT-qPCR. Immortalized fibroblasts were transduced either with lentivirus carrying either *MYOD1* induce transdifferentiation towards myogenic lineage or 3xFlag as a negative control. *DUX4* and its target genes' expression level was normalized to *GUSB* expression level. Each sample was analyzed in biological duplicate. The error bars indicate mean ± SD.

Results

We previously reported on 20 Japanese FSHD2 patients showing pan-D4Z4 hypomethylation, of which 13 had an *SMCHD1* mutation.[10] We sequenced *LRIF1* (GenBank: NM_018372.3) in the seven remaining patients and found a homozygous duplication variant c.869_872dup in exon 2 in one patient, which causes frameshift and leads to a premature stop codon (p.Trp291Ter) (Figure 1B). The six remaining patients did not show evidence for *LRIF1* mutations. Among the two LRIF1 protein isoforms, c.869_872dup is predicted to only affect the longer isoform (LRIF1L; Figure 1B). Western blot of immortalized patient's fibroblasts confirmed the selective absence of LRIF1L (Figure 1C). This variant has not been reported in public databases. Subsequent whole-exome sequencing did not identify pathogenic variants in any of the seven patients in *DNMT3B*, *CAPN3, VCP, FHL1* and *FAT1*, genes that were previously reported to cause or mimic FSHD when mutated (Suppl. Table 1).[1]

The patient, a 53-year-old man born from a consanguineous marriage, experienced difficulty in raising his arms. At age 52, he could not walk fast and felt fatigue when climbing stairs. One year later he suffered from aspiration pneumonia. Muscle weakness of the face, scapular girdle, upper arm, thigh and neck were noted. Serum creatine kinase level was 89 IU/L. On muscle CT, asymmetric involvement of biceps brachii, quadriceps femoris, gastrocnemius and paraspinal muscles was documented (Suppl. Figure 1A). Muscle pathological examination identified a few small angular fibers with high alkaline phosphatase enzyme activity (Suppl. Figure 1B), which is often seen in FSHD. Comparison of clinical phenotype and severity of the proband with 13 FSHD2 patients carrying *SMCHD1* variants from our previous study[10] did not reveal any obvious difference (Suppl. Table 2, Suppl. Table 3 and Suppl. Figure 1C).

In addition to the *LRIF1* variant, the proband carries a D4Z4 repeat of 13 units on a 4qA haplotype (Suppl. Figure 1D), consistent with the digenic inheritance of FSHD2 in which a combination of a D4Z4 chromatin factor mutation and an FSHD-permissive chromosome 4 causes disease.

D4Z4 methylation in proband's blood was 15% as determined by bisulfite pyrosequencing (normal range >25%). D4Z4 methylation of the healthy mother, who is a heterozygous mutation carrier, was within the normal range (60%; Figure 1A). D4Z4 chromatin relaxation (Figure 1D) was confirmed in patient immortalized fibroblasts, showing reduced D4Z4 DNA methylation (Figure 1E and Suppl. Figure 1E), a partial loss of H3K9me3, and a gain of

H3K4me2 and H3K27me3 at D4Z4 (Figure 1F). These D4Z4 chromatin changes are typically found in FSHD2.[1,8]

The molecular hallmark of FSHD is the expression of *DUX4* in myotubes. MYOD1-mediated transdifferentiation of immortalized skin fibroblasts of the proband into myotubes, as evidenced by increased expression of *MYOG* and *MYH3* (Suppl. Figure 1F) resulted in *DUX4* and DUX4 target gene expression at levels comparable to FSHD1 myotubes, confirming that the epigenetic changes of D4Z4 observed in the patient correlate with transcriptional derepression of this locus (Figure 1G).

Consistent with earlier data,[7] LRIF1 interacts with SMCHD1 when co-expressed (Suppl. Figure 1G). Chromatin immunoprecipitation (ChIP) studies showed that LRIF1 binds to D4Z4 repeats in control primary myoblasts and myotubes (Figure 2A and 2B), as does SMCHD1, suggesting that both proteins function together in repressing D4Z4 in muscle tissue. In the proband's immortalized fibroblasts, we observed reduced enrichment of LRIF1 and of SMCHD1 at D4Z4 compared to control immortalized fibroblasts (Figure 2C and 2D) with the SMCHD1 enrichment in the proband being comparable to that in FSHD2 immortalized fibroblasts, despite the normal SMCHD1 protein levels in the proband (Figure 1C).

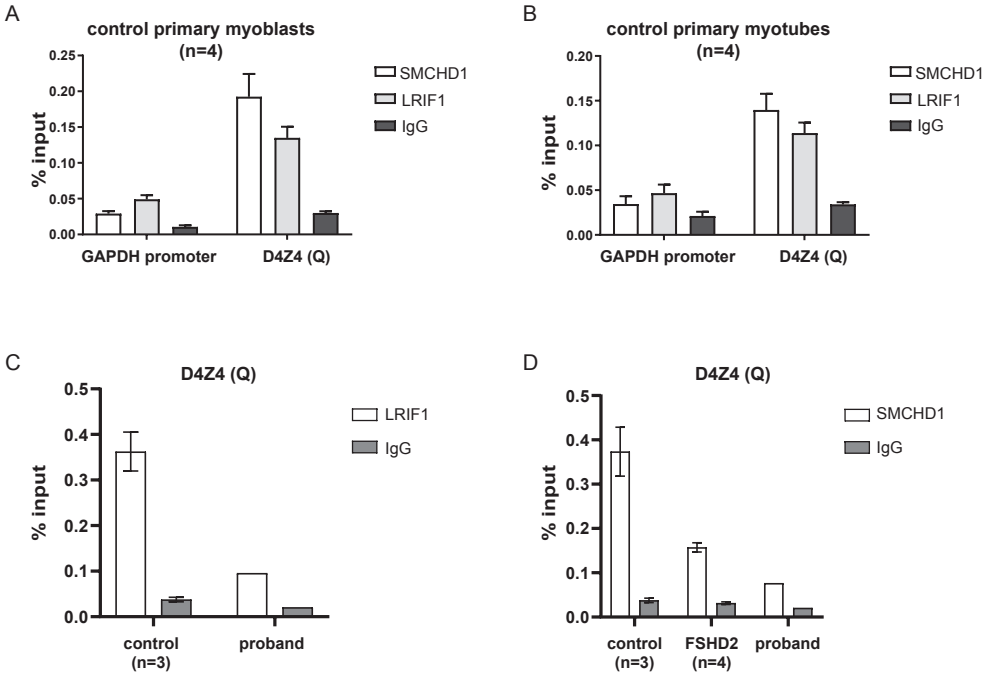


Figure 2. LRIF1 binds to D4Z4 in myogenic cells and its binding is together with SMCHD1 reduced in the proband. A) ChIP was performed with antibodies specific for SMCHD1, LRIF1 or normal anti-rabbit IgG (negative control) in four control primary myoblasts followed by qPCR with primers specific for either Q region or *GAPDH* promoter region, which served as a negative control region for SMCHD1 and LRIF1 enrichment. Data represent the ChIP

values as relative to input in %. Bars represent the mean ± SEM. **B)** ChIP was performed as in (A) but with four control primary myoblasts after their differentiation to myotubes. **C)** LRIF1 ChIP-qPCR for Q region of D4Z4 from three control immortalized fibroblasts and proband's fibroblasts showing reduced LRIF1 enrichment at this region in proband as compared to control fibroblasts. Bars represent the mean ± SEM and enrichment is shown as relative to input in %. **D)** SMCHD1 ChIP-qPCR for Q region of D4Z4 in three control, four FSHD2 and proband's immortalized fibroblasts showing reduced SMCHD1 binding to this region in proband in comparison to control fibroblasts as observed also for FSHD2 samples.

To confirm that the loss of the LRIF1L leads to *DUX4* de-repression in myogenic cells, we performed siRNA knock-down experiments in control, FSHD1 and FSHD2 immortalized myoblasts. When sufficiently reduced, as confirmed by western blot analysis, two out of three independent siRNAs mediated knock-downs of LRIF1L resulted in *DUX4* de-repression in all genetic situations, along with the transcriptional upregulation of well-established DUX4 biomarkers (Figure 3).

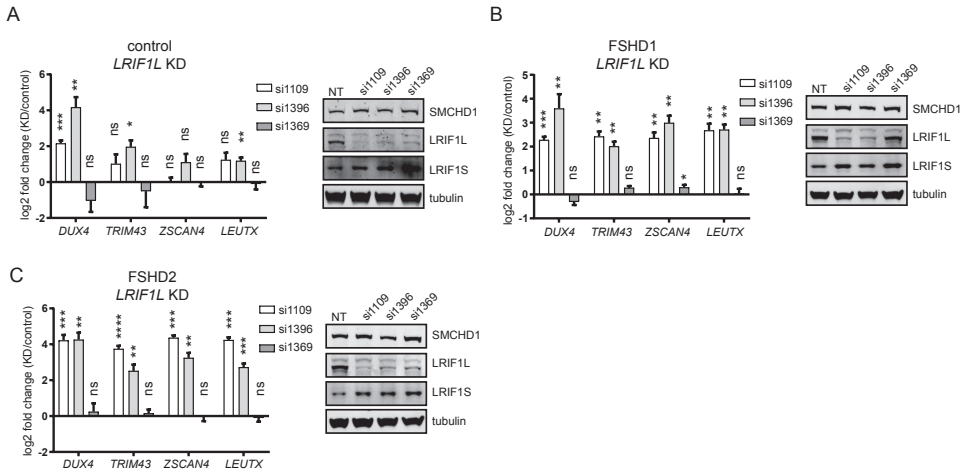


Figure 3. siRNA-mediated depletion of LRIF1L in immortalized myoblasts de-represses *DUX4* locus. RT-qPCR analysis of *DUX4* and its three transcriptional target genes (*TRIM43*, *ZSCAN4* and *LEUTX*) after siRNA-mediated knock-down of LRIF1L in control **A)**, FSHD1 **B)** and FSHD2 **C)** immortalized myoblasts. Expression levels from LRIF1L-specific siRNA treated samples were normalized to those measured in the sample treated with non-targeting (NT) siRNA and were further log2 transformed. *GUSB* mRNA level was used for the RT-qPCR normalization within the samples. Bars represent the mean ± SEM of four independent experiments. Statistical significance was calculated with one-sample t-test comparing LRIF1L-specific siRNA samples vs non-targeting siRNA: ns: not significant, *p-value<0.05, **p<0.01, ***p-value<0.001, ****p-value<0.0001. A representative western blot from one of the four experiments is shown to confirm the successful downregulation of LRIF1L, while LRIF1S and SMCHD1 levels are not decreased. Tubulin was used as a loading control. Note that the least DUX4-responsive siRNA (si1369) also resulted in the mildest LRIF1L protein knock-down.

Discussion

This study identifies *LRIF1* as an FSHD2 disease gene in a patient having a phenotype that is consistent with idiopathic FSHD. *LRIF1* mutations are, like *DNMT3B* mutations, likely a rare

cause of FSHD and should only be considered in FSHD2 when tested negative for *SMCHD1* mutations. Interestingly, while almost all FSHD2 patients show mono-allelic mutations in *SMCHD1* or *DNMT3B*, this patient has a bi-allelic *LRIF1* mutation. This may suggest that a complete loss of full-length LRIF1 is required to de-repress *DUX4* in skeletal muscle and to cause disease. By showing its involvement in D4Z4 chromatin regulation, like the previously identified FSHD2 disease genes *SMCHD1* and *DNMT3B*, this study reinforces the uniform disease mechanism for FSHD that postulates that the disease is caused by inappropriate expression of *DUX4* in skeletal muscle as a result of partial chromatin relaxation of the D4Z4 repeat. FSHD2 patients should therefore equally qualify for current and future therapeutic trials targeting DUX4 expression or function.

Acknowledgement

We thank Ms. K. Goto and Ms. M. Arai in NCNP for technical assistance and Dr. Yasushi Oya for his critical comments. This study was supported partly by Intramural Research Grant (29-4, 29-3) for Neurological and Psychiatric Disorders of NCNP; and AMED under Grant Numbers JP19ek0109285h0003, 18ek0109348s0501, 18kk0205001s0203; and JSPS KAKENHI under Grant Numbers JP19H03156, JP18H04713, JP18H05532, JP17H06426. This study was also supported by funds from the National Institute of Neurological Disorders and Stroke (P01 NS069539) and the Prinses Beatrix Spierfonds (W.OP14-01 and W.OR15-26). These sponsors have no role in study except funding.

Disclosures

All authors approved the final article and declared no conflict of interest.

References

1. Neguembor, M. V., Previtali, S. C. & Gabellini, D. Facioscapulohumeral Dystrophy. in *The Online Metabolic and Molecular Bases of Inherited Disease* (eds. Beaudet, A. L. et al.) (The McGraw-Hill Companies, Inc., 2014). doi:10.1036/ommbid.216.1.
2. Lemmers, R. J. L. F. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–3 (2010).
3. Himeda, C. L. & Jones, P. L. The Genetics and Epigenetics of Facioscapulohumeral Muscular Dystrophy. *Annu. Rev. Genomics Hum. Genet.* **20**, 265–291 (2019).
4. Bosnakovski, D. *et al.* p53-independent DUX4 pathology in cell and animal models of facioscapulohumeral muscular dystrophy. *Dis. Model. Mech.* **10**, 1211–1216 (2017).
5. de Greef, J. C. *et al.* Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum. Mutat.* **30**, 1449–1459 (2009).
6. Lemmers, R. J. L. F. *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat. Genet.* **44**, 1370–1374 (2012).
7. Nozawa, R.-S. *et al.* Human inactive X chromosome is compacted through a PRC2-independent SMCHD1–HBiX1 pathway. doi:10.1038/nsmb.2532.
8. Balog, J. *et al.* Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. *Epigenetics* **10**, 1133–1142 (2015).
9. van den Boogaard, M. L. *et al.* Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am. J. Hum. Genet.* **98**, 1020–1029 (2016).
10. Hamanaka, K. *et al.* Clinical, muscle pathological, and genetic features of Japanese facioscapulohumeral muscular dystrophy 2 (FSHD2) patients with SMCHD1 mutations. *Neuromuscul. Disord.* **26**, 300–308 (2016).

Web Resources

The URLs for data presented are as follow:
Ensembl, http://www.ensembl.org/index.html
OMIM, http://www.omim.org/
1000 Genomes, http://www.1000genomes.org/
ESP6500, http://evs.gs.washington.edu/EVS/
dbSNP138, http://www.ncbi.nlm.nih.gov/projects/SNP/
HGVD, http://www.genome.med.kyoto-u.ac.jp/SnpDB/
Exome Aggregation Consortium (ExAC), http://exac.broadinstitute.org/

Supplementary Information

Materials and Methods

Subjects

Family in Figure 1A was studied after informed consent and after the study protocol had been approved by the relevant institutional review board. Additional clinical information about the proband in comparison to the 13 affected *SMCHD1* mutation carriers can be found in Suppl. Tables 2 and 3.

Of the seven *SMCHD1* mutation negative patients, only the proband in this study had a homozygous mutation in *LRIF1*. Of the six remaining patients, no one had a pathogenic or likely pathogenic variant in *DNMT3B* or *LRIF1* according to the guidelines from the American College of Medical Genetics and Genomics[1]. One patient with a shortened D4Z4 repeat (7 units with A haplotype) had a rare heterozygous truncating variant in *LRIF1* (c.2148_2149del; p.(His718PhefsTer4). However, the *LRIF1* variant or combination of the *LRIF1* variant and shortened D4Z4 repeat did not segregate with FSHD phenotype or D4Z4 hypomethylation status in the family. Therefore, we excluded the possibility that a heterozygous *LRIF1* variant is pathogenic in the family.

Quantification of D4Z4 methylation by pyrosequencing

DNA (500 ng) extracted from the patients' blood was subjected to bisulfite treatment using EpiTect DNA bisulfite kit (QIAGEN) according to the manufacturer's protocol. The level of D4Z4 methylation was quantified using the pyrosequencing technique. Briefly, the polymerase chain reaction (PCR) was performed by PyroMark PCR Kit (QIAGEN), and 10 µl of the biotinylated PCR product was subjected to affinity purification using Streptavidin Sepharose High Performance (GE Healthcare Life Science) and PyroMark Q24 Advanced CpG Reagents (QIAGEN). PCR primers and sequencing primers were designed using PyroMark Assay Design 2.0 Software (QIAGEN). We designed two sets of PCR primers targeting a different sequence within the DR1 region that is reported to be highly hypomethylated in FSHD2 patients [19]. The primer sequences used are as follows: forward primer 1, GAAGGCAGGGAGGAAAAG; biotinylated reverse primer 1, GCTCAGCCTGGGGATGTGCGGTCT; sequencing primer 1, GGTAGGAGGGGTATTATTT; forward primer 2, TAGGGAGGAAAGGAGGGAAAGATAG; biotinylated reverse primer 2, ACTATAAACCCAACCCTCAAC; sequencing primer 2, GGTTTTAGGGAGTAG. Pyrosequencing was performed using PyroMark Q24 Advanced System (QIAGEN). The seven CpG methylation sites were quantified by PyroMark Q24 Advanced Software. Delta 1 score was calculated as previously described.[2]

Methylation analysis of D4Z4 DR1 region by bisulfite PCR and TOPO-TA subcloning

500 ng of genomic DNA was treated with bisulfite reagent using EZ DNA Methylation-Lightning kit (Zymo Research, #D5030) according to manufacturer's protocol. DR1 region was amplified from converted DNA with following primers: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGTTGAGGGTTGGGTTTATA-3' and 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAAAACTCAACCTAAAAATATAC-3' using FastStart™ Taq DNA polymerase (Sigma-Aldrich, #12032902001) according to manufacturer's instructions with the following cycling conditions: initial denaturation at 95°C for 4 min followed by 35 cycles of 95°C for 4 min, 58°C for 20 s and 72°C for 40 s, followed by the final extension step at 72°C for 5 min. PCR products were run on 2% TBE agarose gel, products were purified from the gel using NucleoSpin Gel & PCR Clean-up kit (Bioke, #740609) and subcloned into TOPO-TA vector. Plasmid DNA was isolated from at least 10 independent bacterial colonies for every sample and sent for Sanger sequencing. Methylation levels from Sanger sequencing tracks were analysed using BiQ Analyzer software.

Sanger sequencing of LRIF1

LRIF1 protein coding sequences were amplified from gDNA using partially overlapping primer sets: 5'-GGAAACTCGGCCCCACGC-3' and 5'-CGGGCTCCAACTCCTCTC-3' for exon 1; 5'-CAGCCAGCCAGTTCTTCAAA-3' and 5'-ACTGGCTTTGCTATTTCTGT-3' for the second quartile of exon 2, 5'-CCCAAATGCCAACCGTTATT-3' and 5'-TGGAGTATCAGGAGAAACAGA-3' for the third quartile of exon 2, and 5'-TGGGAAAGTCTATCTGTTGGCT-3' and 5'-AGTCTGTGTGTGATGGGGTT-3' for the fourth quartile of exon 2; 5'-GTGGGTGGTAAGGCAAGGAT-3' and 5'-CTGGGGCCTGGTTGTTTTAA-3' for exon 3; 5'-GGTAGTACCCGGTGCATTTAG-3' and 5'-AGACACTTTCAGAACACACCT-3' for exon 4 using PCR Master Mix (Promega) according to the manufacturer's instructions in 5 min/95', followed by 35 cycles of 30 sec/95', 30 sec/54', and 1 min/72', and 7 min/72'. The first quartile of exon 2 was amplified using primer set: 5'-TCTCATACCCATTGCCTAATCA-3'and 5'-CACACCATGACTCTGAACTT-3' using TaKaRa Ex Taq (Takara Bio Inc.) according to the manufacturer's instructions in 5 min/98', followed by 40 cycles of 10 sec/98', 30 sec/54', and 1 min/72', 7 min/72'. PCR products were directly sequenced with ABI PRISM 3100 automated sequencer (PE Applied Biosystems).

Whole-exome sequencing

Whole-exome sequencing was carried out as previously described. Briefly, genomic DNA was extracted from peripheral blood lymphocytes, then subjected to solution capture (SureSelect Human All Exon V5, Agilent Technologies) to generate barcoded whole-exome sequencing libraries. Libraries were sequenced on an Illumina HiSeq 1000 sequencer employing paired end 100-base reads to a mean target coverage of 170X, and 184X, resulting in 94.10% and 95.50% of the target covered by ≥30 reads. Alignment, variant calling, and annotation were performed with a custom informatics pipeline employing BWA, Picard (http:// picard.sourceforge.net), GATK (ver. 1.6), and ANNOVAR. Known polymorphisms were detected using public database; NHLBI ESP with 6800 exomes, 1000 Genomes Project, dbSNP138, and HGVD for Japanese genetic variants. The human genome reference used for these studies was hg19.

D4Z4 repeat size analysis

D4Z4 repeat size was analyzed as previously described.[3] Briefly, genomic DNA was digested with EcoRI (Takara) or EcoRI/BlnI (Takara) in linear gel electrophoresis (LGE). The digested DNA was electrophoresed in Gel Electrophoresis Apparatus GNA-200 (Amersham), transferred to Hybond-XL (GE Healthcare), and hybridized with ^{32}P-labeled p13E11 probe. The membrane was washed twice in 2×SSC, 0.1% SDS for 20min. Digested bands in EcoRI with 3kb shorter band in EcoRI/BlnI were regarded as 4q-type D4Z4. D4Z4 units were calculated as follows: D4Z4 unit = (D4Z4 length in EcoRI digestion (kb) - 6.6) ÷ 3.3. To determine haplotype of D4Z4, genomic DNA was digested with Hind III (Takara), hybridized with 4qA probe, and washed twice in 1×SSC, 0.1% SDS for 15min, followed by autoradiography.

Cell lines and culturing conditions

Human primary control myoblast lines (1926, 2333, 2417 and 2081) were received from the Fields center biorepository hosted at the University of Rochester (https://www.urmc.rochester.edu/neurology/fields-center.aspx). Immortalized myoblast lines 2401 (control), 073 (FSHD1) and 2440 (FSHD2) were gift from Prof. S. Tapscott. Myoblasts were cultured in Ham's F-10 Nutrient Mix (Gibco, #31550) supplemented with 20% heat-inactivated foetal bovine serum (FBS) (Gibco, #10270106), 1% penicillin/streptomycin (Gibco, #15140), 10 ng/ml FGF-b (Promokine, #C-60240) and 1 µM dexamethasone (Sigma-Aldrich, #D2915). Myogenic differentiation was achieved by switching myoblasts at 100% confluency to DMEM (Gibco, #31966021) supplemented with 2% KnockOut™ serum replacement (Gibco, #10828028). Human primary control and FSHD2 fibroblasts were obtained Fields center biorepository hosted at the University of Rochester. Primary fibroblasts were immortalized by retroviral transduction of hTERT (Addgene #1773) and maintained in DMEM/F-12 GlutaMAX™ Supplement (Gibco, #10565018) supplemented with 20% heat-inactivated FBS, 1% penicillin/streptomycin, 10 mM HEPES (Gibco, #15630056) and 1 mM Sodium Pyruvate (Gibco, #11360070). HEK293T cells were maintained in DMEM (Gibco, #31966021) supplemented with 10% heat-inactivated FBS and 1% penicillin/streptomycin. All cell lines were grown at 37°C and 5% CO_2 and were regularly tested for *Mycoplasma* contamination with MycoAlert™ Mycoplasma detection kit (Lonza, #LT07-318) according to vendor's instructions. Detailed information about used cell lines can be found in Suppl. Table 5.

Whole-cell extract (WCE) preparation and western blot analysis

Cells were washed twice with ice-cold PBS and then lysed in ice-cold RIPA buffer (0.1% SDS, 1% Igepal CA-630, 150mM NaCl, 0.5% Sodium Deoxycholate, 20mM EDTA) supplemented with Complete™, EDTA-free Protease Inhibitor Cocktail (1 tablet/50 ml buffer) (Sigma-Aldrich, #11873580001). Cell lysates were cleared by centrifugation at 13,300 rpm for 10 min at 4°C and supernatant was used to measure protein concentration using Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, #23225). All samples were diluted to the same concentration and denatured by mixing with 6x Laemmli buffer (60% glycerol, 12% SDS, 12% DTT, 0.02% bromphenol blue in 360 mM Tris-HCl pH 6.8) to a final concentration of 1x and boiled at 95°C for 5 min. Samples were resolved on Novex™ NuPAGE™ 4-12% Bis-Tris protein gels (Invitrogen, #NP0321BOX) and transferred to Immobilon-FL PVDF membrane (Merck, #IPFL00010). Membrane was blocked in 4% skim milk in PBS and incubated overnight at 4°C with primary antibodies in Immunobooster solution I (Takara, #T7111A): αSMCHD1 (1:1000, Abcam #ab176731), αLRIF1 (1:1000, Proteintech #26115-1-AP), α-αTubulin (1:4000, Sigma-Aldrich #T6199), αFlag (1:2000, Sigma-Aldrich #F3165). Next day, membranes were washed twice with PBS containing 0.01% Tween 20 and incubated with secondary antibodies IRDye® 800CW goat anti-rabbit IgG and IRDye® 680CW goat anti-mouse IgG (1:10,000, Li-cor #P/N 925-32211 and #P/N 925-68072, respectively) for 1h at room temperature. Membranes were washed again twice and developed using Odyssey® CLx Imaging System (Li-cor).

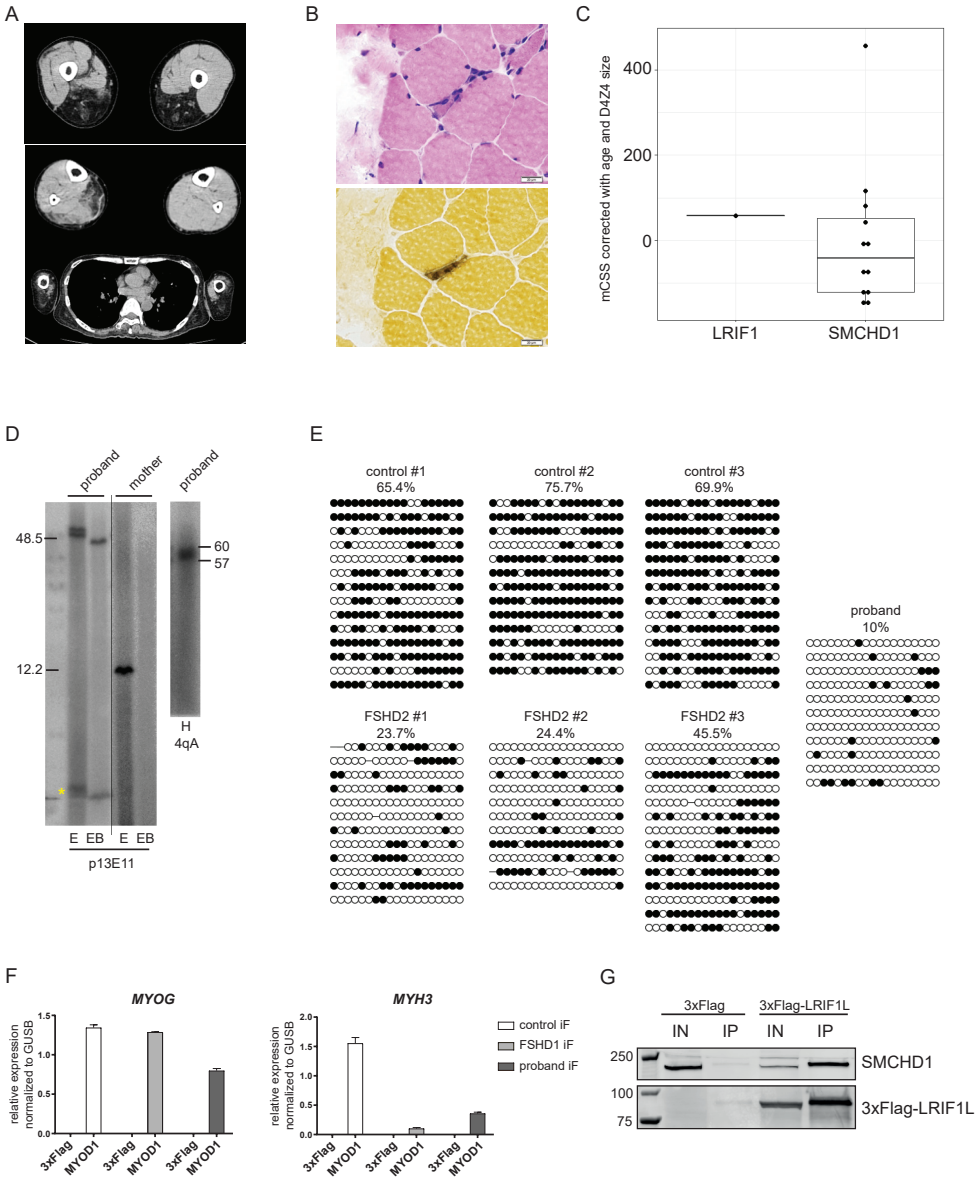
MYOD1-mediated transdifferentiation of human fibroblasts into myogenic cells
Fibroblasts were transduced at 80-90% confluency in the presence of 8 µg/ml polybrene (Sigma-Aldrich, #107689) with lentiviral particles at 15 ng/cm^2 containing either CMV driven *MYOD1* or 3xFlag (pRRL-CMV backbone). A day after transduction, cells were washed once with DPBS (Gibco, #14190) and switched to DMEM (Gibco, #31966021) supplemented with 10% KnockOut™ serum to induce myogenic differentiation.

RNA isolation, cDNA synthesis and qPCR
For RNA expression studies, cells were harvested in QIAzol lysis reagent (Qiagen, #79306) and RNA was isolated with RNeasy mini kit (Qiagen, #74101) with DNase I treatment according to manufacturer's enclosed protocol. cDNA was synthesized from 1-2 µg of RNA and poly-dT primer using RevertAid H Minus First Strand cDNA synthesis kit (Thermo Fisher Scientific, #K1621). Gene expression was measured with CFX384 system in technical triplicates using iQ™ SYBR® Green Supermix (Biorad, #1708887). qPCR primers are listed in Suppl. Table 6 and for every experiment *GUSB* was used as a housekeeping gene control.

Chromatin immunoprecipitation followed by qPCR
Chromatin isolation and chromatin immunoprecipitation was done according to previously published protocol[4]. Shortly, for crosslinking of DNA-protein complexes, formaldehyde was added to the cells to a 1% final concentration and cells were incubated at room temperature (RT) for 10 min. Crosslinking was quenched by adding glycine to a 125 mM final concentration followed by incubation at RT for 5 min. Cells were washed twice with ice-cold PBS containing 0.5 mM PMSF (Sigma-Aldrich, #93482) and harvested by scraping them in PBS with 0.5 mM PMSF followed by centrifugation at 2,500 g for 5 min. To isolate nuclei, cell pellet was resuspended in the ice-cold ChIP buffer (1.5 ml lysis buffer/10 x 10^6 cells) (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5 % Igepal CA-630, 1% Triton X-100) supplemented with cOmplete™ Protease Inhibitor Cocktail table (Sigma-Aldrich, #11697498001), incubated on ice for 10 min and spun down at 8,000 g for 2 min at 4°C. Supernatant was decanted and the same procedure was repeated with the pelleted nuclei. After the second centrifugation, pelleted nuclei were again resuspended in the NP buffer and sonicated at the highest power output for 25 cycles (1 cycle: 30 sec ON/30 sec OFF) using a Bioruptor instrument (Diagenode). Chromatin was pre-cleared with BSA pre-blocked protein A Sepharose beads (GE Healthcare, #17-5280-21) by rotating for 1 h at 4°C and protein-DNA complexes were immunoprecipitated by overnight incubation at 4°C with the following antibodies: αSMCHD1 (Abcam, #ab31865), αLRIF1 (Merck, #ABE1008), αH3 (Abcam, ab1791), αH3K4me2 (Active Motif, #39141), αH3K9me3 (Active Motif, #39161) or αH3K27me3 (Merck, #07-449). Isotype rabbit polyclonal IgG was used as a negative control (Abcam, #ab37415). Next day, the immunocomplexes were pulled down by incubating them for 2 h at 4°C while rotating with BSA pre-blocked protein A Sepharose beads. Beads were then washed with the following buffers: once with low salt wash buffer (1 % Triton X-100, 0.1 % SDS, 2 mM EDTA, 20 mM Tris-HCl, 150 mM NaCl), high salt wash buffer (1 % Triton X-100, 0.1 % SDS, 2 mM EDTA, 20 mM Tris-HCl, 500 mM NaCl), LiCl wash buffer (250 mM LiCl, 1% Igepal CA-630, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl) and twice with TE wash buffer (10 mM Tris-HCl, 1 mM EDTA). After the final wash, 10% (w/v) of Chelex 100 resin was added to the beads and boiled at 95°C while shaking. Supernatant was then further diluted once with MQ and subjected to qPCR analysis with primers amplifying either GAPDH promoter region: 5'- CTGAGCAGTCCGGTGTCACTAC-3' and 5'- GAGGACTTTGGGAACGACTGAG-3' or Q region in exon 1 of *DUX4* 5'-CCGCGTCCGTCCGTGAAA-3' and 5'-TCCGTCGCCGTCCTCGTC-3' as described previously.[5]

siRNA transfections
24h after seeding (1.5 x 10^5/well in 6-well plate), myoblasts were transfected with Stealth RNAi™ siRNAs (Thermo Fisher Scientific) at 10nM final concentration using Lipofectamine RNAiMAX (Invitrogen, #13778075) according to the manufacturer's instructions. 24h after transfection, medium was changed and cells were harvested 72h after transfection for subsequent analysis. All siRNAs used in this study are listed in Suppl. Table 7.

Data availability
All data apart from whole-exome sequencing is contained within the article or supplemental data. WES data is not available due to participants' privacy and consent.



Suppl. Figure 1.

A) Computed tomography analysis of the proband. The upper, middle, and lower panel shows upper arms and thoracic, thighs, and lower limbs, respectively. Bilateral triceps brachii, serratus anterior, and latissimus dorsi muscles are replaced with adipose tissue. Biceps brachii, paraspinal muscle, and gastrocnemius are asymmetrically involved and replaced with adipose tissue. In the legs, bilateral adductor magnus, gracilis, and hamstring muscles are also replaced with fat. The vastus lateralis, vastus medialis, and soleus are partially atrophic with adipose tissue infiltration, only on the right side.

B) Histological analysis of the rectus femoris muscle. The upper and lower panel shows hematoxylin and eosin staining and alkaline phosphatase staining, respectively. Muscle pathology was almost normal except for scattered small angular fibers and type 2C fibers. White bar indicates 20 µm.

C) Comparison of severity among FSHD2 patients. Box plot and dot plot of mCSS corrected with age and D4Z4 size between the patient with *LRIF1* variants and 12 patients with *SMCHD1* variants. The mCSS corrected with age and D4Z4 size was calculated in Table S2. The 12 patients with *SMCHD1* variants were previously reported (Hamanaka et al., 2016). The severity of the patient with *LRIF1* variants was not obviously different from those of patients with *SMCHD1* variants.
D) Southern blot analysis for the D4Z4 repeat. Each panel shows the size of D4Z4 repeat after digestion with restriction enzymes and hybridizion with probe indicated below the panel: E, EB, and H indicates EcoRI, EcoRI and BlnI, and HindIII, respectively. D4Z4 repeats on chromosome 4 shows 3kb reduced size after EB digestion than after E digestion. D4Z4 repeats show 7kb increased size in H digestion than after E digestion. The asterisks indicate the molecular size of a non-specific band derived from the Y chromosome. Marker size is indicated on the right or left side of each panel. In the panel on the right, 57 and 60 indicate the size of 13 and 14 D4Z4 unit array digested by *HindIII*, respectively. DNA extracted from lymphocyte was used in each analysis.
E) Lollipop representation of DR1 methylation data from Figure 1E from three control, three FSHD2 and the proband's immortalized fibroblasts. DR1 site consists of 19 CpGs. Filled circles represent methylated cytosines and open circles represent unmethylated cytosines. Mean methylation in % (methylated cytosines in CpG context/ all CpGs present at the analyzed region) is indicated for every sample. Methylation level for every sample was assessed by analyzing at least 10 independent clones.
F) Expression analysis of *MYOG* and *MYH3* in transdifferentiated fibroblasts by RT-qPCR. MYOD1-mediated myogenic transdifferentiation of immortalized skin fibroblasts was confirmed by measuring mRNA levels of early (MYOG) and late (MYH3) myogenic factors. Expression was normalized to *GUSB* mRNA levels. Bars represent the mean ± SD of two experiments.
G) Anti-Flag co-immunoprecipitation was carried out from HEK293T cells transfected with either 3xFlag or 3xFlag-tagged LRIF1L. Immunoprecipitate was analyzed by western blot with αSMCHD1 and αFlag antibodies. IN – input, IP – immunoprecipitated proteins.

Suppl. Table 1. List of causative genes for myopathy analyzed by whole exome analysis.

ABHD5	CFL2	DNMT3B	HADHA	MYBPC3	SGCA
ACADL	CHAT	DOK7	HADHB	MYH7	SGCB
ACADM	CHKB	DOLK	HSPG2	MYOT	SGCD
ACADS	CHRNA1	DPAGT1	ISCU	NEB	SGCG
ACADVL	CHRNB1	DPM1	ISPD	PFKM	SLC22A5
ACTA1	CHRND	DPM2	ITGA7	PGAM2	SLC25A20
ACVR1	CHRNE	DPM3	KBTBD13	PGK1	SMCHD1
AGL	CHRNG	DYSF	KCNA1	PGM1	SYNE1
AGRN	CLCN1	EMD	KCNE3	PHKA1	SYNE2
ALDOA	CNBP	ENO3	KCNJ12	PLEC	TAZ
ALG13	CNTN1	ETFA	KLHL40	PNPLA2	TCAP
ALG14	COL12A1	ETFB	KLHL9	POMGNT1	TMEM43
ALG2	COL6A1	ETFDH	LAMA2	POMGNT2	TMEM5
ANO5	COL6A2	FAT1	LAMB2	POMK	TNNT1
ATP2A1	COL6A3	FHL1	LARGE	POMT1	TNPO3
B3GALNT1	COLQ	FKRP	LDHA	POMT2	TPM2
B3GALNT2	CPT2	FKTN	LMNA	PRKAG2	TPM3
B3GNT1	CYR61	FLNC	LPIN1	PTPLA	TRAPPC11
BIN1	DAG1	GAA	LRP4	PTRF	TRIM32
CACNA1A	DES	GBE1	MEGF10	PYGM	VCP
CACNA1S	DMD	GFPT1	MICU1	RAPSN	
CAPN3	DMPK	GMPPB	MTM1	RYR1	
CAV3	DNAJB6	GYG1	MTO1	SCN4A	
CCDC78	DNM2	GYS1	MUSK	SEPN1	

Suppl. Table 2. A modified version of clinical severity score (CSS).

Grade	Criteria
1	Only facial muscle weakness
2	Scapular girdle weakness but able to put hands together above the head
3	Unable to put hands together above the head, but able to raise both hands above the head
4	Unable to raise both hands above the head
5	Tibioperoneal weakness and no weakness of pelvic and proximal leg muscles
6	Strength of all pelvic and proximal leg muscles >4 in MMT and able to climb upstairs
7	Strength of all pelvic and proximal leg muscles >3 in MMT and able to climb upstairs
8	Unable to climb upstairs, but able to stand up from a chair
9	Unable to stand up from a chair, but able to walk
10	Unable to walk

MMT: manual muscle testing

Suppl. Table 3. mCSS corrected with age and D4Z4 size in FSHD2 patients. We evaluated severity of FSHD2 patients including 13 with *SMCHD1* variants in a previous study (Hamanaka et al., 2016) and 1 with *LRIF1* variants in this study using mCSS (Table S1). We corrected mCSS with age as previously (age-mCSS, van Overveld PG et al., 2005). Furthermore, we predicted age-mCSS from D4Z4 size using linear regression, and the residual was considered as age-mCSS corrected with D4Z4 size. [1]The patient ID is the same as that in the previous publication (Hamanaka et al., 2016). [2]The haplotype is A except Patient ID 5 and 10 whose haplotype was not confirmed (Hamanaka et al., 2016). mCSS: the modified version of clinical severity scoring; age-mCSS: age-corrected mCSS; hetero: heterozygous; hom: homozygous; NA: not analyzed because D4Z4 size was not definite. The formula for age-mCSS: (mCSS * 2 / (age at examination)) * 1000; the formula for prediction of age-mCSS: 924.43 - 52.39 * (D4Z4 size); the formula for residual of age-mCSS: (age-mCSS) – (Prediction of age-mCSS).

Study	Patient ID[1]	Age (year)	D4Z4 size (unit)[2]	Causative gene	Zygosity	mCSS	Age-mCSS	Prediction of age-mCSS	Residual of age-mCSS
Hamanaka et al., 2016	1	45	9	*SMCHD1*	Hetero	7	311.1	452.9	-141.8
	2	48	9	*SMCHD1*	Hetero	8	333.3	452.9	-119.6
	3	37	13	*SMCHD1*	Hetero	6	324.3	243.4	81.0
	4	63	14	*SMCHD1*	Hetero	6	190.5	191.0	-0.5
	5	44	12	*SMCHD1*	Hetero	5	227.3	295.8	-68.5
	6	50	13	*SMCHD1*	Hetero	3	120.0	243.4	-123.4
	7	64	13	*SMCHD1*	Hetero	3	93.8	243.4	-149.6
	8	14	10	*SMCHD1*	Hetero	6	857.1	400.5	456.6
	9	32	9	*SMCHD1*	Hetero	7	437.5	452.9	-15.4
	10	49	>14	*SMCHD1*	Hom	2	81.6	NA	NA
	11	35	13	*SMCHD1*	Hetero	5	285.7	243.4	42.4
	12	52	14	*SMCHD1*	Hetero	8	307.7	191.0	116.7
	13	55	12	*SMCHD1*	Hetero	6	218.2	295.8	-77.6
This study	-	53	13	*LRIF1*	Hom	8	301.9	243.4	58.5

Suppl. Table 4. List of rare variants identified in the proband by WES analysis. Variants at genes listed in Suppl. Table 1 were filtered with following thresholds for minor allele frequency in ESP6500, 1000G, and HGVD: 0 in AD (autosomal dominant) inheritance model and <0.01 in AR (autosomal recessive) inheritance model. Transcript references: *NEB*: ENST00000397345; *CACNA1A*: ENST00000360228; *HSPG2*: ENST00000374695. Protein references: NEB: ENSP00000380505; CACNA1A: ENSP00000353362.5; HSPG2: ENSP00000363827.3. Homo: homozygous; Het: heterozygous.

Gene	Inheritance	Predicted variants			Minor allele frequency		
		Transcript	Protein	Zygosity	ESP6500	1000G	HGVD
NEB	AR	c.2017T>C	p.Y673H	Homo	0	0	0
NEB	AR	c.25163G>A	p.R8388H	Het	0.000081	0.0014	0
NEB	AR	c.5411C>A	p.A1804E	Het	0	0.0018	0
CACNA1A	AD	c.1412A>G	p.K471R	Het	0	0	0
HSPG2	AD	c.3779G>C	p.G1260A	Het	0	0	0

Suppl. Table 5. Cell line information with detailed genotypes of the FSHD locus on both chromosomes 4 (indicated as 4q1 and 4q2 alleles), which comprises the D4Z4 repeat size, short sequence length polymorphism (SSLP) at proximal region of D4Z4 and the distal D4Z4 ending variant (A or B type). 4A161 haplotypes are considered permissive for *DUX4* expression in skeletal muscle.

					4q1 allele			4q2 allele		
Cell Line ID	**Cell Type**	**Clinical Status and Mutation in *SMCHD1***	**Primary (P) or Immortalized (I)**	**Sex**	**# of D4Z4 units**	**A/B haplotype**	**SSLP**	**# of D4Z4 units**	**A/B haplotype**	**SSLP**
2333	fibroblast	healthy control	I	M	20	A	161	24	A	161
2417	fibroblast	healthy control	I	F	22	A	161	26	B	163
2374	fibroblast	healthy control	I	F	27	B	168	35	B	163
2440	fibroblast	FSHD2 (*SMCHD1* c.1302_1306delTGATA)	I	F	19	A	161	101	B	168
2332	fibroblast	FSHD2 (*SMCHD1* c.3274_3276+1del)	I	M	14	A	161	65	A	161
2337	fibroblast	FSHD2 (*SMCHD1* g.(?_2656075)_(2802551_?)del*)	I	F	11	A	161	35	A	166
	fibroblast	FSHD2 (*SMCHD1* c.4118_4132del)	I	F	12	A	161	27	B	163
1926	myoblast	healthy control	P	F	9	B	163	32	A	161
2333	myoblast	healthy control	P	M	20	A	161	24	A	161
2417	myoblast	healthy control	P	F	22	A	161	26	B	163
2081	myoblast	healthy control	P	F	18	B	163	74	A	161
2401	myoblast	healthy control	I	M	13	A	161	21	B	168
073	myoblast	FSHD1	I	M	7	A	161	36	A	161
2440	myoblast	FSHD2 (*SMCHD1* c.1302_1306delTGATA)	I	F	19	A	161	101	B	168

* DNA change described using hg19/GRCh37 as reference

Suppl. Table 6: Primers' sequences used for qPCR. All primer pairs were used at Tm = 60°C.

Name	5'→3' sequence	remarks
GUSB F	CTCATTTGGAATTTTGCCGATT	used as housekeeping gene
GUSB R	CCGAGTGAAGATCCCCTTTTTA	
pLAMR4	TCCAGGAGATGTAACTCTAATCCA	for 4qA-S haplotype *DUX4* transcript
Dux4RT F2	CCCAGGTACCAGCAGACC	
Exon3pLAMqPCR-F2	CTTCCGTGAAATTCTGGCTGAATG	for *DUX4* transcript from 4qA-S and 4qA-L haplotype
DUX4polyAtail-R	TTTTTTTTTTTTTTTTTCTATAGGATCCACAGG	
hMYOG F	GCCAGACTATCCCCTTCCTC	
hMYOG R	GGGGATGCCCTCTCCTCTAA	
hMYH3 F	TGATCGTGAAAACCAGTCCATTCT	
hMYH3 R	TTGGCCAGGTCCCCAGTAGCT	
TRIM43 F	ACCCATCACTGGACTGGTGT	
TRIM43 R	CACATCCTCAAAGAGCCTGA	
ZSCAN4 F	TGGAAATCAAGTGGCAAAAA	
ZSCAN4 R	CTGCATGTGGACGTGGAC	
LEUTX F	AAGGAGGAGACTCCCTCAGC	
LEUTX R	AAAGAGAGTGGAGGCCCAAG	

Suppl. Table 7: All siRNAs were purchased from Thermo Fisher Scientific. siRNAs targeting exon 2 of *LRIF1* were custom designed using online BLOCK-iT™ RNAi Designer tool from Thermo Scientific Fisher.

Name in the study	Catalogue number	Sequence	remarks
NT	12935300	-	non-targeting siRNA, medium GC content
si1109	custom	CAAGCUGCUCCAGUGAAAUGGAUUU	targeting exon 2 of *LRIF1*
si1369	custom	GACAGAUGUUCUGCCAUCACAAAUU	targeting exon 2 of *LRIF1*
si1396	custom	CCAACAGAAUUCUGUUUCUCCUGAU	targeting exon 2 of *LRIF1*

Suppl. References

1. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
2. Lemmers, R. J. L. F. *et al.* Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum. Mol. Genet.* **24**, 659–669 (2015).
3. Goto, K., Nishino, I. & Hayashi, Y. K. Rapid and accurate diagnosis of facioscapulohumeral muscular dystrophy. *Neuromuscul. Disord.* **16**, 256–261 (2006).
4. Nelson, J. D., Denisenko, O. & Bomsztyk, K. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nat. Protoc.* **1**, 179–185 (2006).
5. Zeng, W. *et al.* Specific Loss of Histone H3 Lysine 9 Trimethylation and HP1γ/Cohesin Binding at D4Z4 Repeats Is Associated with Facioscapulohumeral Dystrophy (FSHD). *PLoS Genet.* **5**, e1000559 (2009).

CHAPTER 4

Locus-specific differences in chromatin recruitment of SMCHD1 and LRIF1

Darina Šikrová[1], Alessandra M. Testa[1,2], Iris Willemsen[1], Anita van den Heuvel[1], Stephen J. Tapscott[3], Lucia Daxinger[1], Judit Balog[1] and Silvère M. van der Maarel[1]

[1]Department of Human Genetics, Leiden University Medical Center, 2333ZC Leiden, The Netherlands

[2]present address: Department of Biomedical Sciences, University of Padua, 35100 Padua, Italy

[3]Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA

Manuscript in revision

Abstract

Germline mutations in *SMCHD1* or *LRIF1* are causative for Facioscapulohumeral muscular dystrophy (FSHD). FSHD results from a partial failure in epigenetic silencing of the D4Z4 repeat in the 4q subtelomere. This results in inappropriate *DUX4* expression from the repeat in skeletal muscle and leads to muscle wasting. The mechanism of SMCHD1- and LRIF1-mediated D4Z4 repression in myogenic cells is not fully elucidated. We show that SMCHD1 and LRIF1, despite their binding to D4Z4 in somatic cells, do not play a role in heterochromatin maintenance of this locus as defined by H3K9me3 and DNA methylation. Furthermore, we show that SMCHD1 recruitment to D4Z4 is LRIF1-independent while LRIF1 requires SMCHD1 for its D4Z4 association in somatic cells. In addition, we present evidence that SMCHD1 and LRIF1 form an auxiliary layer of *DUX4* repression on top of the known D4Z4 repressive mechanisms, even at already epigenetically compromised D4Z4 repeats. Lastly, we uncover that SMCHD1 together with the long isoform of LRIF1 negatively regulates expression of *LRIF1* by binding to its promoter region. The interdependency of SMCHD1 and LRIF1 binding thus seems locus-specific and shows different sensitivity to either early developmentally or somatically perturbed SMCHD1 function.

Introduction

Facioscapulohumeral muscular dystrophy (OMIM #158900 & #158901) is a heterogeneous disorder caused by misexpression of the transcription factor DUX4 in skeletal muscle [1,2]. One of the key physiological roles of DUX4 is its involvement in zygotic genome activation at the human 4-cell cleavage stage [3,4]. The short burst of *DUX4* expression during cleavage stage is followed by the activation of specific classes of retroelements and a cleavage stage-specific gene set. Indeed, this DUX4-sensitive transcriptional signature is also present in skeletal muscle biopsies or muscle cell cultures derived from individuals with FSHD or upon ectopic *DUX4* expression in control myoblasts [5–7]. This and other evidence suggests that DUX4 is a pioneer transcription factor able to overwrite the existing chromatin environment in differentiated cell types and to activate its native transcriptional program [8–12]. *DUX4* is encoded by a multicopy retrogene organized into the D4Z4 macrosatellite repeat located in the 4q and 10q subtelomeres [13]. While the exact origin of *DUX4* expression at the cleavage stage has not been determined yet, typically only 4q D4Z4-derived *DUX4* transcripts are associated with FSHD [1]. Furthermore, two major 4q subtelomeric allelic variants exist (4qA and 4qB), with only 4qA alleles being permissive for *DUX4* expression in skeletal muscle tissue due to the existence of a polymorphic *DUX4* polyadenylation signal [14].

DUX4 expression is restricted to the 4-cell cleavage stage after which it is quickly attenuated[4] and the *DUX4* locus remains transcriptionally silent in most somatic tissues [15,16]. In general, macrosatellite repeats in the genome, like D4Z4, display a heterochromatic structure in soma marked by high levels of DNA methylation and repressive histone modifications such as H3K9me3 [17]. A partial failure in the establishment and/or maintenance of this epigenetic landscape at D4Z4 results in variegated *DUX4* expression in FSHD myogenic cultures [1]. Successful D4Z4 repeat silencing is largely dependent on the repeat copy number [18]. In the non-affected population, the D4Z4 repeat is polymorphic in size and consists of 8-100 repeat units. In FSHD individuals, two distinct but partially overlapping genetic mechanisms lead to a failure in epigenetic silencing of this locus, allowing for *DUX4* expression in skeletal muscle and disease manifestation. In the majority of FSHD cases (FSHD1; 95%), a contraction of the repeat to a size of 1-10 units on a 4qA allele occurs which is associated with partial D4Z4 chromatin relaxation in somatic cells [19,20]. In the remaining 5% of FSHD cases (FSHD2), the epigenetic deregulation of D4Z4 occurs in a D4Z4 repeat contraction-independent manner as it results from *in trans* mutations in chromatin factors that act on D4Z4 [21–23]. In the latter case, the epigenetic landscape of both 4q and 10q D4Z4 repeats is affected, whereas in FSHD1 cases, only the contracted 4qA-D4Z4 repeat is epigenetically compromised [24,25]. Although the FSHD2 disease mechanism is considered D4Z4 repeat contraction-independent, it is not repeat size-independent as mutations in D4Z4 chromatin modifiers only result in disease presentation when combined with repeat sizes <20 D4Z4 units [26].

Mutations in three genes have been linked to FSHD2 so far, namely *SMCHD1* [21,27,28], *DNMT3B* [22] and *LRIF1* [23]. The most frequently mutated gene in FSHD2 cases is *SMCHD1*, accounting for >85% of FSHD2 individuals [21]. The SMCHD1 protein has been shown to undergo homodi-

merization via its C-terminal hinge domain [29] and several studies reported its role in the architectural organization of chromatin, predominantly at the inactive X chromosome [30–34]. However, the mechanism by which SMCHD1 imposes silencing on D4Z4 has not been fully answered yet. The somatic D4Z4 chromatin profile in FSHD2 cases with heterozygous *SMCHD1* mutations is characterized, apart from DNA hypomethylation, by increased H3K4me2 levels and decreased H3K9me3 levels [23], similar to contracted D4Z4 repeats in FSHD1. In addition, increased levels of H3K27me3 are specifically found in FSHD2 [35]. While mutations have been identified over the entire *SMCHD1* locus in FSHD2, heterozygous missense mutations in the ATPase domain of *SMCHD1* can also cause the rare developmental syndrome termed Bosma Arhinia and Microphtalmia Syndrome (BAMS; MIM603457) [36,37]. *SMCHD1* mutations in BAMS patients also result in D4Z4 hypomethylation and *DUX4* transcripts have been detected in some BAMS individuals [36–38]. There is an ongoing debate about the molecular basis for a clinical phenotype difference arising from *SMCHD1* mutations. Some studies suggest a gain-of-function model for BAMS mutations with FSHD2 mutations rather causing a loss-of-function [39]. This can, however, not explain the observation of two identical mutations in unrelated BAMS and FSHD2 patients [26,36,40,41].

The second gene identified as FSHD2 disease gene in a D4Z4 contraction-independent manner is *DNMT3B*. Heterozygous mutations in *DNMT3B* have been linked to FSHD2, while biallelic mutations in *DNMT3B* have been shown to cause the Immunodeficiency, Centromeric instability, Facial anomalies type I (ICF1) syndrome (OMIM #242860) [42,43]. In both disease situations, 4q and 10q D4Z4 repeats are hypomethylated [22,44] and *DUX4* expression has been also observed in ICF1 individuals who have at least one 4qA allele, which puts them at risk for FSHD [22]. As with BAMS, a full explanation as to why mutations in a single gene can cause such disparate disease phenotypes is missing, however, the simplest explanation could be a lower *DNMT3B* dosage in biallelic mutation carriers compared to heterozygous carriers thus resulting in worsen phenotype.

More recently, we have identified an individual presenting with symptoms consistent with FSHD [23] caused by a homozygous frameshift mutation in the *LRIF1* gene combined with 11 unit-long repeat on a 4qA chromosome. This homozygous frameshift mutation leads to the specific loss of the long LRIF1 isoform, while the short isoform persists. The D4Z4 chromatin profile of the proband resembles that of FSHD2 individuals with heterozygous *SMCHD1* mutations including increased H3K27me3 levels consistent with the presence of DUX4 in myogenic cell cultures [23].

The initiation of the D4Z4 epigenetic abnormalities in FSHD1 and 2 is not well known. However, in case of ICF1 and FSHD2 individuals with germline *DNMT3B* mutations, it is most likely during early embryonic developmental stages when DNMT3B is under normal circumstances responsible for establishing the cells' methylation profiles. The time window as well as the particular molecular action of SMCHD1 and LRIF1 that enforces a repressive D4Z4 chromatin structure in somatic cells is less clear. On one hand, ectopic expression of

SMCHD1 in FSHD1 and 2 myoblasts [35] as well as its mutation correction in FSHD2 myoblasts [45] was shown to result in *DUX4* downregulation suggesting that SMCHD1 does have a role in *DUX4* repression also in somatic cells, although the D4Z4 chromatin state after modulating SMCHD1 levels was not thoroughly examined in these studies. On the other hand, it was recently shown that knocking out SMCHD1 in HCT116 colon carcinoma cells leads to *DUX4* de-repression and that this *DUX4* transcriptional response cannot be attributed to changes in DNA methylation or H3K9me3 levels at D4Z4 [38]. This suggests that SMCHD1 is not required for DNA methylation or H3K9me3 maintenance in somatic cells. In addition, transient knock-down of the long LRIF1 isoform results in *DUX4* transcriptional de-repression in control as well as in FSHD1 and FSHD2 myoblasts [23] suggesting that it too has a *DUX4* expression modifying role in somatic cells albeit with unknown effect on the D4Z4 chromatin. Therefore, it is imperative to examine the role of SMCHD1 and LRIF1 in *DUX4* repression in somatic cells as well as studying the limiting chromatin requirements for their D4Z4 recruitment. Here, we examined SMCHD1 and LRIF1-mediated *DUX4* repression in different somatic cell model systems with distinct D4Z4 chromatin environments and demonstrate that they provide an auxiliary layer of chromatin repression on top of DNA methylation and H3K9me3. We also uncover an SMCHD1 and LRIF1-mediated transcriptional regulation of the *LRIF1* locus itself in somatic cells and show that this regulation differs from the action that SMCHD1 and LRIF1 impose on D4Z4 suggesting different sensitivity and mode of repression imposed by these two proteins at different genomic loci.

Results

The somatic loss of LRIF1 or SMCHD1 in control myoblasts leads to mild *DUX4* de-repression

We have previously shown that SMCHD1 and LRIF1L maintain repression of the D4Z4 repeat as short-term siRNA-mediated knock-down of LRIF1L or shRNA-mediated knock-down of SMCHD1 in control muscle cells having a D4Z4 repeat of <20 units on a 4qA allele results in transcriptional de-repression of *DUX4* [21,23,35]. To further study the mechanism of repression imposed by SMCHD1 and LRIF1 at D4Z4 in somatic cells, we employed CRISPR/Cas9 genome editing to generate somatic knock-out conditions for SMCHD1 ($SMCHD1^{KO}$), the long isoform of LRIF1 ($LRIF1L^{KO}$) or both isoforms of LRIF1 (long + short isoform, hereafter referred to as $LRIF1L+S^{KO}$) (Figure 1A) in the control immortalized myoblast cell line (1926iMB), which has a 32 unit-long 4qA FSHD permissive allele (Figure 1B). All three somatic knock-out conditions caused transcriptional de-repression of *DUX4*, albeit to very low levels which were insufficient to elicit a significant transcriptional response of DUX4 target genes (Figure 1C). We could only observe a mild increase in *MBD3L2* mRNA levels, but the expression levels of the other two DUX4 target genes tested (*KHDC1L* and *TRIM43*) remained at the level of WT clones. Interestingly, knocking out simultaneously the long and short LRIF1 isoform did not lead to additional *DUX4* de-repression compared to specific long isoform knock-out. Expression of *DUX4* has been shown to be positively influenced by myogenic differentiation [35].

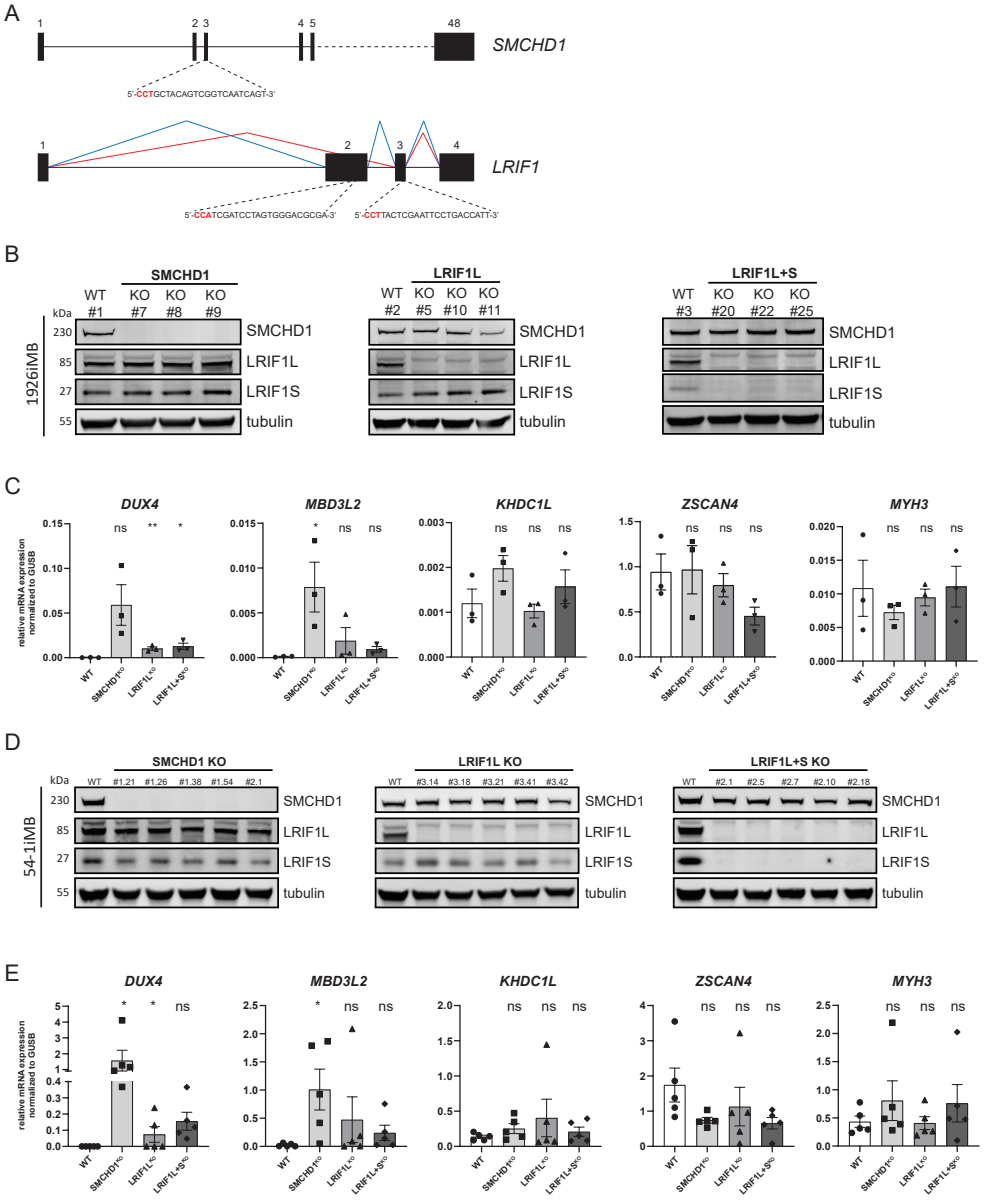


Figure 1. Knock-out of *SMCHD1* or *LRIF1* in control immortalized myoblasts have only a mild effect on *DUX4* de-repression. A) Gene structure of human *SMCHD1* (top) and *LRIF1* (bottom) and the position of the sgRNAs used for creating respective KOs (PAM sequence labelled in red). Two different LRIF1 isoforms are produced by differential splicing of exon 2 as denoted by different splicing patterns (blue = long isoform, red = short isoform). **B)** Confirmation of successful SMCHD1 and LRIF1 knock-outs in 1926iMB by western blot. Tubulin was used as a loading control. **C)** RT-qPCR of *DUX4*, three of its target genes (*MBD3L2*, *KHDC1L* and *ZSCAN4*) and a myogenic marker *MYH3* in differentiated 1926iMB WT and knock-out clones. Bars represent mean ± SEM. Each dot represents one clone. Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (**p<0.01,*p<0.05, ns - not significant). **D)** Western blot confirmation of successful SMCHD1 and different

LRIF1 KOs in 54-1iMB. Tubulin was used as loading control. **E)** RT-qPCR of *DUX4,* three of its target genes (*MBD3L2*, *KHDC1L* and *ZSCAN4*) and myogenic marker (*MYH3*) in differentiated 54-1iMB WT and different KO clones. Bars and whiskers represent mean ± SEM. Each dot represents one clone. Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (*p<0.05, ns - not significant).

MYH3 mRNA levels as well as fusion index, both markers of myogenic differentiation, did not reveal major differences between WT and KO clones (Figure 1C, Suppl. Figure 1A, B). We only detected a mild significant increase in fusion index in the case of SMCHD1^{KO}. This rules out the possibility that knock-out of SMCHD1 or LRIF1 profoundly impairs or accelerates myogenesis, both situations which would confound a direct effect on *DUX4* expression.
Next, we tested the hypothesis that the mild *DUX4* de-repression in these knock-out cell lines is caused by the relatively long 4qA permissive repeat (32U). Sizing of D4Z4 permissive alleles in an FSHD2 cohort with germline heterozygous *SMCHD1* mutations revealed that the majority of disease allele sizes are between 11-20U [26]. Moreover, individuals with germline heterozygous *SMCHD1* mutations and longer D4Z4 permissive alleles tend to be asymptomatic [26]. Therefore, we generated an additional set of knock out clones from the control immortalized myoblast line 54-1iMB with a permissive 4qA allele of 13 U (Figure 1D). However, despite the FSHD2-sized D4Z4 repeat, we did not observe more pronounced *DUX4* de-repression with concomitant DUX4 target genes upregulation in any of the knock-out conditions (Figure 1E), suggesting that repeat length cannot explain the mild *DUX4* de-repression in control somatic knock-out cells.

The somatic loss of LRIF1 or SMCHD1 in control myoblasts does not result in D4Z4 chromatin changes typical for FSHD2

The lack of a robust transcriptional *DUX4* response upon *SMCHD1* or *LRIF1* knock-out prompted us to investigate the D4Z4 chromatin features that are characteristic of FSHD2 D4Z4 alleles. First, we examined possible DNA methylation changes as germline defects in SMCHD1 or LRIF1 in FSHD2 lead to pronounced pan-D4Z4 hypomethylation especially of 19 CpGs within the previously reported DR1 region [23,46]. We analyzed three independent clones from each 1926iMB and 54-1iMB knock-out condition. Bisulfite PCR of the DR1 region followed by subcloning and sequencing did not, however, reveal noticeable changes in either overall DNA methylation levels (Figure 2A and 2B) or at individual CpGs in the DR1 amplicon in any of the knock-out conditions compared to WT (Suppl. Figure 2A and 2B). This finding corroborates and extends on a previous study showing that SMCHD1 knock-out in HEK293T cells or HCT116 cells does not result in D4Z4 hypomethylation [38].

Next, we performed chromatin immunoprecipitation of three histone marks (H3K9me3, H3K4me2 and H3K27me3), which are known to be deregulated at D4Z4 in FSHD2 due to germline mutations either in *SMCHD1* or in *LRIF1* [23,35,47]. As was the case for DNA methylation, *SMCHD1* and *LRIF1* somatic knock-outs in 1926iMB did not show altered levels of histone H3 itself (Figure 2C) or any of the examined H3-associated histone modifications as compared to WT clones (Figure 2D). Similar observations, i.e. largely unchanged H3K9me3 levels at

D4Z4, have been made upon *SMCHD1* knock-out in HCT116 cells [38]. Taken together, these findings may thus explain the observed limited transcriptional response of the *DUX4* locus resulting from decreased levels of either SMCHD1 or LRIF1 in control myoblasts, since the examined repressive mechanisms in the form of DNA methylation and repressive histone modifications remained intact.

LRIF1 recruitment to D4Z4 in somatic cells is partially SMCHD1-dependent while SMCHD1 recruitment to D4Z4 is independent of LRIF1

LRIF1 and SMCHD1 have been found to be associated with H3K9me3 in independent proteomic studies aimed at identifying factors associated with specific histone modifications [48–50]. At D4Z4 it was shown that reducing H3K9me3 levels in control myoblasts results in reduced SMCHD1 occupancy, placing SMCHD1 downstream of H3K9me3 [51]. In mouse embryonic stem cells, a predominant mechanism for Smchd1 recruitment to H3K9me3-marked chromatin depends on Lrif1 and this study proposed that SMCHD1 recruitment to D4Z4 could be also mediated by LRIF1 as LRIF1 recognizes HP1-bound H3K9me3 enriched heterochromatin [29]. To test if this proposed LRIF1-dependent SMCHD1 chromatin recruitment to D4Z4 mechanism holds true, we performed SMCHD1 and LRIF1 chromatin immunoprecipitation in our somatic SMCHD1 and LRIF1 knock-out 1926iMB clones, where, as we show, the H3K9me3 levels at D4Z4 are preserved (Figure 2C). This allowed us to interrogate both the inter-dependency of these two factors in their D4Z4 recruitment, and the H3K9me3-dependency of this mechanism. In agreement with a previous study [38], SMCHD1 is mostly enriched at the DR1 region of D4Z4 with a gradual decrease in 3' direction in the WT situation and this enrichment pattern is also observed for LRIF1 (Figure 3A). Interestingly, after having examined three different regions along the D4Z4 unit, we did not observe reduced SMCHD1 binding to D4Z4 in either LRIF1 knock-out condition (Figure 3A). Therefore, the presence of SMCHD1 at D4Z4 in 1926iMB cells with unperturbed D4Z4 heterochromatin is independent of LRIF1. On the other hand, we detected decreased LRIF1 enrichment at D4Z4 in 1926iMB SMCHD1^{KO} cells to the same degree as in LRIF1L+S^{KO} cells, which served as a baseline for the ChIP antibody background. This suggests that the presence of LRIF1 at D4Z4 is at least in part SMCHD1-dependent. Since H3K9me3 levels at D4Z4 were not reduced in SMCHD1^{KO} cells (Figure 2C), this implies that H3K9me3 alone is not sufficient for LRIF1 to be present on D4Z4 in 1926iMB cells.

To further examine the SMCHD1-dependency of LRIF1 at D4Z4, we studied the reverse situation and tested whether increased SMCHD1 binding to D4Z4 would also result in increased LRIF1 binding. To address this, we used a previously described FSHD2 myoblast cell line, which carries a heterozygous germline mutation (c.4347-236A>G) in intron 34 of the *SMCHD1* locus [45]. This mutation creates a cryptic splice site which leads to exonisation of 53 bp of intronic sequence thereby disturbing the open reading frame of SMCHD1 and causing its haploinsufficiency (Figure 3B). We recently showed that we can correct this genetic lesion in myoblasts derived from this individual by removing the pseudo-exon with a dual Cas9 strategy, which restores SMCHD1 splicing and protein levels and results in *DUX4* suppression [45].

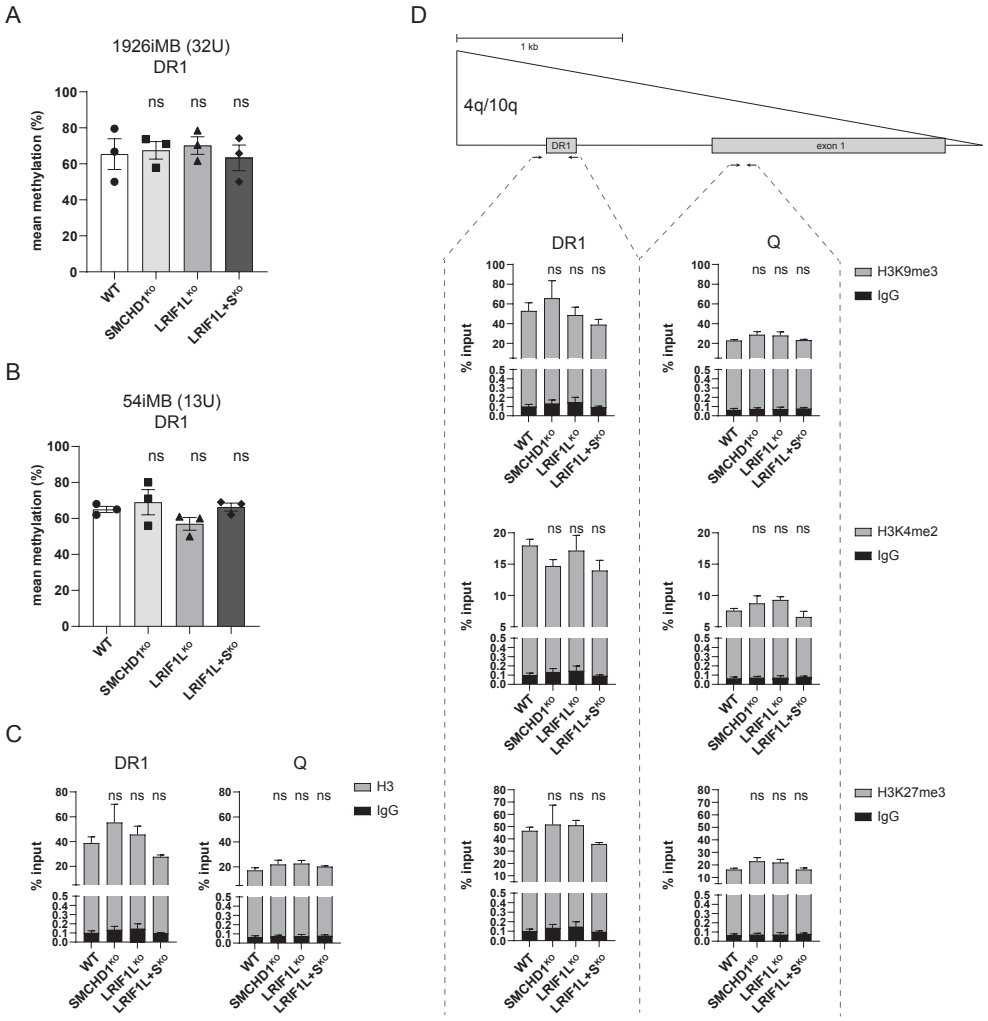


Figure 2. Somatic SMCHD1 and LRIF1 knock-outs do not recapitulate perturbations of heterochromatin marks known in FSHD2. A) Mean DNA methylation level of the D4Z4 DR1 region in different 1926iMB knock-out clones as determined by bisulfite Sanger sequencing. Bars and whiskers represent mean ± SEM of three independent clones**,** respectively. **B)** DNA methylation of the DR1 region in different knock-out 54-1iMB clones as determined by bisulfite Sanger sequencing. Bars and whiskers represent mean ± SEM of three independent clones, respectively. **C)** ChIP-qPCR of histone H3 at the DR1 and Q region in different 1926iMB knock-out conditions. Isotype specific IgG served as a background control. **D)** ChIP-qPCR of selected H3 modifications at the D4Z4 DR1 and Q region in different 1926iMB knock-out conditions. Isotype specific IgG served as a background control. Schematic of a D4Z4 unit with position of the DR1 and Q region within D4Z4 examined by ChIP-qPCR indicated. Bars represent mean ± SEM (ns = 3 per genotype). Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (ns – not significant).

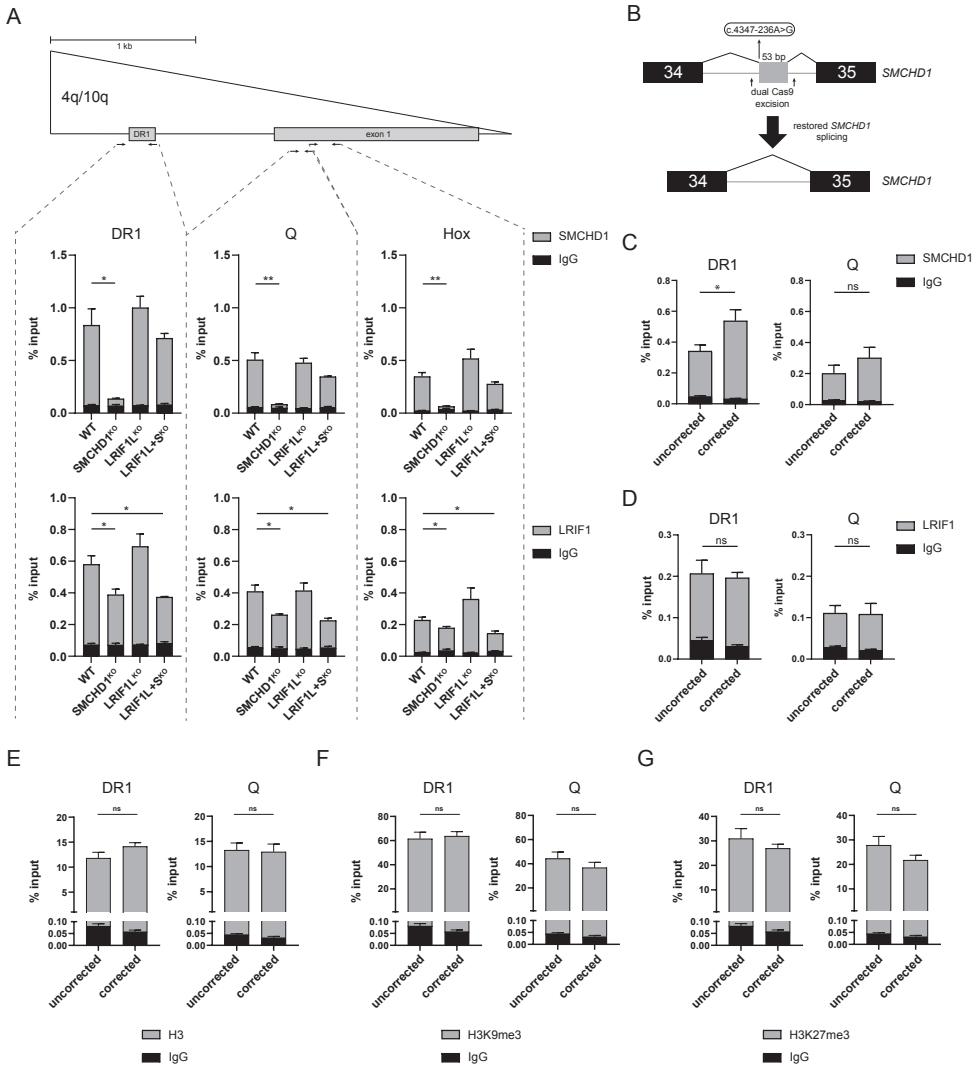


Figure 3. SMCHD1 binding to D4Z4 in somatic cells is independent of LRIF1. A) SMCHD1 and LRIF1 ChIP-qPCR in different 1926iMB knock-out conditions. Schematic of one D4Z4 unit and the position of three regions within D4Z4 examined by ChIP-qPCR is indicated. Bars and whiskers represent mean ± SEM of three independent clones. Isotype specific IgG was used for background control. Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (**p<0.01,*p<0.05). Only significant p-values are shown. **B)** Schematic representation of splicing of the mutant *SMCHD1* allele carrying the intronic SNP variant indicated in the box. **C)** SMCHD1 ChIP-qPCR of two D4Z4 regions (DR1 and Q) from four *SMCHD1* intron unedited and four *SMCHD1* intron edited clones that restores WT *SMCHD1* splicing. Bars and whiskers represent mean ± SEM. **D)** LRIF1 ChIP-qPCR of two D4Z4 regions (DR1 and Q) from four *SMCHD1* intron unedited and four *SMCHD1* intron edited clones. Bars and whiskers represent mean ± SEM. Statistical significance was calculated with unpaired t-test (**p<0.01,*p<0.05, ns – not significant). **E)** H3 ChIP-qPCR of the D4Z4 DR1 and Q region from four *SMCHD1* unedited and four *SMCHD1* intron edited clones. Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. **F)** H3K9me3 ChIP-qPCR of the D4Z4 DR1 and Q region from four *SMCHD1*

unedited and four *SMCHD1* intron edited clones. Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. **G)** H3K27me3 ChIP-qPCR of the D4Z4 DR1 region from four *SMCHD1* unedited and four *SMCHD1* intron edited clones. Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. Statistical significance was calculated with an unpaired t-test (ns - not significant).

We performed chromatin immunoprecipitation studies of SMCHD1 and LRIF1 in four *SMCHD1* uncorrected and four corrected clones, which were previously characterized [45]. We found that SMCHD1 enrichment at D4Z4 was indeed increased in the corrected myoblast clones with the strongest rescue being at the DR1 region thus explaining the previously observed *DUX4* repression in the corrected cells (Figure 3C). Next, we tested whether this increased SMCHD1 binding was associated with increased LRIF1 binding to D4Z4. However, LRIF1 enrichment at D4Z4 at two examined sites (DR1 and Q) did not change significantly in *SMCHD1* corrected clones (Figure 3D).

To investigate why LRIF1 enrichment at D4Z4 was not restored together with increased SMCHD1 levels, we examined the chromatin state of D4Z4 in *SMCHD1* corrected clones. We previously showed that restoring SMCHD1 levels in these corrected clones does not lead to re-methylation of D4Z4 [45]. Further examination of the H3K9me3 and H3K27me3 histone modifications showed that correction of the *SMCHD1* mutation did not result in the re-establishment of a D4Z4 histone modification pattern observed in healthy individuals (Figure 3E, F and 3G). These results suggest that modulating SMCHD1 levels in somatic cells does not affect D4Z4 chromatin state defined by DNA methylation, H3K9 trimethylation and H3K27 trimethylation. In addition, this might indicate that LRIF1 binding to D4Z4 does not solely depend on SMCHD1 and that other chromatin factors or marks may play a role in somatic cells.

SMCHD1 and LRIF1 provide auxiliary repression of *DUX4* at epigenetically compromised D4Z4 repeats

Since modulating SMCHD1 levels in FSHD2 myoblasts affects *DUX4* levels, we further explored this in two other unrelated conditions in which the D4Z4 repeat is hypomethylated due to either dysfunctional DNA methylation maintenance or its establishment. This allowed us to assess if SMCHD1 and LRIF1 can bind to hypomethylated D4Z4 repeats and enforce *DUX4* repression in a situation where the epigenetic disturbance of D4Z4 is not due to germline mutations in *SMCHD1* or *LRIF1*. First, we focused on a situation in which hypomethylated D4Z4 arose due to inactivation of the DNA methylation maintenance machinery in somatic cells. For this, we used the colorectal cancer line (HCT116) and its *DNMT1* and *DNMT3B* double knock-out (DKO) derivative. D4Z4 hypomethylation in HCT116 DKO cells is accompanied by a reduction in H3K9me3 and gain in H3K4me2 ultimately resulting in *DUX4* de-repression [16]. Somatic loss of DNA methylation in HCT116 DKO cells leads to 5' to 3' redistribution of SMCHD1 along the D4Z4 unit (Suppl. Figure 3), in agreement with previous findings [38]. Similarly, the LRIF1 enrichment pattern followed the one of SMCHD1 (Suppl. Figure 3).

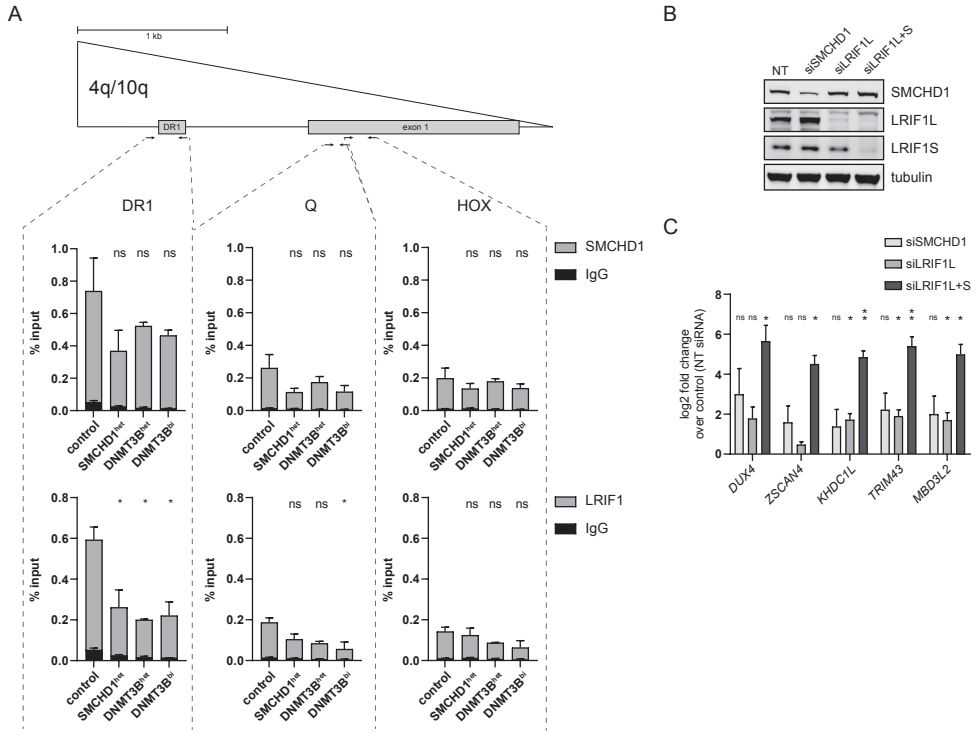


Figure 4. SMCHD1 and LRIF1 have residual repressive action at hypomethylated D4Z4. A) SMCHD1 and LRIF1 ChIP-qPCR in primary control (n = 3, lines: 2524, 2397, 2333) fibroblasts and fibroblasts carrying either a heterozygous *SMCHD1* mutation (n = 3, lines: 2440, 2337, 2332), a heterozygous *DNMT3B* mutation (n = 2, lines: v294, b974) or biallelic *DNMT3B* mutations (n = 2, lines: Rf699.3, Rf286.3). Schematic of one D4Z4 unit in which the position of three regions within D4Z4 examined by ChIP-qPCR is indicated (DR1, Q, HOX). Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. Statistical significance between WT and mutant groups was calculated by one-way ANOVA with Dunnett's post hoc test (*p<0.05, ns - not significant). **B)** Western blot confirmation of successful siRNA-mediated knock-down of SMCHD1, LRIF1L or LRIF1L+S in primary ICF1 myoblasts. Tubulin was used as a loading control. **C)** RT-qPCR of *DUX4* and four of its target genes (ZSCAN4, *KHDC1L, TRIM43* and *MBD3L2*) after siRNA-mediated KD of SMCHD1, LRIF1L or LRIF1L+S in ICF1 myoblasts. Expression levels detected in KD cells were normalized to cells transfected with non-targeting (NT) siRNA and further log2 transformed. *GUSB* was used as a housekeeping gene for intra-sample normalization. Bars and whiskers represent mean ± SEM of three independent experiments. Statistical significance was calculated by one sample t-test (**p<0.01, *p<0.05, ns - not significant).

Next, we used a cellular model system in which the D4Z4 repeat is hypomethylated in somatic cells derived from individuals with germline mutations in *DNMT3B* thus representing a case of failed DNA methylation establishment at D4Z4. For this we studied primary fibroblasts from individuals having either heterozygous (*DNMT3B*^{het}) or biallelic *DNMT3B* mutations (*DNMT3B*^{bi}). All *DNMT3B*^{bi} fibroblasts are derived from individuals diagnosed with ICF1 syndrome. These individuals present with more pronounced D4Z4 hypomethylation as compared to their heterozygous unaffected relatives [22]. First, we characterized the D4Z4 chromatin in these samples to examine if the DNA hypomethylation is accompanied by histone modification changes typical for FSHD2. We performed ChIP-qPCR for H3K4me2, H3K9me3 and H3K27me3 as well as histone H3. Already the H3 level itself was slightly reduced compared to

primary fibroblasts from control individuals suggesting a possible loosening or remodelling of nucleosomes at D4Z4 in fibroblasts from individuals with mono- or biallelic *DNMT3B* mutations (Suppl. Figure 4A). In addition, H3K9me3 levels were decreased while those of H3K4me2 and H3K27me3 were increased similar to what has been observed in FSHD2 fibroblasts carrying either *SMCHD1* or *LRIF1* mutations (Suppl. Figure 4B). Interestingly, DNMT3B^{bi} fibroblasts displayed a tendency towards more pronounced changes than DNMT3B^{het} fibroblasts. Next, we performed ChIP-qPCRs for SMCHD1 and LRIF1. We included also primary FSHD2 fibroblasts, which have heterozygous *SMCHD1* mutations (*SMCHD1*^{het}), and where SMCHD1 and LRIF1 occupancy at D4Z4 is expected to be reduced based on the previous studies [21,23,35]. Interestingly, whereas the SMCHD1 and LRIF1 D4Z4 enrichment profile in HCT116 DKO cells showed evidence for a redistribution (Suppl. Figure 3), in primary DNMT3B mutant fibroblasts their occupancy was reduced at all three tested D4Z4 regions with the strongest impact observed at the D4Z4 DR1 site, while at the Q and Hox region the enrichment difference did not reach statistical significance (Figure 4A). Altogether, this shows that both SMCHD1 and LRIF1 recruitment to D4Z4 is sensitive to chromatin changes associated with DNA hypomethylation either at the somatic stage as represented by the results from our studies in HCT116 DKO cells or by a failure in DNA methylation establishment during early development as represented by the results from our studies in fibroblasts carrying *DNMT3B* mutations.

Additionally, we tested if there is a synergistic effect of heterochromatin marks and SMCHD1 and LRIF1 on *DUX4* repression. We used a mix of siRNAs to deplete either SMCHD1, LRIF1L or LRIF1L+S in ICF1 myoblasts (Rf285.3) derived from an individual who carries an 11 unit-long permissive 4qA allele (Figure 4B). Since these myoblasts have biallelic mutations in the *DNMT3B* gene, the D4Z4 heterochromatin is already compromised as shown above. All three knock-down scenarios lead to variable *DUX4* transcriptional upregulation and further activation of four *DUX4* target genes (*ZSCAN4*, *KHDC1L*, *TRIM43*, *MBD3L2*) as compared to cells which were treated with a non-targeting siRNA mix (Figure 4C). This suggests that despite the decreased SMCHD1 and LRIF1 enrichment at D4Z4 in ICF1, these proteins still provide residual repression and their depletion leads to further transcriptional de-repression.

SMCHD1 and the long isoform of LRIF1 negatively regulate *LRIF1* expression

Lastly, to evaluate a genome-wide repressive function of SMCHD1 and LRIF1, we performed poly-A RNA-seq in WT and respective KO clones derived from the 1926iMB line. For this we used undifferentiated myoblasts to avoid transcriptional differences which could arise due to different myogenic differentiation of individual clones as well as to avoid any possible DUX4-driven signature as *DUX4* is expressed, albeit in a low levels, in myotubes of the knock-out clones (Figure 1C). Differential expression analysis did not reveal major transcriptional changes in any of the knock-out conditions when compared to WT clones, with SMCHD1^{KO} having the strongest impact out of the three knock-out conditions (Figure 5A, Suppl. Figure 5A and 5B, Suppl. Table 1, 2 and 3). These results extend on the previously reported lack of transcriptional deregulation after siRNA-mediated knock-down of SMCHD1 or LRIF1L+S in female RPE1-hTERT cells [52] and suggest that neither short term nor permanent depletion of SMCHD1 or LRIF1 in somatic cells has a major impact on the transcriptome.

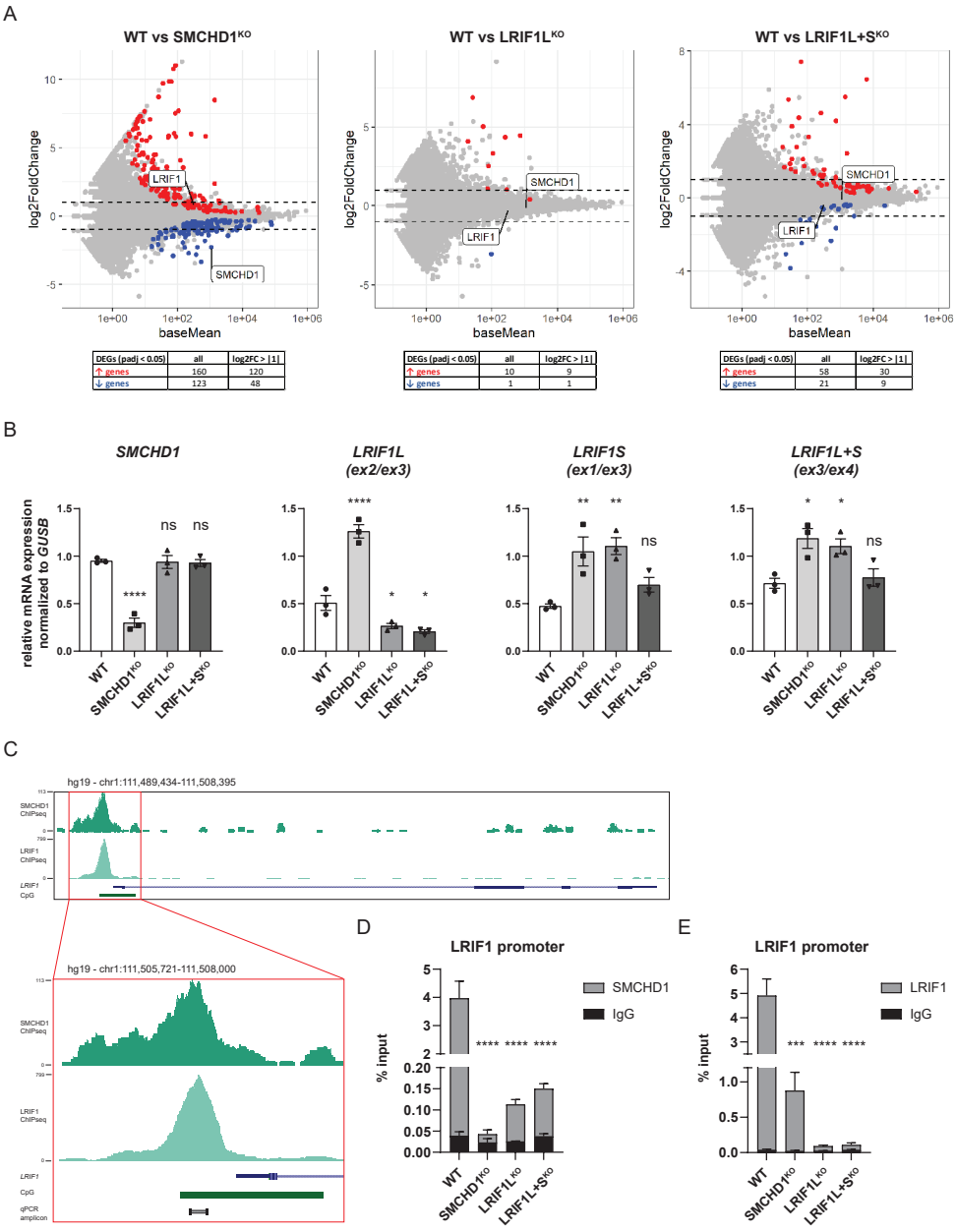


Figure 5. SMCHD1 and LRIF1 long isoform negatively regulate *LRIF1* expression. A) MA plots of RNA-seq experiments performed on three independent WT, SMCHD1^{KO}, LRIF1L^{KO} or LRIF1L+S^{KO} clones derived from the 1926iMB cell line. Differentially upregulated genes are highlighted in red and differentially downregulated genes are in blue (p-adjusted < 0.05) with summary of differentially expressed genes provided in a table format below each MA plot. Dashed lines mark $\log_2$ fold change of |1|. *SMCHD1* and *LRIF1* transcripts are indicated. **B)** RT-qPCR of *SMCHD1* and different exon junctions of *LRIF1* to differentiate between expression of different *LRIF1* isoforms (ex2/ex3 = long isoform, ex1/ex3 = short isoform, ex3/4 = both isoforms). Bars and whiskers represent mean ±

SEM. Each dot represents one clone. Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (****<0.0001, ***<0.001, **p<0.01, *p<0.05, ns – not significant). **C)** SMCHD1 and LRIF1 ChIP-seq from RPE1 cells showing SMCHD1 and LRIF1 enrichment over the LRIF1 promoter region. A zoom of the promoter region is presented to depict the amplicon used for ChIP-qPCR. **D)** SMCHD1 ChIP-qPCR of the *LRIF1* promoter in different 1926iMB WT and KO conditions. Bars and whiskers represent mean ± SEM of three independent clones. Isotype specific IgG was used as background control. **E)** LRIF1 ChIP-qPCR of the *LRIF1* promoter in different 1926iMB WT and KO conditions. Bars and whiskers represent mean ± SEM of three independent clones. Isotype specific IgG was used as background control. Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (****<0.0001, ***<0.001).

Interestingly, we noticed that the *LRIF1* gene was differentially upregulated in SMCHD1^{KO} clones and furthermore, was not differentially downregulated in either of the two LRIF1^{KO} situations as might be expected from the CRISPR/Cas9 induced indels leading to a premature stop codon often leading to non-sense mediated decay (NMD) of transcripts as is the case for *SMCHD1* transcripts in the SMCHD1^{KO} condition (Figure 5A). We validated this observation by RT-qPCR using exon junction primers specifically detecting *LRIF1* long (ex2/3), short (ex1/ex3) or all isoforms (ex3/4) (Figure 5B). Indeed, we detected elevated transcript levels of both *LRIF1* isoforms in SMCHD1^{KO} clones and even increased levels of the *LRIF1* short isoform in LRIF1L^{KO} clones. This prompted us to examine if LRIF1 and SMCHD1 directly regulate the *LRIF1* locus. Examining previously published SMCHD1 and LRIF1 ChIP-seq datasets from RPE1-hTERT cells [52] revealed enrichment of both SMCHD1 and LRIF1 immediately upstream of *LRIF1* exon 1, coinciding with the CpG island (Figure 5C). We confirmed this ChIP-seq peak with SMCHD1 and LRIF1 ChIP-qPCR also in 1926iMB cells suggesting that this transcriptional regulation might be conserved between different cell types (Figure 5D and 5E). The enrichment of both proteins is reduced in SMCHD1^{KO} clones (Figure 5D) and already in LRIF1L^{KO} clones which still express the short isoform (Figure 5E). Furthermore, the enrichment of SMCHD1 or LRIF1 is not further reduced in LRIF1L+S^{KO} cells suggesting that there is no additive effect in binding reduction after depleting cells also of the short LRIF1 isoform. This differs from the situation at D4Z4 where SMCHD1 binding is affected neither in LRIF1L^{KO} or LRIF1L+S^{KO} clones (Figure 3A). In addition, the overall enrichment of LRIF1 is not affected at D4Z4 in LRIF1L^{KO} cells in contrast to the situation at the *LRIF1* promoter suggesting different binding properties of LRIF1 isoforms to these two loci.

Promoters of expressed genes are known to be decorated by active histone marks such as H3K4me3 and H3K4me2, while promoters of silent genes are marked with repressive histone marks such as H3K9me3 and H3K27me3. As LRIF1 and SMCHD1 are known to be associated with H3K9me3 and we show that both proteins bind to the LRIF1 promoter in 1926iMB cells, we examined the histone marks at this locus. Interestingly, despite the *LRIF1* gene being expressed in these cells, its promoter is characterized by the active H3K4me2 mark and the repressive histone marks, H3K9me3 and H3K27me3, as opposed to the promoter of *GAPDH*, a constitutively expressed housekeeping gene, which is only enriched in the active H3K4me2 mark (Suppl. Figure 5C). As *LRIF1* expression is upregulated in SMCHD1 and LRIF1 knock-out 1926iMB cells, we wondered if we can find underlying changes in histone marks which would explain such transcriptional response, possibly increased levels in active marks and/or decreased levels of repressive marks. Surprisingly, the H3 level itself was reduced in

each knock-out condition as well as all the examined histone marks coupled to this histone (Suppl. Figure 5D). This indicates that a nucleosome displacement from the *LRIF1* promoter could explain the observed transcriptional upregulation.

Lastly, since somatic depletion of SMCHD1 in 1926iMB results in reduced LRIF1 binding at the *LRIF1* promoter and subsequent *LRIF1* upregulation, we examined if SMCHD1 and LRIF1 enrichment at the *LRIF1* promoter would also be decreased in FSHD2 primary fibroblasts with heterozygous *SMCHD1* mutations or in primary fibroblasts carrying either heterozygous or biallelic *DNMT3B* mutations similarly to what we observed at the D4Z4 repeat. The ChIP-qPCR of the *LRIF1* promoter in the same set of primary fibroblasts as in Figure 4A did not reveal differences in SMCHD1 or LRIF1 enrichment at this locus (Suppl. Figure 5E and 5F). This observation is consistent with unchanged *LRIF1* transcript levels in different examined cell types (primary fibroblasts, myoblasts or differentiated myotubes) derived from FSHD2 individuals with SMCHD1 haploinsufficiency compared to their control counterparts (Suppl. Figure 5G). These results suggest a different binding mechanism of SMCHD1 and LRIF1 to different H3K9me3-marked genomic regions as well as different sensitivity of these regions to either germline or somatic dosages of these genes as evidenced by the expression regulation of the *DUX4* gene organized in a repetitive macrosatellite structure and the single-copy *LRIF1* locus.

Discussion

To date, mutations in three genes, namely *SMCHD1*, *DNMT3B* and *LRIF1*, have been identified to cause FSHD type 2: a disease in which the chromatin structure of the D4Z4 repeat is compromised leading to inappropriate expression of *DUX4* in skeletal muscle. Therefore, understanding the role of these proteins in establishing or maintaining a repressive D4Z4 epigenetic landscape in somatic cells is not only important from a biological perspective, but also of clinical importance.

While the exact biological roles of SMCHD1 and LRIF1 are less defined, the predominant function of DNMT3B is to establish the DNA methylation pattern during early embryonic development [53]. While the expression levels of the catalytically active isoform DNMT3B1 sharply decline during pluripotent stem cell differentiation, cells continue to express its catalytically inactive isoforms [54,55], albeit at low levels. These catalytically inactive isoforms are thought to act as accessory proteins to catalytically active DNMT1, thus aiding the DNA methylation maintenance process in somatic cells [56]. Interestingly, two studies also reported a role for the catalytically active DNMT3B1 isoform in skeletal muscle cells [57,58]. However, whether catalytically active or inactive DNMT3B isoforms have a physiologically relevant function in D4Z4 repression after methylation patterns have been established in early embryogenesis, remains to be addressed. In contrast, we have previously demonstrated that SMCHD1 and LRIF1 have a *DUX4* expression modifying role in somatic cells having observed that altering their levels in FSHD1 and FSHD2 myoblasts influences the expression of *DUX4* by yet unknown mechanisms [21,23,35]. This provided the rational to only focus on the knock-out of SMCHD1 and LRIF1 in somatic cells.

Here, we aimed to further study the role of SMCHD1 and LRIF1 in D4Z4 repression. In order to do so, we evaluated their repressive activity at D4Z4 in different D4Z4 genetic and chromatin contexts. First, we created SMCHD1 and LRIF1 knock-out sets in two independent control immortalized myoblast lines with different D4Z4 repeat sizes and performed expression and chromatin studies of the D4Z4 repeat. Removing these factors from control cells derived from healthy individuals which have undergone uncompromised epigenetic establishment trajectories during development showed that once the D4Z4 epigenetic landscape is established, these factors do not play a role in its heterochromatin maintenance. Rather, they provide an auxiliary molecular seal on top of the existing chromatin structure thus increasing the robustness of a locus against leaky transcription. A complementary experiment supports this conclusion since in *SMCHD1* gene-corrected FSHD2 patient myoblasts, in which *SMCHD1* haploinsufficiency is rescued, *DUX4* downregulation is achieved in the absence of a reversal of the chromatin landscape as determined by DNA methylation, H3K9me3 and H3K4me2. These factors thus control *DUX4* expression by other yet unknown mechanism, possibly by promoting further chromatin condensation or higher-order chromatin conformation as was recently reported for SMCHD1 in the process of inactive X formation and *Hox* gene cluster regulation [30,33,34,59].

The association of Smchd1 with H3K9me3-enriched chromatin was previously shown to be dependent on Lrif1 in mouse embryonic stem cells [29] and it was speculated that this interaction may also facilitate the recruitment of SMCHD1 to D4Z4 chromatin. In line with this hypothesis, we have previously observed decreased SMCHD1 levels at D4Z4 in somatic cells derived from an FSHD2 individual in whom the long isoform of LRIF1 is absent to similar levels as observed in FSHD2 cases with an *SMCHD1* defect [23]. In contrast, here we show that knocking out specifically the long isoform of LRIF1 or both LRIF1 isoforms in control immortalized myoblasts does not affect SMCHD1 binding to D4Z4, which suggests that SMCHD1 recruitment, at least in somatic cells with properly established D4Z4 heterochromatin, is not dependent on LRIF1. On the other hand, we show that the loss of SMCHD1 in somatic cells leads to decreased LRIF1 enrichment at D4Z4 and similarly, LRIF1 enrichment at D4Z4 is decreased in FSHD2 cases with *SMCHD1* mutations. This is in line with the findings reported by a recent preprint, where association of Lrif1 with DAPI-dense heterochromatin was shown to be lost in the absence of Smchd1 in mouse differentiated cells [60]. However, LRIF1 recruitment to D4Z4 does not seems solely depend on SMCHD1 as rescuing SMCHD1 levels in FSHD2 cells and increasing its levels at D4Z4 in its de-repressed state does not lead to higher LRIF1 levels. This might suggest that LRIF1 recruitment to D4Z4 needs some other chromatin factor apart from SMCHD1 that was not restored upon *SMCHD1* gene correction, such as H3K9me3 or factor(s) dependent on this mark like HP1 proteins or alternatively, a newly gained modification or factor at D4Z4 in these FSHD2 cells even following *SMCHD1* correction impedes LRIF1 recruitment.

D4Z4 is decorated with H3K9me3 in somatic cells and this repressive histone mark is significantly decreased at this locus in FSHD2, ICF1 as well as in HCT116 DKO cells. Others have shown that the presence of this mark is crucial for SMCHD1 recruitment to D4Z4 in somatic cells [51] and thus could explain the decreased levels of SMCHD1 and LRIF1 at

D4Z4 in its hypomethylated state as DNA hypomethylation concomitantly results in lower H3K9me3 levels as evidenced by results from HCT116 DKO and samples with heterozygous or homozygous *DNMT3B* mutations.

The remaining H3K9me3 at hypomethylated D4Z4 could provide an explanation for residual SMCHD1 and LRIF1 binding to this locus and also explain the previously observed reduced binding of SMCHD1 in cells from an FSHD2 individual in whom the LRIF1 long isoform is absent or from *DNMT3B* mutation carriers, all conditions in which the H3K9me3 mark is reduced at D4Z4. This also suggest a more fine-tuning role for SMCHD1 and LRIF1 in *DUX4* repression in somatic cells as correctly established D4Z4 repeat displays a large degree of resistance to its transcriptional de-repression with majority of this repression block coming from the chromatin state itself (Figure 6).

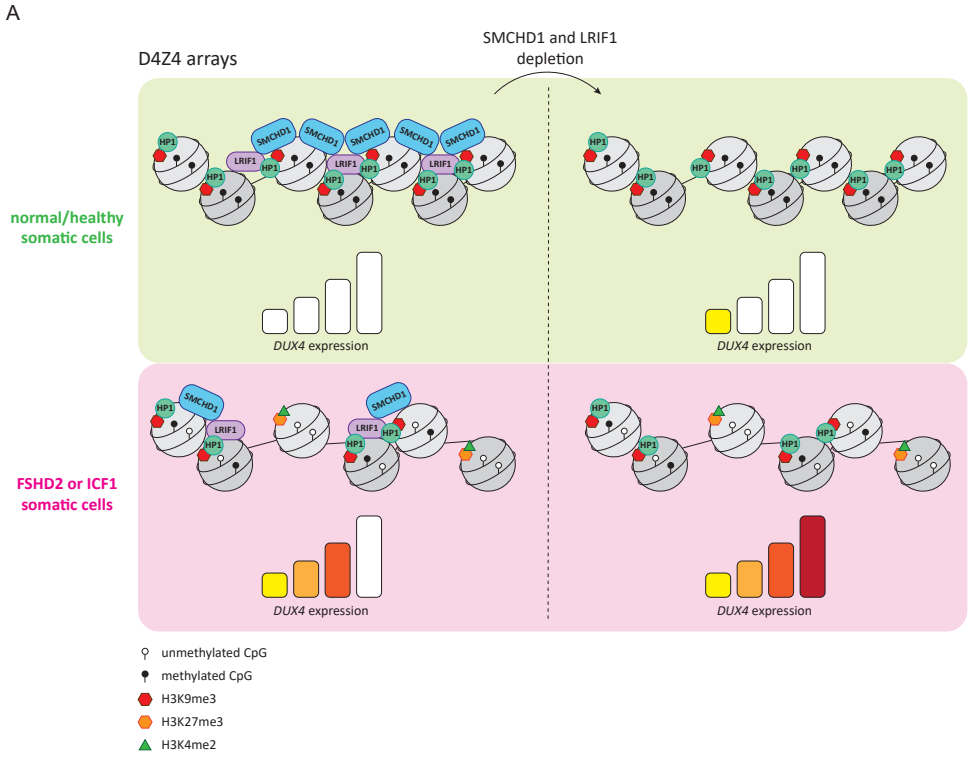


Figure 6. Model for SMCHD1 and LRIF1-mediated D4Z4 repression in somatic cells. In somatic cells from unaffected individuals, D4Z4 is marked by high levels of CpG methylation, H3K9me3, HP1 proteins as well as LRIF1 and SMCHD1, resulting in transcriptional silencing of *DUX4*. LRIF1 recruitment is partially dependent on SMCHD1 and possibly stabilized by HP1, while the mechanism of SMCHD1 recruitment to D4Z4 requires H3K9me3. In ICF1 or FSHD2 somatic cells, both SMCHD1 and LRIF1 occupancy is reduced at D4Z4 due to lower H3K9me3 resulting in *DUX4* transcriptional de-repression. Additional depletion of SMCHD1 or LRIF1 in both situations results in further *DUX4* upregulation.

Apart from D4Z4, we have also uncovered an unexpected regulation of the *LRIF1* gene by SMCHD1 and the long isoform of LRIF1 itself through their binding to the *LRIF1* promoter in somatic cells. This *LRIF1* gene regulation is likely not relevant for FSHD2 pathology since SMCHD1 nor LRIF1 binding was affected in FSHD2 cells which is consistent with the unchanged expression levels of *LRIF1* in these cells. Together with previous reports, our work extends the knowledge about the versatile involvement of SMCHD1 in regulating different types of chromatin (euchromatin as represented by *LRIF1* locus; facultative heterochromatin as represented by the inactive X [30,34,61–65] and tissue-specific expression attenuation of developmental genes such as clustered *PCDH* [63,64,66,67] or *HOX* genes [30,67] and constitutive heterochromatin exemplified by D4Z4). Interestingly, knocking out SMCHD1 in the 1926iMB cell line did not lead to dysregulated expression of clustered *PCDH* or *HOX* genes or genes on the inactive X, which is consistent with findings obtained from near-diploid RPE1 cells upon SMCHD1 depletion [52] but opposed to findings from HEK293T cells, where SMCHD1 depletion lead to upregulation of *PCDH* β cluster and preferential upregulation of X chromosomal genes [68]. This begs the question what underlies this different sensitivity of SMCHD1-regulated loci to its gene dosage in early development versus in differentiated somatic stage as well as in different cell types.

Materials and Methods

Cell lines and culturing

Immortalized myoblasts were cultured in DMEM/F-10 medium (#31550, Gibco/Life Technologies) supplemented with 20% fetal bovine serum (FBS #10270, Gibco/Life Technologies), 1% Penicillin/Streptomycin (Pen/Strep #15140, Gibco/Life Technologies), with addition of 10 ng/ml rhFGF (#G5071, Promega) and 1 μM dexamethasone (#D2915, Sigma-Aldrich). Myoblasts were fused at 80% confluency by replacing growth medium with DMEM/F-12 Glutamax medium (#31331, Gibco/Life Technologies) containing 1% penicillin/streptomycin and 2% KnockOut serum replacement formulation (#10828, Gibco/Life Technologies) for 2 to 5 days depending on the cell line. The HEK293T cells were grown in Gibco DMEM, High Glucose, Pyruvate (#119950, Gibco/Life Technologies) with addition of 10% FBS and 1% Penicillin/streptomycin. Primary fibroblasts were cultured in DMEM/F-12 GlutaMAX™ Supplement (Gibco, #10565018) supplemented with 20% FBS, 1% penicillin/streptomycin, 10 mM HEPES (Gibco, #15630056) and 1 mM Sodium Pyruvate (Gibco, #11360070). The human colon carcinoma HCT116 (WT and DKO) cell lines were grown in McCoy's 5A medium (ThermoFisher Scientific, #16600082) supplemented with 10% FBS and 1% penicillin/streptomycin. Additional information about cell lines is provided in Suppl. Table 4.

Generation of knock-out cell lines with CRISPR/Cas9

The sgRNA sequences targeting exon 3 of *SMCHD1*, exon 2 of *LRIF1* (LRIF1 long isoform specific knock-out) or exon 3 of *LRIF1* (both LRIF1 isoforms knock-out) were designed using the CRISPOR online design tool[69] (available at http://crispor.tefor.net/crispor.py). The sgRNA oligonucleotides (sequences in Suppl. Table 5) were cloned into the pX458 vector (Addgene #458138) via BbsI sites as described previously.[70] Immortalized myoblasts were seeded in 6-well plates to 60-70% confluency one day prior to transfection. Cells were transfected with 2.5 μg/well of pX458 vector containing gene-specific sgRNAs with Lipofectamine 3000 reagent according to the manufacturer instructions. 24h after transfection medium was exchanged and 3 days post-transfection GFP positive cells were single-cell sorted to 96-well plates using a BD FACS Aria™ III cell sorter. Single cells were expanded and knock-outs were confirmed by Western blot. As WT control clones were used single-cell sorted cells derived either from untransfected pool or a pool transfected with vector encoding only Cas9 without sgRNA.

siRNA transfections

One day prior transfection, 2 x 10^5 cells were seeded in 6-well plate. The next day, cells were transfected with 25 pmol of gene-specific siRNA mix using RNAiMAX (Thermo Fisher Scientific, #13778075) according to

manufacturer's instructions. A non-targeting siRNA was used as a negative control. Cells were harvested three days post-transfection for respective analysis.

RNA isolation, cDNA synthesis and RT-qPCR

Cells were lysed in Qiazol (Qiagen, #79306) and total RNA was isolated with RNeasy mini kit (Qiagen, #74101) with on-column DNase I treatment. 1-2 µg of RNA was used for cDNA synthesis with poly-dT primer using RevertAid H Minus First Strand cDNA synthesis kit (Thermo Fisher Scientific, #K1621). Gene expression was analyzed in technical triplicates using iQ™ SYBR® Green Supermix (Biorad, #1708887) on CFX384 Touch Real-Time PCR Detection System. All primers used for RT-qPCR are listed in Suppl. Table 6. *GUSB* was used as a housekeeping gene.

SDS-PAGE followed by Western blot

Cells were washed twice with ice-cold PBS and resuspended in RIPA buffer (0.1% SDS, 1% Igepal CA-630, 150mM NaCl, 0.5% Sodium Deoxycholate, 20mM EDTA) supplemented with Complete™, EDTA-free Protease Inhibitor Cocktail (1 tablet/50 ml buffer) (Sigma-Aldrich, #11873580001). Samples were kept on ice for 10 min followed by centrifugation at 16,000g for 10 min at 4°C. Protein concentration of the supernatant was determined with Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, #23225). For western blotting, samples were resolved on Novex™ NuPAGE™ 4-12% Bis-Tris protein gels (Invitrogen, #NP0321BOX) and transferred to Immobilon-FL PVDF membrane (Merck, #IPFL00010). The membrane was blocked for 1 h in 4% skim milk in PBS followed by incubation overnight at 4°C with primary antibodies: RαSMCHD1 (1:1000, Abcam #ab176731), RαLRIF1 (1:1000, Proteintech #26115-1-AP) and Mα-αTubulin (1:4000, Sigma-Aldrich #T6199). The next day, membranes were washed twice with PBS-T (0.01% Tween 20) and incubated with following secondary antibodies: IRDye® 800CW goat anti-rabbit IgG (1:10,000, Li-cor #P/N 925-32211) and IRDye® 680CW donkey anti-mouse IgG (1:10,000, #P/N 925-68072) for 1h at room temperature. Membranes were washed twice with PBS-T prior scanning on Odyssey® CLx Imaging System (Li-cor).

DR1 region methylation analysis by bisulfite PCR followed by TOPO-TA subcloning

Bisulfite conversion of genomic DNA was carried out using the EZ DNA Methylation-Lightning kit (Zymo Research, #D5030) according to the manufacturer's protocol. Converted DNA was used to amplify the DR1 region using FastStart™ Taq DNA polymerase (Sigma-Aldrich, #12032902001) with the following primers: 5'-TCGTCGGCAGCGT-CAGATGTGTATAAGAGACAGGGGTTGAGGGTTGGGTTTATA-3' and 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGA-CAAAACTCAACCTAAAAATATAC-3'. PCR products were resolved on 2% TBE agarose gel, gel extracted with NucleoSpin Gel & PCR Clean-up kit (Bioke, #740609) and subcloned into the TOPO-TA vector (Thermo Fisher Scientific, #45-064-1) according to manufacturer's protocol. Plasmids were isolated from independent bacterial colonies and sent for Sanger sequencing (Macrogen). BiQ Analyzer software was used for the methylation analysis.

Chromatin immunoprecipitation (ChIP)

Cells were crosslinked for 10 min at room temperature with formaldehyde of 1% final concentration. The reaction was quenched by adding glycine to 125 mM final concentration. Cells were washed twice with PBS containing 0.5 mM PMSF (Sigma-Aldrich, #93482), collected by scraping and spun at 500g for 10 min at 4°C. Cell pellets were resuspended in the ice-cold ChIP buffer (1.5 ml lysis buffer/10 x 10^6 cells) (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5 % Igepal CA-630, 1% Triton X-100) supplemented with Complete™ Protease Inhibitor Cocktail table (Sigma-Aldrich, #11697498001). After 10 min incubation on ice, samples were spun down at 8,000 g for 2 min at 4°C. The nuclear pellets were again resuspended in ChIP buffer, incubated for 5 min on ice and followed by another round of centrifugation. Final nuclear pellets were resuspended in the ChIP buffer and sonicated at the highest power output for 25 cycles (1 cycle: 30 sec ON/30 sec OFF) using a Bioruptor instrument (Diagenode). For ChIP, chromatin was first pre-cleared with BSA-blocked protein A Sepharose beads (GE Healthcare, #17-5280-21) by rotating for 30-60 min at 4°C. For histone ChIP, 3 µg of chromatin was used and for SMCHD1 and LRIF1 ChIP, 30 µg of chromatin was used in a final volume of 500 µl. 50 µl (10%) of each chromatin was kept as the input sample for later normalization. ChIP was carried out by rotation at 4°C with following primary antibodies: RαSMCHD1 (Abcam, #ab31865), RαLRIF1 (Merck, #ABE1008), RαH3 (Abcam, ab1791), RαH3K4me2 (Active Motif, #39141), RαH3K9me3 (Active Motif, #39161) or RαH3K27me3 (Merck, #07-449). As a negative control, isotype rabbit polyclonal IgG was used (Abcam, #ab37415). The second day, 20 µl of BSA-blocked protein A Sepharose beads were added to all samples and incubated for 2 h at 4°C while rotating. Afterwards, beads were washed as follows: once with low salt wash buffer (1 % Triton X-100, 0.1 % SDS, 2 mM EDTA, 20 mM Tris-HCl, 150 mM NaCl), high salt wash buffer (1 % Triton X-100, 0.1 % SDS, 2 mM EDTA, 20 mM Tris-HCl, 500 mM NaCl), LiCl wash buffer (250 mM LiCl, 1% Igepal CA-630, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl) and twice with TE wash buffer (10 mM Tris-HCl, 1

mM EDTA). For DNA extraction, 10% (w/v) of Chelex 100 resin was added to the beads and boiled at 95°C for 10 min while shaking. Supernatant was used for qPCR analysis. Primers that were used can be found in Suppl. Table 7.

Immunofluorescent staining

Cells were grown on collagen-coated glass-bottom 96-well plates (Greiner Bio-One, #655892) and differentiated for 2-3 days prior staining. Cells were fixed with 2% paraformaldehyde diluted in 1x PBS for 7 min at RT, followed by permeabilization with 1% Triton X-100 diluted in 1x PBS for 10 min at RT. The primary antibody against MYH1E (MF20, Hybridoma Bank, Iowa University) was diluted 1:200 in 1x PBS and incubated with the fixed cells over-night at 4°C. Next day, primary antibody was washed away with 1xPBS and cells were incubated with the secondary antibody (1:500 dilution in 1xPBS) goat-anti-mouse Alexa 594 (Thermo Fisher Scientific, # A21203). Cells were washed with 1x PBS containing 1:1000 dilution of DAPI (Sigma-Aldrich, #268298) for nuclei visualization. Stained cells were imaged with Thermo Cellomics ArrayScan VTI HCS Reader and 100 images per cell line were taken at 20× magnification. Images were analyzed using CellProfiler Software (v2.1.1) with a custom made analysis pipeline. In short, nuclei were segmented based on DAPI staining and individual nuclei were identified based on shape and size. Myotubes were segmented based on MYH1E staining and used as mask overlay to discriminate myotube nuclei from myoblast nuclei. Fusion index was calculated as the percentage of myotube nuclei as compared to the total number of nuclei per image.

Poly-A RNA-seq and data analysis

Total RNA was isolated as described above and poly-A RNA-seq was outsourced to GenomeScan B.V.. Sequencing libraries were prepared with NEBNext® Ultra™ II RNA Library Prep Kit for Illumina® kit (New England Biolabs, #E7775) according to the manufacturer's manual. Samples were sequenced as 150 bp paired-end on a NovaSeq6000 instrument. Quality assessment of the raw sequencing reads was done using FastQC v0.11.6. Adapters were removed by TrimGalore v0.4.5 with option --paired. The remaining quality-filtered reads were aligned to the human reference genome (version hg38) with the corresponding annotation file from Ensemble using the STAR aligner v2.5.1. Read count table was obtained with HTSeq-count v0.9.1 using the GENCODE V29 annotation with the option "–stranded no". The differential expression statistical analysis was done with DESeq2 v1.24.0 (R package) with default settings. The final list of differentially expressed genes contains genes for which the adjusted p-value (Benjamini-Hochberg correction) is < 0.05. RNA-seq plots were generated with ggplot2 v3.3.3 (R package). Raw sequencing files have been under GEO accession number GSE185511.

Acknowledgements

We thank members of the Van der Maarel lab for all their helpful suggestions and we are thankful to LUMC Flow cytometry Core Facility for their technical support.

Funding

This study was supported by grants from the US National Institute of Arthritis and Musculoskeletal and Skin Diseases (R01AR066248) and the Prinses Beatrix Spierfonds (W.OP14–01; W.OR15–26). DSS, IW, AVDH, JB and SMVDM are members of the European Reference Network for Rare Neuromuscular Diseases [ERN EURO-NMD].

Conflict of interest

Authors declare no conflict of interest.

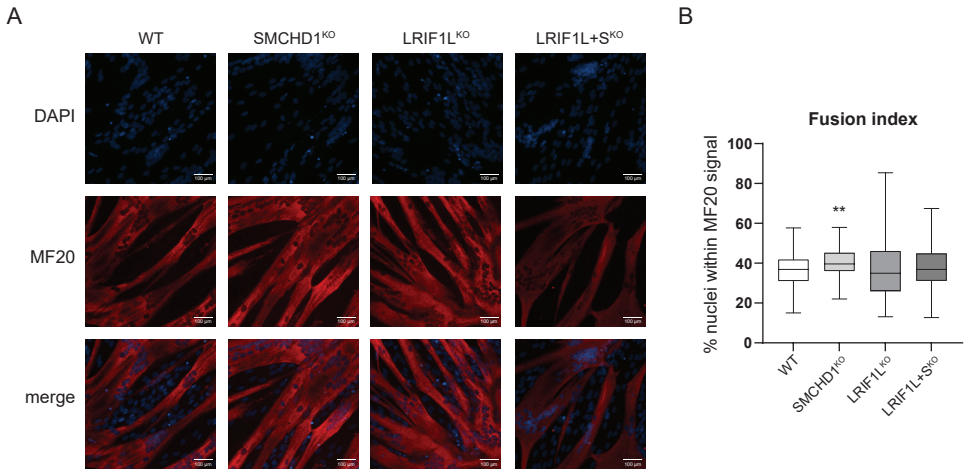
References

1. Snider, L. *et al.* Facioscapulohumeral Dystrophy: Incomplete Suppression of a Retrotransposed Gene. *PLoS Genet.* **6**, e1001181 (2010).
2. Dixit, M. *et al.* DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18157–18162 (2007).
3. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
4. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
5. Jagannathan, S. *et al.* Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. *Hum. Mol. Genet.* **25**, ddw271 (2016).
6. Wong, C.-J. *et al.* Longitudinal measures of RNA expression and disease activity in FSHD muscle biopsies. *Hum. Mol. Genet.* **29**, 1030–1043 (2020).
7. Wang, L. H. *et al.* MRI-informed muscle biopsies correlate MRI with pathology and DUX4 target gene expression in FSHD. *Hum. Mol. Genet.* **28**, 476–486 (2019).
8. Geng, L. N. *et al.* DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev. Cell* **22**, 38–51 (2012).
9. Yao, Z. *et al.* DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Hum. Mol. Genet.* **23**, 5342–5352 (2014).
10. Young, J. M. *et al.* DUX4 Binding to Retroelements Creates Promoters That Are Active in FSHD Muscle and Testis. *PLoS Genet.* **9**, e1003947 (2013).
11. Resnick, R. *et al.* DUX4-Induced Histone Variants H3.X and H3.Y Mark DUX4 Target Genes for Expression. *Cell Rep.* **29**, 1812-1820.e5 (2019).
12. Choi, S. H. *et al.* DUX4 recruits p300/CBP through its C-terminus and induces global H3K27 acetylation changes. *Nucleic Acids Res.* **44**, 5161–73 (2016).
13. Lemmers, R. J. L. F. *et al.* Worldwide Population Analysis of the 4q and 10q Subtelomeres Identifies Only Four Discrete Interchromosomal Sequence Transfers in Human Evolution. *Am. J. Hum. Genet.* **86**, 364–377 (2010).
14. Lemmers, R. J. L. F. *et al.* Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nat. Genet.* **32**, 235–236 (2002).
15. Gannon, O. M., Merida de Long, L. & Saunders, N. A. DUX4 Is Derepressed in Late-Differentiating Keratinocytes in Conjunction with Loss of H3K9me3 Epigenetic Repression. *J. Invest. Dermatol.* **136**, 1299–1302 (2016).
16. Das, S. & Chadwick, B. P. Influence of Repressive Histone and DNA Methylation upon D4Z4 Transcription in Non-Myogenic Cells. *PLoS One* **11**, e0160022 (2016).
17. Dumbovic, G., Forcales, S. V. & Perucho, M. Emerging roles of macrosatellite repeats in genome organization and disease development. *Epigenetics* vol. 12 515–526 (2017).
18. Garrick, D., Fiering, S., Martin, D. I. K. & Whitelaw, E. Repeat-induced gene silencing in mammals. *Nat. Genet.* **18**, 56–59 (1998).
19. Wijmenga, C. *et al.* Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet.* **2**, 26–30 (1992).
20. Deutekom, J. C. T. V. *et al.* FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum. Mol. Genet.* **2**, 2037–2042 (1993).
21. Lemmers, R. J. L. F. *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat. Genet.* **44**, 1370–1374 (2012).
22. van den Boogaard, M. L. *et al.* Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *Am. J. Hum. Genet.* **98**, 1020–1029 (2016).
23. Hamanaka, K. *et al.* Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology* **94**, e2441–e2447 (2020).

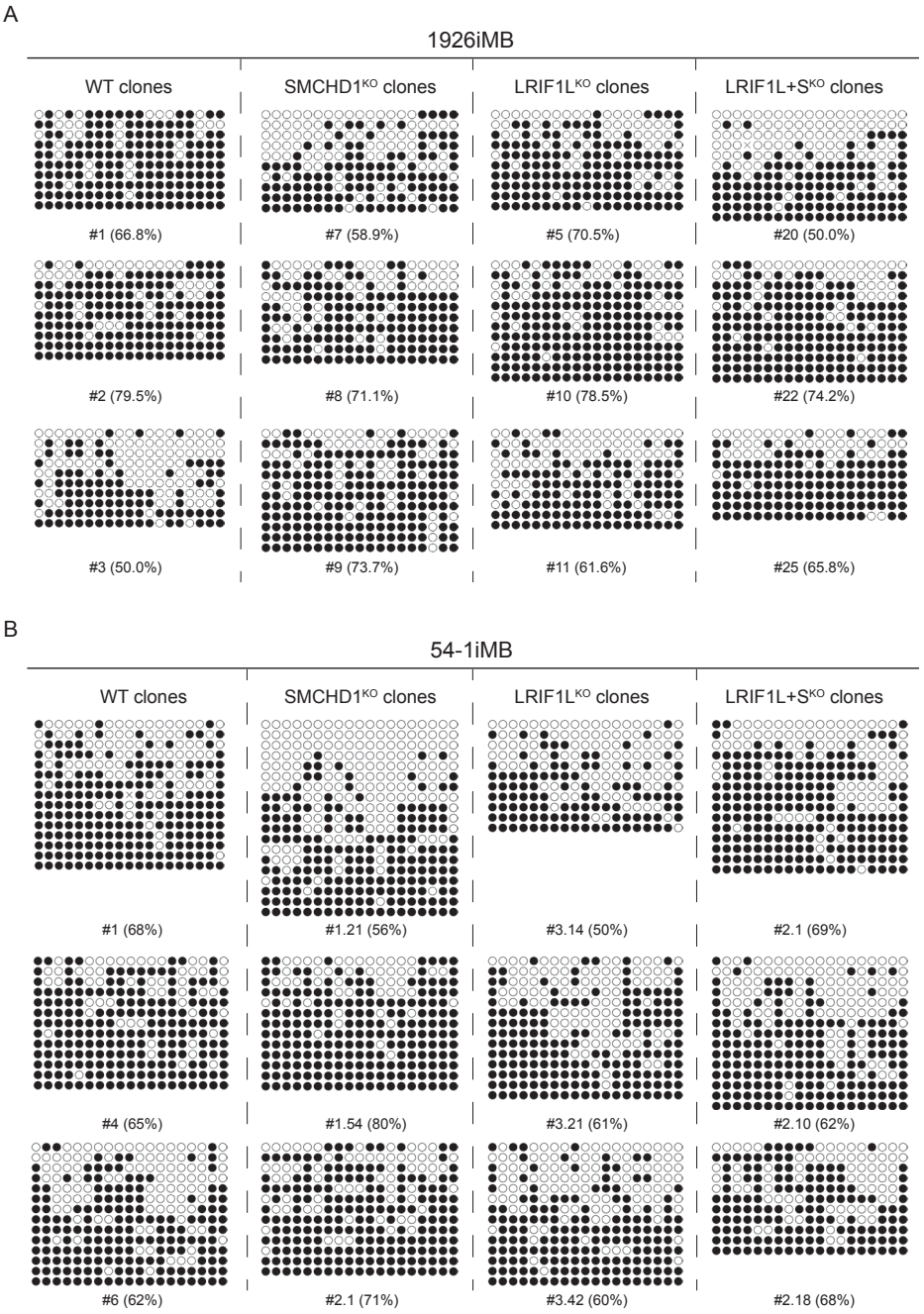
24. de Greef, J. C. *et al.* Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum. Mutat.* **30**, 1449–1459 (2009).
25. van Overveld, P. G. M. *et al.* Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat. Genet.* **35**, 315–317 (2003).
26. Lemmers, R. J. L. F. *et al.* Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum. Mol. Genet.* **24**, 659–669 (2015).
27. Lemmers, R. J. L. F. *et al.* Hemizygosity for *SMCHD1* in Facioscapulohumeral Muscular Dystrophy Type 2: Consequences for 18p Deletion Syndrome. *Hum. Mutat.* **36**, 679–683 (2015).
28. van den Boogaard, M. L. *et al.* Double SMCHD1 variants in FSHD2: the synergistic effect of two SMCHD1 variants on D4Z4 hypomethylation and disease penetrance in FSHD2. *Eur. J. Hum. Genet.* **24**, 78–85 (2016).
29. Brideau, N. J. *et al.* Independent Mechanisms Target SMCHD1 to Trimethylated Histone H3 Lysine 9-Modified Chromatin and the Inactive X Chromosome. *Mol. Cell. Biol.* **35**, 4053–68 (2015).
30. Jansz, N. *et al.* Smchd1 regulates long-range chromatin interactions on the inactive X chromosome and at Hox clusters. *Nat. Struct. Mol. Biol.* **25**, 766–777 (2018).
31. Jansz, N. *et al.* Smchd1 Targeting to the Inactive X Is Dependent on the Xist-HnrnpK-PRC1 Pathway. *Cell Rep.* **25**, 1912-1923.e9 (2018).
32. Wang, C.-Y., Colognori, D., Sunwoo, H., Wang, D. & Lee, J. T. PRC1 collaborates with SMCHD1 to fold the X-chromosome and spread Xist RNA between chromosome compartments. *Nat. Commun.* **10**, 2950 (2019).
33. Gdula, M. R. *et al.* The non-canonical SMC protein SmcHD1 antagonises TAD formation and compartmentalisation on the inactive X chromosome. *Nat. Commun.* **10**, 30 (2019).
34. Wang, C.-Y., Jégu, T., Chu, H.-P., Oh, H. J. & Lee, J. T. SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X. *Cell* **174**, 406-421.e25 (2018).
35. Balog, J. *et al.* Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. *Epigenetics* **10**, 1133–1142 (2015).
36. Shaw, N. D. *et al.* SMCHD1 mutations associated with a rare muscular dystrophy can also cause isolated arhinia and Bosma arhinia microphthalmia syndrome. *Nat. Genet.* **49**, 238–248 (2017).
37. Gordon, C. T. *et al.* De novo mutations in SMCHD1 cause Bosma arhinia microphthalmia syndrome and abrogate nasal development. *Nat. Genet.* **49**, 249–255 (2017).
38. Dion, C. *et al.* SMCHD1 is involved in *de novo* methylation of the *DUX4* -encoding D4Z4 macrosatellite. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz005.
39. Gurzau, A. D. *et al.* FSHD2- and BAMS-associated mutations confer opposing effects on SMCHD1 function. *J. Biol. Chem.* **293**, 9841–9853 (2018).
40. Lemmers, R. J. L. F. *et al.* SMCHD1 mutation spectrum for facioscapulohumeral muscular dystrophy type 2 (FSHD2) and Bosma arhinia microphthalmia syndrome (BAMS) reveals disease-specific localisation of variants in the ATPase domain. *J. Med. Genet.* **56**, 693–700 (2019).
41. Mul, K. *et al.* FSHD type 2 and Bosma arhinia microphthalmia syndrome. *Neurology* **91**, e562–e570 (2018).
42. Xu, G.-L. *et al.* Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187–191 (1999).
43. Hansen, R. S. *et al.* The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14412–14417 (1999).
44. Kondo, T. *et al.* Whole-genome methylation scan in ICF syndrome: hypomethylation of non-satellite DNA repeats D4Z4 and NBL2. *Hum. Mol. Genet.* **9**, 597–604 (2000).
45. Goossens, R. *et al.* Intronic SMCHD1 variants in FSHD: Testing the potential for CRISPR-Cas9 genome editing. *J. Med. Genet.* **56**, 828–837 (2019).
46. Hartweck, L. M. *et al.* A focal domain of extreme demethylation within D4Z4 in FSHD2. *Neurology* **80**, 392–399 (2013).
47. Zeng, W. *et al.* Specific Loss of Histone H3 Lysine 9 Trimethylation and HP1γ/Cohesin Binding at D4Z4 Repeats Is Associated with Facioscapulohumeral Dystrophy (FSHD). *PLoS Genet.* **5**, e1000559 (2009).

48. Villaseñor, R. *et al.* ChromID identifies the protein interactome at chromatin marks. *Nat. Biotechnol.* **38**, 728–736 (2020).
49. Vermeulen, M. *et al.* Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers. *Cell* **142**, 967–980 (2010).
50. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M. & Mann, M. A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics. *Mol. Cell* **49**, 368–378 (2013).
51. Zeng, W. *et al.* Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. *Hum. Mutat.* **35**, 998–1010 (2014).
52. Nozawa, R.-S. *et al.* Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* **20**, 566–573 (2013).
53. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).
54. Huntriss, J. *et al.* Expression of mRNAs for DNA Methyltransferases and Methyl-CpG-Binding Proteins in the Human Female Germ Line, Preimplantation Embryos, and Embryonic Stem Cells. *Mol. Reprod. Dev.* **67**, 323–336 (2004).
55. Liao, J. *et al.* Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* **47**, 469–478 (2015).
56. Duymich, C. E., Charlet, J., Yang, X., Jones, P. A. & Liang, G. DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. *Nat. Commun.* **7**, 11453 (2016).
57. Zhou, J. *et al.* H19 lncRNA alters DNA methylation genome wide by regulating S-adenosylhomocysteine hydrolase. *Nat. Commun.* **6**, 1–13 (2015).
58. Barrès, R. *et al.* Non-CpG Methylation of the PGC-1α Promoter through DNMT3B Controls Mitochondrial Density. *Cell Metab.* **10**, 189–198 (2009).
59. Sakakibara, Y. *et al.* Role of SmcHD1 in establishment of epigenetic states required for the maintenance of the X-inactivated state in mice. *Development* **145**, dev166462 (2018).
60. Ichihara, S., Nagao, K., Sakaguchi, T., Obuse, C. & Sado, T. SmcHD1 underlies the formation of H3K9me3 blocks on the inactive X chromosome in mice. *bioRxiv* 2021.08.23.457321 (2021) doi:10.1101/2021.08.23.457321.
61. Blewitt, M. E. *et al.* SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nat. Genet.* **40**, 663–669 (2008).
62. Gendrel, A.-V. *et al.* Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome. *Dev. Cell* **23**, 265–279 (2012).
63. Gendrel, A.-V. *et al.* Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. *Mol. Cell. Biol.* **33**, 3150–65 (2013).
64. Mould, A. W. *et al.* Smchd1 regulates a subset of autosomal genes subject to monoallelic expression in addition to being critical for X inactivation. *Epigenetics Chromatin 2013 61* **6**, 1–16 (2013).
65. Wang, C. Y., Colognori, D., Sunwoo, H., Wang, D. & Lee, J. T. PRC1 collaborates with SMCHD1 to fold the X-chromosome and spread Xist RNA between chromosome compartments. *Nat. Commun.* **10**, (2019).
66. Mason, A. G. *et al.* SMCHD1 regulates a limited set of gene clusters on autosomal chromosomes. *Skelet. Muscle 2017 71* **7**, 1–13 (2017).
67. Chen, K. *et al.* Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3535-44 (2015).
68. Massah, S. *et al.* Epigenetic Characterization of the Growth Hormone Gene Identifies SmcHD1 as a Regulator of Autosomal Gene Clusters. *PLoS One* **9**, e97535 (2014).
69. Concordet, J. P. & Haeussler, M. CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
70. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

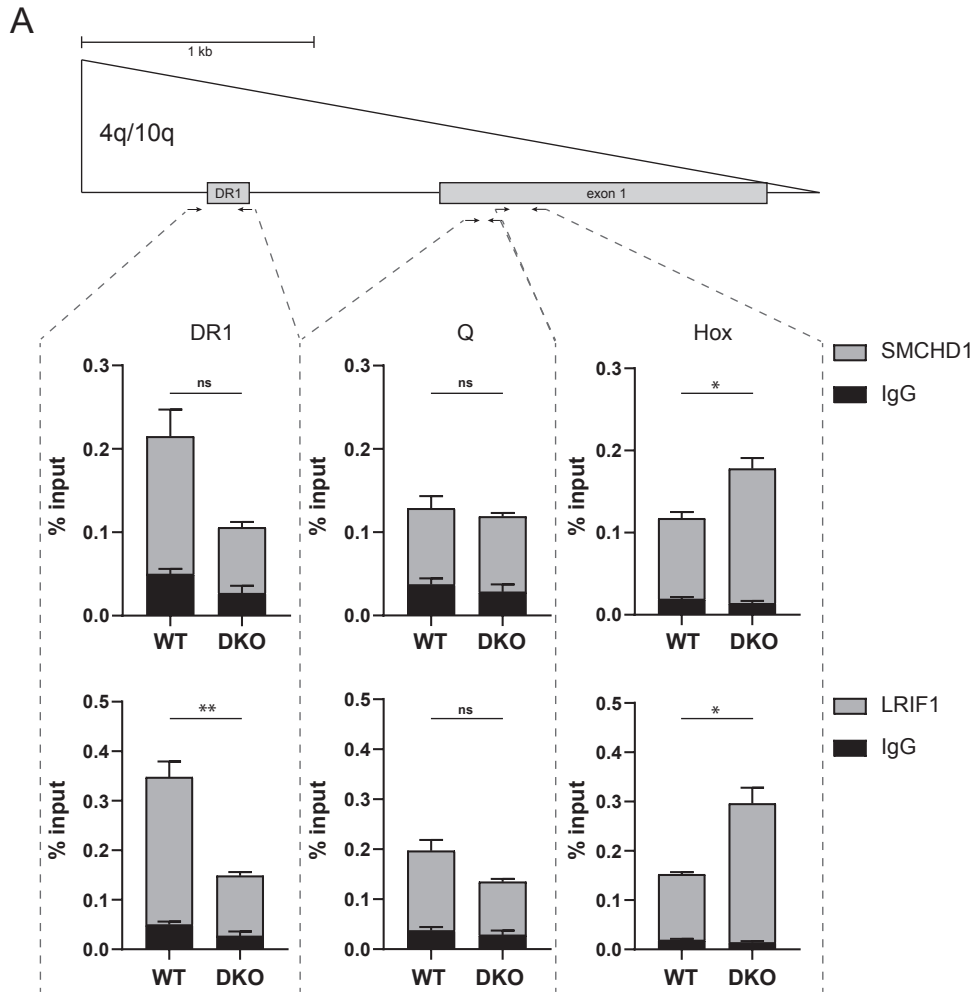
Supplementary Information



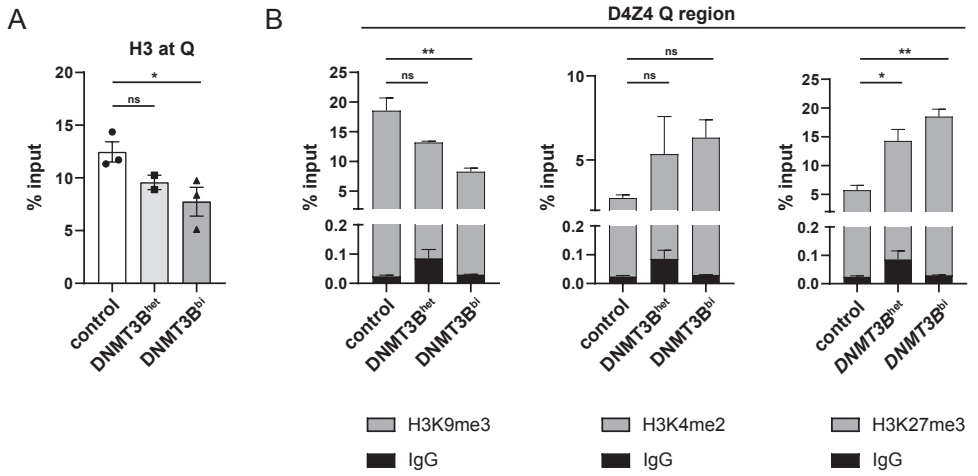
Suppl. Figure 1. A) MYH1E staining (in red) of one WT and one clone of each knock-out condition of the 1926iMB cell line. Nuclei are counterstained with DAPI (in blue). Merged images show overlay of DAPI and MYH1E staining. Scale bar is 100 µm. **B)** Fusion index (=number of nuclei inside myotubes as a percentage of the total number of nuclei) calculated for each 1926 clone that is depicted in A). Box represents 25th to 75th percentile and line represents a median value of all fusion indexes calculated from 100 images per clone, totalling on average to 10,000 analysed nuclei positions per clone. Statistical significance between WT and KO groups was calculated by one-way ANOVA with Dunnett's post hoc test (**p<0.01).



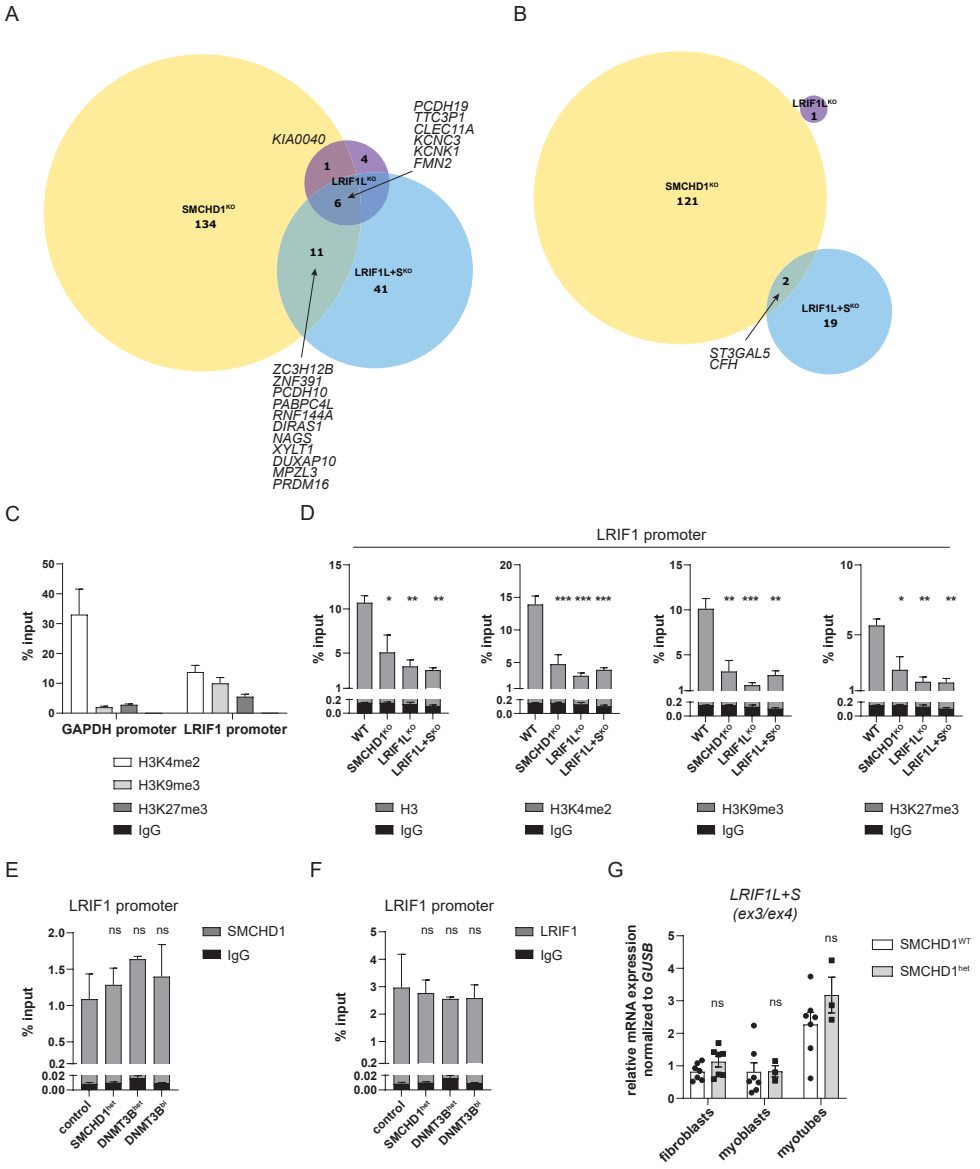
Suppl. Figure 2. Methylation of individual CpGs in D4Z4 DR1 is unchanged in different immortalized myoblast KO clones related to Figure 2A and 2B. A) Lollipop representation of DR1 site methylation in different clones derived from 1926iMB cell line. **B)** Lollipop representation of DR1 site methylation in different clones derived from 54-1iMB cell line. Full circles represent methylated CpGs and open circles represent unmethylated CpGs. Mean methylation is calculated in the brackets and plotted in Figure 2A for 1926 clones and 2B for 54-1 clones.



Suppl. Figure 3. SMCHD1 and LRIF1 binding to D4Z4 is reduced in HCT116 DKO cells. A) SMCHD1 and LRIF1 ChIP-qPCR in HCT116 WT and DKO cells. Schematic of one D4Z4 unit with the position of the three regions within D4Z4 examined by ChIP-qPCR are indicated (DR1, Q, HOX). Bars and whiskers represent mean ± SEM of three experiments. Isotype specific IgG was used for background control. Statistical significance was calculated with an unpaired t-test (**p<0.01,*p<0.05, ns – not significant).



Suppl. Figure 4. Histone mark profiles in primary fibroblasts carrying heterozygous (*DNMT3B*^{het}) or biallelic *DNMT3B* mutation (*DNMT3B*^{bi}) resembles those reported in FSHD2 cases due to *SMCHD1* or *LRIF1* mutations. **A)** H3 ChIP-qPCR of the D4Z4 Q region in primary control fibroblasts (n = 3, lines: 2374, 2417, 2397) or fibroblasts carrying either heterozygous *DNMT3B* mutation (n = 2, lines: v294, b974) or biallelic *DNMT3B* mutations (n = 3, lines: GM08714, Rf614, Rf699.3). Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. **B)** ChIP-qPCR of selected histone marks at the D4Z4 Q region in the same primary fibroblast sets as in A). Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (**p<0.01, *p<0.05, ns – not significant).



Suppl. Figure 5. *LRIF1* expression is sensitive to somatic LRIF1 and SMCHD1 gene dosage perturbations. A) Venn diagram of differentially upregulated genes overlapping between the knock-out conditions. **B)** Venn diagram of differentially downregulated genes overlapping between the knock-out conditions. **C)** Selected histone marks ChIP-qPCRs of *GAPDH* and *LRIF1* promoter in WT 1926iMB clones. Bars and whiskers represent mean ± SEM (ns = 3). **D)** H3 and selected H3-coupled marks ChIP-qPCR of the LRIF1 promoter in WT and different 1926iMB knock-out conditions. Bars and whiskers represent mean ± SEM (ns = 3). Isotype specific IgG was used for background control. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (***<0.001, **p<0.01, *p<0.05). **E)** SMCHD1 ChIP-qPCR of the LRIF1 promoter in primary control fibroblasts (n = 3, lines: 2524, 2397, 2333) and fibroblasts carrying either a heterozygous *SMCHD1* mutation (n = 3, lines: 2440, 2337, 2332), a heterozygous *DNMT3B* mutation (n = 2, lines: v294, b974) or biallelic *DNMT3B* mutations (n = 2, lines: Rf699.3,

Rf286.3). Bars and whiskers represent mean ± SEM. Isotype specific IgG was used for background control. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (ns - not significant). **I)** LRIF1 ChIP-qPCR of the *LRIF1* promoter in the same samples as in H)**.** Statistical significance between WT and mutant groups was calculated by one-way ANOVA with Dunnett's post hoc test (ns – not significant). **J)** RT-qPCR of *LRIF1* in different control and SMCHD1 haploinsufficient primary cell lines (fibroblasts, myoblasts or differentiated myotubes). Bars and whiskers represent mean ± SEM. Each dot represents cell line derived from a unique individual. Statistical significance between control and $SMCHD1^{het}$ group was calculated with an unpaired t-test (ns – not significant).

Suppl. Table 1. Differentially expressed genes upon SMCHD1^{KO} in 1926iMB.

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000101596	18	2655738	2805017	1	SMCHD1	1087,959937	-2,290156456	1,06319E-54
ENSG00000235750	1	175156986	175192999	-1	KIAA0040	80,1925354	6,038995895	5,20483E-23
ENSG00000165194	X	100291644	100410273	-1	PCDH19	107,10547	7,705851884	2,47299E-22
ENSG00000112964	5	42423439	42721878	1	GHR	88,11332663	2,640684145	3,6315E-19
ENSG00000108691	17	34255218	34257203	1	CCL2	74,01263878	10,76120685	3,73681E-15
ENSG00000114200	3	165772904	165837462	-1	BCHE	87,97791929	11,00922465	1,86384E-14
ENSG00000227115	18	51346249	51643939	1	LINC01630	35,46821318	9,700681245	8,05043E-12
ENSG00000215105	X	75740831	75746911	-1	TTC3P1	285,3096611	1,424714373	1,23995E-10
ENSG00000138650	4	133149294	133208606	1	PCDH10	1405,670404	8,485124069	1,20881E-09
ENSG00000142611	1	3069168	3438621	1	PRDM16	73,83955733	2,989932254	1,98212E-09
ENSG00000122824	X	51332231	51337525	1	NUDT10	41,54354926	2,127695142	2,48146E-09
ENSG00000135750	1	233614004	233672512	1	KCNK1	26,03652631	8,720818993	1,67144E-08
ENSG00000204442	13	107163510	107866735	-1	FAM155A	89,38264416	5,792525693	2,50078E-08
ENSG00000105472	19	50723364	50725718	1	CLEC11A	258,7588469	6,002938287	2,8165E-08
ENSG00000254535	4	134196333	134201789	-1	PABPC4L	46,96301519	3,577973869	6,41914E-08
ENSG00000081803	7	122318411	122886759	-1	CADPS2	18,89032563	4,470147246	7,20791E-08
ENSG00000112378	6	138088505	138107419	-1	PERP	487,2955328	1,488232072	1,01831E-07
ENSG00000146242	6	82363206	82370828	1	TPBG	657,8327384	1,165228919	1,42125E-07
ENSG00000181381	4	168356735	168537786	-1	DDX60L	937,5520335	-0,946809136	1,66189E-07
ENSG00000133985	14	70641916	70675366	1	TTC9	799,1466894	-1,006669079	2,59181E-07
ENSG00000176485	11	63573195	63616883	-1	PLAAT3	778,4551035	-0,94212166	3,67694E-07
ENSG00000154529	9	41890314	42129510	-1	CNTNAP3B	266,4868505	5,985254449	3,67694E-07
ENSG00000114541	3	69168782	69542583	-1	FRMD4B	375,2836016	-2,83411861	4,02022E-07
ENSG00000155816	1	240014348	240475187	1	FMN2	736,055994	5,819976795	6,4658E-07
ENSG00000189221	X	43654907	43746824	1	MAOA	85,35921434	7,504824313	7,37507E-07
ENSG00000067141	15	73051710	73305205	1	NEO1	82,7668465	1,590848537	8,09496E-07
ENSG00000158813	X	69616067	70039472	1	EDA	20,48380635	3,320312352	1,06407E-06
ENSG00000115525	2	85837120	85905199	-1	ST3GAL5	529,368235	-0,803048943	1,46085E-06
ENSG00000100344	22	43923739	43964488	1	PNPLA3	556,1629158	-1,340012265	3,94501E-06

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000113361	5	31193686	31329146	1	CDH6	16,8537064	3,926119946	4,41601E-06
ENSG00000118515	6	134169246	134318112	-1	SGK1	1962,388374	0,955412853	6,37975E-06
ENSG00000154654	21	20998409	21543329	1	NCAM2	32,18749002	3,207090531	1,02343E-05
ENSG00000149256	11	78652831	79440948	-1	TENM4	115,6078876	1,97390139	1,1537E-05
ENSG00000184347	5	168661733	169301129	-1	SLIT3	39,4042427	2,882142522	1,24928E-05
ENSG00000235034	19	50649445	50659310	1	C19orf81	14,22100712	7,05682584	1,70279E-05
ENSG00000131398	19	50311937	50333515	-1	KCNC3	53,96334835	5,869713777	1,81771E-05
ENSG00000277406	1	146321214	146401647	-1	SEC22B4P	73,70358476	-1,517010544	2,07298E-05
ENSG00000151692	2	6917412	7068286	1	RNF144A	17,32295849	4,101930896	2,07298E-05
ENSG00000150471	4	61201258	62078335	1	ADGRL3	11,1107094	5,508573686	2,07298E-05
ENSG00000102043	X	64268081	64395452	-1	MTMR8	7,142108533	7,29129059	2,28411E-05
ENSG00000280241	4	153948718	154300500	1		80,45469104	1,984521034	2,29682E-05
ENSG00000176049	5	147585438	147782775	-1	JAKMIP2	9,569566539	7,665252203	2,4737E-05
ENSG00000124613	6	27374615	27403904	1	ZNF391	42,65278443	1,737168913	2,66336E-05
ENSG00000244306	14	19268853	19337730	-1		64,05503278	9,820557004	2,66336E-05
ENSG00000145423	4	153780591	153789083	-1	SFRP2	14,29179085	5,764181795	3,47067E-05
ENSG00000237283	1	188705623	188732208	-1		6,743009248	7,285903473	3,67967E-05
ENSG00000112379	6	138161939	138344663	1	ARFGEF3	168,2205383	0,962712525	6,16236E-05
ENSG00000121904	1	33513999	34165842	-1	CSMD2	5,990094358	7,102787175	0,000119071
ENSG00000180875	1	240489573	240612155	-1	GREM2	51,96010687	5,762951101	0,000158483
ENSG00000132821	20	37903111	37945350	1	VSTM2L	171,7786347	-1,537253595	0,000168764
ENSG00000269893	4	118278709	118285316	1	SNHG8	485,5631687	0,59589091	0,000199648
ENSG00000035664	15	63907036	64072033	-1	DAPK2	91,31082351	-0,990592224	0,000239863
ENSG00000158258	3	139935185	140577397	1	CLSTN2	62,15514291	2,580678383	0,000347978
ENSG00000161638	12	54395261	54419266	-1	ITGA5	29192,19314	0,638898417	0,000367516
ENSG00000105419	19	47403124	47419523	-1	MEIS3	258,8057118	1,649119047	0,000370511
ENSG00000183873	3	38548057	38649673	-1	SCN5A	4822,77124	-1,02891125	0,000371797
ENSG00000119280	1	230837119	230869589	-1	C1orf198	1734,633101	-0,83288701	0,000371797
ENSG00000038945	8	16107878	16567490	-1	MSR1	18,66739334	3,591526977	0,000371797
ENSG00000131737	17	41377650	41382403	-1	KRT34	343,6404735	-1,059930649	0,000396403

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000213023	19	50621307	50639881	-1	SYT3	5,621316482	6,914561022	0,000429388
ENSG00000144677	3	37861880	37984469	1	CTDSPL	1677,254794	-0,498002397	0,000447053
ENSG00000064300	17	49495293	49515008	1	NGFR	219,8250889	-2,726413483	0,000448198
ENSG00000168427	2	238138668	238152947	1	KLHL30	72,59713771	-2,910995477	0,000449921
ENSG00000274428	1	146409903	146410062	1		31,97200575	-2,490858487	0,0004572
ENSG00000175764	9	121821928	122093606	-1	TTLL11	277,16322	0,627804439	0,000555683
ENSG00000100784	14	90847861	91060641	-1	RPS6KA5	169,8701662	-1,060108047	0,000669453
ENSG00000261534	9	121815674	121819452	-1		268,707564	1,229973657	0,000669453
ENSG00000154493	10	126424997	126798708	-1	C10orf90	22,776382	1,829136645	0,000669453
ENSG00000133055	1	203167811	203175826	-1	MYBPH	529,0224602	-3,364236112	0,000718873
ENSG00000091409	2	172427354	172506459	1	ITGA6	1189,984463	-1,02989359	0,000743285
ENSG00000182272	11	369499	382117	1	B4GALNT4	48,16787163	2,259304191	0,000763988
ENSG00000092096	14	23346306	23352912	-1	SLC22A17	706,3685183	0,740690626	0,000778416
ENSG00000170396	2	184598529	184939492	1	ZNF804A	5,824890831	7,058182185	0,000778416
ENSG00000100311	22	39223359	39244751	-1	PDGFB	910,208924	-1,386055453	0,000910176
ENSG00000086289	7	37683843	37951936	1	EPDR1	693,0245029	-1,08086894	0,000910176
ENSG00000157111	5	73120569	73131809	1	TMEM171	334,8664535	-0,928511858	0,000910176
ENSG00000151748	14	50632058	50668331	-1	SAV1	1124,107547	-0,593754838	0,000910176
ENSG00000156103	8	88032011	88328025	-1	MMP16	119,9060171	1,3528882	0,000910176
ENSG00000102053	X	65366638	65507887	1	ZC3H12B	33,47085801	2,167510707	0,000978708
ENSG00000163531	1	204828651	205022822	1	NFASC	24,67127622	3,430210596	0,000978708
ENSG00000116337	1	109616104	109632053	1	AMPD2	1374,61669	0,387877329	0,001224593
ENSG00000173621	11	66856647	66860475	1	LRFN4	311,6829343	0,599770515	0,001347461
ENSG00000169760	3	173396284	174286644	1	NLGN1	30,38682558	2,53543795	0,001347461
ENSG00000119866	2	60450520	60555154	-1	BCL11A	16,67933934	3,461014047	0,001347461
ENSG00000004478	12	2794970	2805423	1	FKBP4	1647,383163	0,405254081	0,001429817
ENSG00000173597	4	69721167	69787961	-1	SULT1B1	149,2170238	3,552922491	0,001583608
ENSG00000163623	4	84491987	84498450	-1	NKX6-1	36,13187194	1,740621019	0,00165347
ENSG00000198300	19	56810077	56840728	-1	PEG3	6,434528548	7,189568623	0,001672547
ENSG00000106714	9	39072767	39288315	-1	CNTNAP3	121,0836841	4,060557274	0,001707784

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000116132	1	170662728	170739421	1	PRRX1	603,6838239	1,285775863	0,001882157
ENSG00000116183	1	176463171	176845601	1	PAPPA2	47,52161087	3,289277066	0,001931334
ENSG00000107796	10	88935074	88991339	-1	ACTA2	1392,201328	2,355523753	0,001955073
ENSG00000250007	15	34755084	34813505	1		25,020447	-2,088219746	0,002015724
ENSG00000240065	6	32844136	32859851	1	PSMB9	126,2637475	-1,186173305	0,002466571
ENSG00000110660	11	107790991	107928293	-1	SLC35F2	452,8732867	0,787231859	0,002466571
ENSG00000184986	14	105526583	105530202	1	TMEM121	181,7362314	0,798864314	0,00247804
ENSG00000057294	12	32790745	32896840	-1	PKP2	5,134997462	6,713706855	0,00247804
ENSG00000103489	16	17101769	17470881	-1	XYLT1	1840,556503	1,138589197	0,002648457
ENSG00000162595	1	68045896	68051631	-1	DIRAS3	29,02332411	1,479249567	0,002648457
ENSG00000132357	5	40841308	40860175	1	CARD6	73,26275133	-1,4364391	0,002742259
ENSG00000242516	3	75672300	75742048	1	LINC00960	9,63353942	3,703131832	0,002742259
ENSG00000150556	2	149038107	149215262	1	LYPD6B	696,4440213	-1,82148605	0,002770278
ENSG00000135636	2	71453722	71686768	1	DYSF	13248,38823	-0,89324774	0,002771292
ENSG00000148426	10	11823339	11872277	1	PROSER2	439,7709753	-1,113950826	0,002998797
ENSG00000155324	5	126360132	126496494	1	GRAMD2B	884,3162947	-0,801397096	0,002998797
ENSG00000006740	17	12789498	12991643	1	ARHGAP44	17,3109827	2,112095896	0,002998797
ENSG00000134138	15	36889204	37101299	-1	MEIS2	157,4214016	0,914101122	0,003127714
ENSG00000131831	X	17800049	17861337	-1	RAI2	24,8181068	2,870834161	0,003127714
ENSG00000136997	8	127735434	127742951	1	MYC	3172,545108	0,359041629	0,003140396
ENSG00000003096	X	117897813	118117340	-1	KLHL13	4,233838986	6,304302218	0,003151407
ENSG00000168334	3	39183210	39192596	-1	XIRP1	147,5646127	-1,9471375	0,003251479
ENSG00000163513	3	30606502	30694142	1	TGFBR2	24391,62002	-0,390224517	0,003272958
ENSG00000169213	1	51907956	51990700	-1	RAB3B	3771,617585	-0,50901613	0,003435832
ENSG00000176749	17	32486993	32491253	1	CDK5R1	430,0277535	-0,646728642	0,003591604
ENSG00000169851	4	30720415	31146805	1	PCDH7	346,5784849	1,484403771	0,003757749
ENSG00000138795	4	108047545	108168956	-1	LEF1	8,912752445	3,40366203	0,003757749
ENSG00000235961	X	153072482	153075018	1	PNMA6A	72,93953754	2,057267716	0,003944033
ENSG00000160588	11	118226690	118252365	-1	MPZL3	29,51893106	2,183287641	0,004037196
ENSG00000078114	10	20779973	21174187	-1	NEBL	25,72672554	2,00087894	0,004087604

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000156140	4	72280969	72569221	-1	ADAMTS3	6,687282146	4,458557682	0,004328777
ENSG00000249158	5	140868183	141012344	1	PCDHA11	4,508331315	6,389380261	0,004478888
ENSG00000126790	14	59460363	59484408	-1	L3HYPDH	657,6427526	0,39579392	0,004489233
ENSG00000132688	1	156668763	156677407	-1	NES	75514,84143	-0,435240177	0,004558024
ENSG00000176490	19	2714567	2721372	-1	DIRAS1	79,37025803	1,398086684	0,004621233
ENSG00000172403	4	118850688	119061247	1	SYNPO2	2355,129099	-0,988629854	0,0046527
ENSG00000146426	6	154832697	155257723	1	TIAM2	340,4237131	-0,682049939	0,004888942
ENSG00000110921	12	109573255	109598125	1	MVK	1086,348895	-0,567252289	0,004888942
ENSG00000118276	18	31622247	31685836	-1	B4GALT6	21,36983076	4,486189073	0,004888942
ENSG00000182179	3	49805209	49813953	-1	UBA7	543,2557896	-0,779560881	0,005037187
ENSG00000003402	2	201116154	201176687	1	CFLAR	2761,755454	-0,362118351	0,005037187
ENSG00000187123	2	149329985	149474138	1	LYPD6	864,7463315	-1,0276921	0,005868698
ENSG00000116774	1	113979391	114035572	1	OLFML3	16727,82316	-0,767892378	0,005868698
ENSG00000106123	7	142855061	142871094	1	EPHB6	83,67342155	1,333718572	0,005868698
ENSG00000055163	5	157266079	157395595	1	CYFIP2	105,9457552	1,545352799	0,005868698
ENSG00000066735	14	104138723	104180894	1	KIF26A	5,07825739	6,387532795	0,005868698
ENSG00000196954	11	104942866	104969366	-1	CASP4	834,0533463	-0,383284181	0,006069234
ENSG00000226087	2	47225781	47240886	1		172,4008039	-1,02170088	0,006091176
ENSG00000123360	12	54549601	54579239	1	PDE1B	8,219516043	4,603924625	0,006091176
ENSG00000110042	11	59171430	59208588	1	DTX4	370,23164	-0,802712463	0,006163382
ENSG00000149212	11	95165513	95232541	-1	SESN3	8,800603665	4,514013148	0,006313313
ENSG00000137726	11	117836976	117877486	-1	FXYD6	12,1199786	4,291433954	0,006422093
ENSG00000284977	9	117759455	117879988	1		25,90832275	-1,57420254	0,006519435
ENSG00000272872	22	15823197	15823890	1		8,552207589	6,926686942	0,006782856
ENSG00000165996	10	17589032	17617374	-1	HACD1	663,536182	-0,542682078	0,006835207
ENSG00000163840	3	122564338	122575203	1	DTX3L	488,1293482	-0,557667781	0,006887589
ENSG00000172201	6	19837370	19842197	1	ID4	7,996889842	5,555080913	0,007075379
ENSG00000127329	12	70515866	70637440	-1	PTPRB	120,3735356	-1,258948986	0,00718646
ENSG00000206195	22	15784959	15829984	1	DUXAP8	55,59443033	9,850354265	0,007338464
ENSG00000074964	1	17539698	17697874	1	ARHGEF10L	2478,000634	-0,307092923	0,007553801

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000084764	2	26970612	27027196	1	MAPRE3	76,75029806	0,853195469	0,007553801
ENSG00000147036	X	37571569	37684463	1	LANCL3	51,46601781	1,311238897	0,007553801
ENSG00000105088	19	9853718	9936552	-1		12,75364502	2,614741268	0,008833257
ENSG00000119771	2	23385179	23708611	1	KLHL29	230,9876323	0,628502211	0,008836133
ENSG00000044524	3	89107621	89482134	1	EPHA3	5,850489604	5,157657915	0,008913862
ENSG00000129910	16	89171767	89195492	1	CDH15	6690,064605	-1,040709707	0,009001501
ENSG00000142408	19	53963040	53990215	1	CACNG8	8,604455735	4,631920419	0,009001501
ENSG00000187098	3	69739435	69968337	1	MITF	205,6651075	-0,64540724	0,009304993
ENSG00000118564	4	15604539	15681679	-1	FBXL5	2292,209215	-0,311524216	0,009633534
ENSG00000130222	9	89605012	89606555	1	GADD45G	5,291588875	5,732707846	0,00978631
ENSG00000116117	2	204545475	205620162	1	PARD3B	873,8975214	-0,624679217	0,010192682
ENSG00000011347	11	61515313	61581148	-1	SYT7	11,62963356	2,755937949	0,010192682
ENSG00000123500	6	116118909	116158747	-1	COL10A1	16,24084211	-2,208646818	0,010264202
ENSG00000161653	17	44004546	44009063	1	NAGS	383,0744498	0,790797109	0,010504457
ENSG00000166016	11	34150987	34358010	-1	ABTB2	231,0457999	-0,703063939	0,01149597
ENSG00000120149	5	174724582	174730896	1	MSX2	277,2782152	-0,888368635	0,011621381
ENSG00000134508	18	23134564	23260467	1	CABLES1	218,9168856	-1,336428859	0,011717281
ENSG00000184349	5	107376889	107670937	-1	EFNA5	225,4454764	-0,58760296	0,012310598
ENSG00000167549	17	29614756	29622907	-1	CORO6	1455,599022	-0,919253533	0,012484548
ENSG00000172379	15	80404350	80597933	1	ARNT2	1610,951036	-1,113286001	0,012693432
ENSG00000060140	12	10618923	10674318	-1	STYK1	65,05493216	-1,047859952	0,012697423
ENSG00000143341	1	185734391	186190949	1	HMCN1	18,63544559	2,430577075	0,012863893
ENSG00000013016	2	31234152	31269451	1	EHD3	1000,705064	0,41174549	0,012888746
ENSG00000104361	8	98189826	98294393	-1	NIPAL2	446,0983077	-0,59024164	0,01321272
ENSG00000111199	12	109783085	109833401	-1	TRPV4	156,0393371	-0,773788482	0,013376565
ENSG00000100342	22	36253010	36267530	1	APOL1	156,3268208	-1,312276282	0,013628425
ENSG00000121769	1	31365625	31376850	-1	FABP3	417,2454364	-1,274486173	0,013628425
ENSG00000166482	17	19383442	19387240	-1	MFAP4	205,7394448	1,060851687	0,013628425
ENSG00000161681	19	50661827	50719450	-1	SHANK1	5,808981085	6,439494246	0,013628425
ENSG00000103888	15	80779343	80951776	1	CEMIP	3575,112821	-1,436861365	0,013862274

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000154914	17	9644698	9729687	1	USP43	6,570979281	6,004717455	0,013961714
ENSG00000198879	10	7158624	7411486	-1	SFMBT2	5,65258205	3,72664849	0,014615901
ENSG00000146094	5	177501907	177511274	-1	DOK3	310,1533971	-0,595575492	0,014860934
ENSG00000000971	1	196651878	196747504	1	CFH	2049,525726	-0,392008869	0,015226976
ENSG00000101040	20	47209214	47356889	-1	ZMYND8	3903,640516	0,288247606	0,016657617
ENSG00000158710	1	159918107	159925732	-1	TAGLN2	29878,75721	0,40319687	0,01700813
ENSG00000108018	10	106573663	107164534	-1	SORCS1	12,78464003	2,909967006	0,01714321
ENSG00000140092	14	91869412	91947987	-1	FBLN5	926,5942237	0,674045902	0,01790892
ENSG00000168497	2	191834310	191847088	-1	CAVIN2	8750,516249	-0,691707057	0,018627768
ENSG00000135324	6	84033772	84090881	1	MRAP2	7,782017648	2,420667603	0,0208869
ENSG00000140853	16	56989485	57083531	1	NLRC5	505,397748	-0,481362193	0,021152066
ENSG00000128335	22	36226209	36239954	-1	APOL2	1191,083454	-0,30775726	0,021152066
ENSG00000135486	12	54280193	54287088	1	HNRNPA1	32436,38736	0,250861441	0,021152066
ENSG00000145147	4	20251905	20620561	1	SLIT2	645,8999212	0,752843322	0,021152066
ENSG00000132932	13	25371974	26025851	1	ATP8A2	208,9982347	-0,739388806	0,022162097
ENSG00000186868	17	45894382	46028334	1	MAPT	49,42420169	1,622577529	0,022162097
ENSG00000211445	5	151020438	151028992	1	GPX3	185,0690729	-0,799588196	0,022199813
ENSG00000175294	11	66016752	66026517	-1	CATSPER1	304,2128436	-0,511332997	0,022199813
ENSG00000134369	1	201622885	201826969	1	NAV1	9463,496005	0,334478525	0,022199813
ENSG00000169918	15	31475398	31870789	-1	OTUD7A	10,32558492	2,841202419	0,022199813
ENSG00000197182	22	46053869	46113928	1	MIRLET7BHG	940,9648215	-0,404580724	0,022345216
ENSG00000134057	5	69167135	69178245	1	CCNB1	2999,453482	0,615519292	0,022345216
ENSG00000222022	2	237421420	237425276	-1		177,3268007	0,806392491	0,022345216
ENSG00000160145	3	124080023	124726325	1	KALRN	37,19343774	1,460362354	0,022528729
ENSG00000080493	4	71062667	71572087	1	SLC4A4	220,8015253	-1,05827644	0,023113761
ENSG00000167889	17	76868456	76950393	1	MGAT5B	15,39511178	1,913265269	0,023113761
ENSG00000163349	1	113929324	113977869	1	HIPK1	2957,766033	-0,310864583	0,023144079
ENSG00000175792	3	128064778	128153914	-1	RUVBL1	2184,027459	0,329496693	0,02381512
ENSG00000184164	22	49918167	49927540	1	CRELD2	2591,089225	0,389189324	0,023905606
ENSG00000173432	11	18266260	18269977	1	SAA1	67,12739687	-1,51322549	0,023918134

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000139970	14	59595976	59870776	-1	RTN1	18,08304653	2,28560468	0,023918134
ENSG00000102383	X	75368427	75523502	-1	ZDHHC15	41,29389648	2,371723763	0,023918134
ENSG00000153391	18	35452230	35497991	-1	INO80C	236,7908101	-0,52750076	0,024140664
ENSG00000163171	2	37641882	37738468	-1	CDC42EP3	15683,78207	-0,307298364	0,024262787
ENSG00000110104	11	60842113	60851081	1	CCDC86	964,8666016	0,490676492	0,02453116
ENSG00000158246	1	27005020	27012850	-1	TENT5B	210,2810118	0,597196167	0,024770776
ENSG00000137502	11	82973133	83071923	-1	RAB30	571,7239339	-0,501333855	0,025128163
ENSG00000143603	1	154697455	154870281	-1	KCNN3	16,0591828	-2,241352355	0,025198698
ENSG00000137501	11	85694224	85811159	-1	SYTL2	794,2802728	-0,634394778	0,025198698
ENSG00000221968	11	61873519	61892051	-1	FADS3	9202,639999	-0,375648232	0,025198698
ENSG00000162981	2	14632700	14650814	1	LRATD1	9,199505095	2,473149827	0,025271024
ENSG00000171552	20	31664452	31723989	-1	BCL2L1	6415,105339	-0,409452247	0,0263095
ENSG00000197106	1	110150494	110202202	1	SLC6A17	44,33277171	2,03129622	0,0263095
ENSG00000221866	7	132123332	132648688	-1	PLXNA4	2,546011049	5,497966513	0,0263095
ENSG00000237813	7	116238260	116499465	-1		17,63962745	-1,679833057	0,026663933
ENSG00000145390	4	119212587	119295517	1	USP53	340,2711779	-0,836839659	0,026916016
ENSG00000175274	11	44885903	44951306	-1	TP53I11	808,7351937	0,624885771	0,028170928
ENSG00000119862	2	64453969	64461381	1	LGALSL	10,22637469	2,266847803	0,028398135
ENSG00000137273	6	1389576	1395603	1	FOXF2	6,904790801	2,76201443	0,028706562
ENSG00000197180	X	154424380	154428479	-1		44,08060015	-1,013551632	0,029844001
ENSG00000084774	2	27217369	27243943	1	CAD	4335,364513	0,321449763	0,029844001
ENSG00000184838	5	120464300	120687332	1	PRR16	2337,501621	-0,340970529	0,030178584
ENSG00000186193	9	137062127	137070557	-1	SAPCD2	252,2252086	0,976635722	0,030179464
ENSG00000105383	19	51225064	51243860	1	CD33	5,506258121	4,803165601	0,030656757
ENSG00000281969	6	39818751	39823227	1		66,23610952	-0,974012503	0,030728692
ENSG00000130600	11	1995176	2001470	-1	H19	78913,62493	-0,635736698	0,030728692
ENSG00000143013	1	87328880	87348923	1	LMO4	774,8311658	0,379026465	0,030728692
ENSG00000172348	6	46220736	46491972	-1	RCAN2	48,74546571	1,220243377	0,030728692
ENSG00000164099	4	118280038	118353003	-1	PRSS12	103,9835071	1,466560499	0,031427331
ENSG00000178038	3	46668995	46693704	-1	ALS2CL	25,01140922	-1,643618959	0,032277291

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000185070	14	85530144	85654428	1	FLRT2	403,1990111	-0,680515906	0,03228218
ENSG00000277156	14	19267115	19268164	1		3,174251335	6,021346946	0,032910903
ENSG00000204264	6	32840717	32844679	-1	PSMB8	573,7040458	-0,542425617	0,032933164
ENSG00000146592	7	28299321	28825894	1	CREB5	49,25917448	1,593888897	0,032933164
ENSG00000284731	3	130002789	130003007	-1		44,27595476	-1,612620878	0,034646372
ENSG00000154736	21	26917922	26967088	-1	ADAMTS5	181,5400873	-1,462064183	0,034646372
ENSG00000161642	12	54369133	54391298	-1	ZNF385A	194,9769672	-0,511670884	0,034646372
ENSG00000083799	16	50742050	50801935	1	CYLD	1468,179089	-0,291201527	0,034646372
ENSG00000163485	1	203090654	203167405	1	ADORA1	165,7800627	-2,297445674	0,035579766
ENSG00000181722	3	114314501	115147271	-1	ZBTB20	320,9721197	-0,698767881	0,035579766
ENSG00000196730	9	87497228	87708634	1	DAPK1	1020,620144	-1,033387086	0,035888724
ENSG00000158292	1	6247353	6261098	-1	GPR153	726,5479316	-0,903182992	0,035888724
ENSG00000132965	13	30713478	30764426	1	ALOX5AP	90,49029539	0,854155137	0,036447206
ENSG00000158089	2	30910467	31155202	-1	GALNT14	4,839301024	4,106085708	0,036447206
ENSG00000138600	15	50702266	50765709	-1	SPPL2A	1513,978457	-0,262928213	0,03803244
ENSG00000213694	9	88990863	89005155	1	S1PR3	1053,326935	0,80377373	0,038548613
ENSG00000162631	1	107140007	107483458	1	NTNG1	109,673862	1,427042522	0,038548613
ENSG00000226592	7	145269514	145270384	1		4,721086935	6,378032544	0,038548613
ENSG00000168528	1	31409565	31434680	1	SERINC2	6688,754409	-0,825469646	0,038574367
ENSG00000125347	5	132481609	132490777	-1	IRF1	413,8828739	-0,499225222	0,038574367
ENSG00000021355	6	2832332	2841959	-1	SERPINB1	350,1641376	-0,584064275	0,038685878
ENSG00000153071	5	39371675	39462300	-1	DAB2	2773,870863	-0,486231375	0,038812592
ENSG00000179715	12	47079603	47236662	1	PCED1B	292,0089599	-0,396378688	0,039137795
ENSG00000109436	4	140620782	140756385	-1	TBC1D9	829,7509657	0,29070166	0,040670639
ENSG00000143322	1	179099330	179229684	-1	ABL2	4118,652464	-0,368567088	0,041751717
ENSG00000177133	1	3059615	3068437	-1	PRDM16-DT	8,552844351	2,605702328	0,04225573
ENSG00000226380	7	130876809	130913310	-1		28,55341707	-1,207160609	0,042914397
ENSG00000128284	22	36140330	36166177	-1	APOL3	44,57734063	-1,170489768	0,043503695
ENSG00000189108	X	104566315	105767829	1	IL1RAPL2	82,10763372	-0,89708886	0,044001922
ENSG00000050438	12	51391317	51515763	1	SLC4A8	251,9517022	0,803122764	0,044321431

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000177679	7	76201896	76287288	1	SRRM3	31,23196035	1,580110047	0,044641226
ENSG00000167508	16	88651935	88663161	-1	MVD	3041,36659	-0,45181474	0,045289083
ENSG00000141232	17	50862223	50867978	-1	TOB1	880,42068	-0,421060632	0,046502288
ENSG00000237807	8	53493523	53524336	-1		152,2930638	-0,608902442	0,046546119
ENSG00000121931	1	110947190	110963965	-1	LRIF1	316,103634	0,778898039	0,046546119
ENSG00000127252	3	193241125	193277738	1	PLAAT1	4,078019001	3,844369681	0,046546119
ENSG00000160255	21	44885953	44931989	-1	ITGB2	1146,087023	-0,74290305	0,047076119
ENSG00000136295	7	2631986	2664802	1	TTYH3	6106,302434	0,26980754	0,047076119
ENSG00000175920	4	3463306	3501473	1	DOK7	411,6235509	-2,288381379	0,047629005
ENSG00000137727	11	110577042	110713189	-1	ARHGAP20	3,296295863	5,832790518	0,047629005
ENSG00000174738	3	23945286	23980617	1	NR1D2	479,5372326	-0,505930318	0,047769125
ENSG00000133321	11	63536808	63546462	1	PLAAT4	31,77319646	-1,341823111	0,04784121
ENSG00000188641	1	97077743	97995000	-1	DPYD	913,2689045	-0,47208081	0,049131113
ENSG00000151876	5	41925254	41941743	1	FBXO4	359,378805	-0,426650274	0,049131113

Suppl. Table 2. Differentially expressed genes upon LRIF1L^{KO} in 1926iMB.

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000125657	19	6531026	6535924	1	TNFSF9	74,4144078	1,081349517	0,026966033
ENSG00000235750	1	175156986	175192999	-1	KIAA0040	80,1925354	2,52029463	0,026966033
ENSG00000100593	14	77474394	77498816	-1	ISM2	18,8067049	4,09042583	0,026557188
ENSG00000081760	12	125065434	125143333	1	AACS	1448,907478	0,411842874	0,025813372
ENSG00000165194	X	100291644	100410273	-1	PCDH19	107,10547	3,32357456	0,025813372
ENSG00000155816	1	240014348	240475187	1	FMN2	736,055994	4,449396622	0,008434473
ENSG00000131398	19	50311937	50333515	-1	KCNC3	53,96334835	5,01674292	0,008434473
ENSG00000105472	19	50723364	50725718	1	CLEC11A	258,7588469	4,337706657	0,00634564
ENSG00000215105	X	75740831	75746911	-1	TTC3P1	285,3096611	1,025022965	0,00067202
ENSG00000135750	1	233614004	233672512	1	KCNK1	26,03652631	6,867499109	0,00067202
ENSG00000093072	22	17178790	17258235	-1	ADA2	97,7754348	-3,071026749	1,33614E-07

Suppl. Table 3. Differentially expressed genes upon LRIF1L+S^{KO} in 1926iMB.

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000244306	14	19268853	19337730	-1		64,05503278	7,421851595	0,003818178
ENSG00000049540	7	74027789	74069907	1	ELN	6286,324051	6,464142439	7,9065E-05
ENSG00000138650	4	133149294	133208606	1	PCDH10	1405,670404	5,503026441	0,003231218
ENSG00000135750	1	233614004	233672512	1	KCNK1	26,03652631	5,364339231	0,014060871
ENSG00000105472	19	50723364	50725718	1	CLEC11A	258,7588469	4,628334817	0,000400145
ENSG00000131398	19	50311937	50333515	-1	KCNC3	53,96334835	4,368248633	0,000514778
ENSG00000155816	1	240014348	240475187	1	FMN2	736,055994	4,192104541	0,002719559
ENSG00000250312	4	124501	202303	1	ZNF718	33,03018702	3,891990163	7,9065E-05
ENSG00000165194	X	100291644	100410273	-1	PCDH19	107,10547	3,313602309	0,005426219
ENSG00000254535	4	134196333	134201789	-1	PABPC4L	46,96301519	2,769645614	0,000714102
ENSG00000151692	2	6917412	7068286	1	RNF144A	17,32295849	2,760371119	0,04168387
ENSG00000105048	19	55132698	55149206	-1	TNNT1	1586,713656	2,423922212	0,012320574
ENSG00000102053	X	65366638	65507887	1	ZC3H12B	33,47085801	2,125986977	0,010167363
ENSG00000142611	1	3069168	3438621	1	PRDM16	73,83955733	2,102478318	0,002244177
ENSG00000160588	11	118226690	118252365	-1	MPZL3	29,51893106	1,847651557	0,013714524
ENSG00000176490	19	2714567	2721372	-1	DIRAS1	79,37025803	1,730804983	9,86202E-05
ENSG00000115896	2	197804593	198572581	1	PLCL1	19,54316259	1,659993481	0,036377607
ENSG00000131037	19	55072020	55087923	1	EPS8L1	31,49987618	1,571117004	0,012493192
ENSG00000128965	15	40952962	40956512	1	CHAC1	320,5982669	1,536231521	0,00356336
ENSG00000177614	1	230314490	230426332	-1	PGBD5	142,5722663	1,455419262	0,033313213
ENSG00000124613	6	27374615	27403904	1	ZNF391	42,65278443	1,435236361	0,007332035
ENSG00000269706	19	49050217	49050846	-1		43,38791936	1,428107204	0,045752734
ENSG00000133216	1	22710839	22921500	1	EPHB2	158,1940262	1,321653001	0,040766593
ENSG00000120594	10	19816239	20289856	1	PLXDC2	317,6554282	1,262683419	0,046557051
ENSG00000006432	14	70722526	70809534	-1	MAP3K9	241,9562663	1,232585509	0,007021285
ENSG00000160200	21	43053191	43076943	-1	CBS	295,2310397	1,230714656	0,014060871
ENSG00000107562	10	44370165	44386493	-1	CXCL12	716,6752537	1,132374626	0,017216183
ENSG00000107731	10	71212570	71302864	1	UNC5B	739,1442003	1,103498205	0,005426219
ENSG00000215105	X	75740831	75746911	-1	TTC3P1	285,3096611	1,069564538	7,9065E-05

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000136010	12	105019784	105107643	-1	ALDH1L2	1778,538842	1,047238152	0,008515968
ENSG00000075643	18	36187497	36272157	1	MOCOS	609,2309164	0,905723806	0,013714524
ENSG00000103489	16	17101769	17470881	-1	XYLT1	1840,556503	0,844372887	0,014060871
ENSG00000127125	1	42456117	42473385	1	PPCS	634,7119207	0,832149615	0,010463973
ENSG00000115902	2	64988477	65023865	1	SLC1A4	1725,406463	0,829296551	0,007438333
ENSG00000178814	8	144051266	144063965	-1	OPLAH	221,7255457	0,814025343	0,033115548
ENSG00000173546	15	75674322	75712848	-1	CSPG4	3417,365589	0,797768319	0,006921527
ENSG00000161653	17	44004546	44009063	1	NAGS	383,0744498	0,74057004	0,046099917
ENSG00000064655	20	46894624	47188844	1	EYA2	270,4645254	0,691171176	0,017563462
ENSG00000109107	17	28573115	28576948	-1	ALDOC	1621,567368	0,671281364	0,005426219
ENSG00000154175	3	100749156	100993515	-1	ABI3BP	7210,206635	0,637482671	0,016303943
ENSG00000166123	16	46884362	46931289	1	GPT2	1981,903682	0,630337837	0,014060871
ENSG00000177374	17	2054154	2063241	1	HIC1	726,6684234	0,626802708	0,009932035
ENSG00000100234	22	32801701	32863043	1	TIMP3	4935,413685	0,618442216	0,033313213
ENSG00000105281	19	46774883	46788594	-1	SLC1A5	5982,352046	0,602398046	0,03299471
ENSG00000183010	17	81932384	81942412	-1	PYCR1	3894,292316	0,591038828	0,004415495
ENSG00000140105	14	100333790	100376805	-1	WARS	3087,414681	0,572086052	0,033313213
ENSG00000109846	11	111908619	111923722	-1	CRYAB	2381,877159	0,563526191	0,03299471
ENSG00000140044	14	75427716	75474111	1	JDP2	1030,080076	0,540825872	0,023148724
ENSG00000149257	11	75562056	75572783	1	SERPINH1	29928,90969	0,511215375	0,000153099
ENSG00000004399	3	129555175	129606818	-1	PLXND1	9012,682468	0,483161061	0,045988009
ENSG00000090861	16	70252295	70289543	-1	AARS	7746,116445	0,428342451	0,008515968
ENSG00000113657	5	147390808	147510068	-1	DPYSL3	2565,074735	0,425519633	0,005062084
ENSG00000117385	1	42746335	42767084	-1	P3H1	6692,361951	0,386117046	0,019154907
ENSG00000153815	16	81445170	81711762	1	CMIP	1563,668258	0,378944167	0,013623288
ENSG00000135596	6	109444062	109465968	-1	MICAL1	7413,291499	0,347605646	0,010842502
ENSG00000196924	X	154348524	154374638	-1	FLNA	205505,4418	0,321374957	0,003363165
ENSG00000109089	17	74987632	75005800	1	CDR2L	2027,442474	0,302456199	0,033845665
ENSG00000089248	12	112013348	112023449	1	ERP29	2651,254486	0,268999401	0,023540323
ENSG00000112096	6	159669069	159745186	-1	SOD2	1487,346244	-0,37135412	0,036071758

ensembl_gene_id	chr	start	end	strand	hgnc_symbol	baseMean	log2FoldChange	padj
ENSG00000000971	1	196651878	196747504	1	CFH	2049,525726	-0,402990621	0,033115548
ENSG00000196776	3	108043091	108091862	-1	CD47	1300,945727	-0,406920582	0,002604437
ENSG00000104368	8	42175233	42207724	-1	PLAT	22657,50421	-0,422005345	0,03943701
ENSG00000106355	7	32485338	32495283	-1	LSM5	647,0384688	-0,428519065	0,036071758
ENSG00000188342	13	45120510	45284893	1	GTF2F2	1708,059089	-0,463624672	0,008515968
ENSG00000115525	2	85837120	85905199	-1	ST3GAL5	529,368235	-0,520664027	0,03299471
ENSG00000155099	8	90993802	91040872	-1	PIP4P2	279,3604339	-0,533601313	0,024208816
ENSG00000172927	11	69294151	69367726	1	MYEOV	881,3784439	-0,604145349	0,046557051
ENSG00000162645	1	89106132	89150456	-1	GBP2	313,0588505	-0,612551295	0,033115548
ENSG00000154188	8	107249482	107498055	-1	ANGPT1	867,9421459	-0,6398038	0,010463973
ENSG00000169908	3	149369022	149377692	-1	TM4SF1	2629,146715	-0,986605179	0,003114631
ENSG00000156103	8	88032011	88328025	-1	MMP16	119,9060171	-1,174852982	0,018408933
ENSG00000188906	12	40196744	40369285	1	LRRK2	60,83423199	-1,191448792	0,01945222
ENSG00000173801	17	41754604	41786931	-1	JUP	91,86027144	-1,334919978	0,023132633
ENSG00000175879	2	176129694	176132695	1	HOXD8	186,5258921	-1,562190625	0,002398265
ENSG00000095303	9	122370530	122395703	1	PTGS1	742,263139	-1,649661703	0,033845665
ENSG00000149131	11	57597387	57614853	1	SERPING1	522,0542459	-2,347339672	1,09938E-10
ENSG00000041353	18	54717860	54895516	1	RAB27B	66,35166965	-2,478540487	0,020432508
ENSG00000128709	2	176122719	176124937	1	HOXD9	21,91260757	-3,07049735	0,04061903
ENSG00000007933	1	171090901	171117819	1	FMO3	30,22774261	-3,839629436	0,000153099

Suppl. Table 4. Additional information about cell lines used in this study. The percentage value in the Remarks columns refers to 4q+10q D4Z4 methylation expressed as the Delta1 score rounded to the nearest integer. The Delta1 score is determined by measuring methylation-sensitive FseI digestion efficiency with Southern blotting and adjusting the observed methylation for the cumulative size of both 4q and 10q repeats. For orientation, the average Delta1 in control individuals is almost 0, whereas for *SMCHD1* mutation carriers it is −31.3% as previously reported by Lemmers et al [1].

Cell Line ID	Cell Type	Clinical Status	Male (M) or Female (F)	Primary (P) or Immortalized (I)	Remarks
1926	myoblast	healthy control	F	I	-4%
54-1	myoblast	healthy control	M	I	N.D.%
2445	myoblast	FSHD2 (*SMCHD1* c.4347-236A>G)	M	I	-29%
Rf285.3	myoblast	ICF1 (*DNMT3B*, c.2421-11G>A, c.2421-11G>A)	M	I	-42%, 11U 4qA161S
2524	fibroblast	healthy control	F	P	19%
2397	fibroblast	healthy control	M	P	12%
2333	fibroblast	healthy control	M	P	4%
2440	fibroblast	FSHD2 (*SMCHD1* c.1302_1306delTGATA)	F	P	-25%
2332	fibroblast	FSHD2 (*SMCHD1* c.3274_3276+1del)	M	P	-36%
2337	fibroblast	FSHD2 (*SMCHD1* 1.2 Mb deletion)	F	P	-29%
Rf732 (v294)	fibroblast	FSHD2 (*DNMT3B* c.1579T>C)	M	P	-21%
Rf210 (b974)	fibroblast	Healthy control (*DNMT3B* c.2072C>T)	M	P	-29%
Rf699.3	fibroblast	ICF1 (*DNMT3B* c.1918G>C, c.1918G>C)	F	P	-46%
Rf286.3	fibroblast	ICF1 (*DNMT3B* c.2177T>G, c.1918G>C)	M	P	-37%
2374	fibroblast	Healthy control	F	P	-1%
2417	fibroblast	Healthy control	F	P	-7%
GM08714	fibroblast	ICF1 (*DNMT3B* c.1807G>A, c.2232-11G>A)	F	P	-34%
Rf614	fibroblast	ICF1 (*DNMT3B* c.2292G>T, c.2342_2343del)	F	P	-39%
HCT116	colon carcinoma	NA	M	NA	57%, First described here [2]
HCT116 DKO	colon carcinoma	Double knock-out of *DNMT3B* and *DNMT1*	M	NA	-37%, First described here [2]

Suppl. Table 5. Oligonucleotides used for sgRNAs cloning into pX458 vector. The extra G (underlined) was added upstream of the sgRNA sequence if the sgRNA sequence itself did not start with one to ensure transcription from the U6 promoter. The sequence specific to the targeted DNA region is in bold.

Name	5'->3'
SMCHD1 ex3 sgRNA3 F	CACCG**ACTGATTGACCGACTGTAGC**
SMCHD1 ex3 sgRNA3 R	AAAC**GCTACAGTCGGTCAATCAGT**C
LRIF1 ex2 gRNA948 F	CACCG**TCGCGTCCCACTAGGATCGA**
LRIF1 ex2 gRNA948 R	AAAC**TCGATCCTAGTGGGACGCGA**C
LRIF1 ex3 gRNA153 F	CACCG**AATGGTCAGGAATTCGAGTA**
LRIF1 ex3 gRNA153 R	AAAC**TACTCGAATTCCTGACCATT**C

Suppl. Table 6. Primers used for RT-qPCR analyses. All primer pairs were used at Tm = 60°C. *GUSB* was used as a house-keeping gene.

Name	5'->3'
GUSB F	CTCATTTGGAATTTTGCCGATT
GUSB R	CCGAGTGAAGATCCCCTTTTTA
Dux4RT F2	CCCAGGTACCAGCAGACC
pLAM R4	TCCAGGAGATGTAACTCTAATCCA
hMYH3 F	TGATCGTGAAAACCAGTCCATTCT
hMYH3 R	TTGGCCAGGTCCCCAGTAGCT
TRIM43 F	ACCCATCACTGGACTGGTGT
TRIM43 R	CACATCCTCAAAGAGCCTGA
ZSCAN4 F	TGGAAATCAAGTGGCAAAAA
ZSCAN4 R	CTGCATGTGGACGTGGAC
MBD3L2 F	GCGTTCACCTCTTTTCCAAG
MBD3L2 R	GCCATGTGGATTTCTCGTTT
KHDC1L F	TGAATCAGGTGGGAGCACAG
KHDC1L R	CAATGCAGCGAAGGTACGTG
SMCHD1 ex47 F	CGACAGATTGTCCAGTTCCTC
SMCHD1 ex48 R	CCAATGGCCTCTTCTCTCTG
LRIF1 ex2/3 F	GTGTCCTCCAGAGCATAGAG
LRIF1 ex2/3 R	GCCATCTCATTATGGATCTTTGG
LRIF1 ex1/3 F	TCGCGTTGATCCATAATGAG
LRIF1 ex1/3 R	CACTCTTCAGATGTAATGCCT
LRIF1 ex3/4 F	GTTTATGGTGAAGGAAGGAGAG
LRIF1 ex3/4 R	ACCGGTGACATTAGCTTCC

Suppl. Table 7. Primers used for ChIP-qPCR analyses.

Name	5'->3'	Note (Tm, reference)
DR ChIP F2	GGCAGGGAGGAAAAGCGGTCC	60°C, this paper
DR ChIP R2	CTGTGAACCGCGCGGGTGAAG	
Q ChIP F	CCGCGTCCGTCCGTGAAA	65°C, [3]
Q ChIP R	TCCGTCGCCGTCCTCGTC	
Hox ChIP F	CGAGGACGGCGACGGAGAC	58°C, [3]
Hox ChIP R	ACCCTGTCCCGGGTGCCTG	
LRIF1 prom 679 F	AAGGTGACTGGCTCGCTAAA	60°C, this paper
LRIF1 prom 830 R	TTTATGATTGACCCCGGAAA	
GAPDH prom F	CTGAGCAGTCCGGTGTCACTAC	60°C, this paper
GAPDH prom F	GAGGACTTTGGGAACGACTGA	

Supplementary References

1. Lemmers, R. J. L. F. *et al.* Inter-individual differences in CpG methylation at D4Z4 correlate with clinical variability in FSHD1 and FSHD2. *Hum. Mol. Genet.* **24**, 659–669 (2015).
2. Rhee, I. *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**, 552–556 (2002).
3. Zeng, W. *et al.* Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. *Hum. Mutat.* **35**, 998–1010 (2014).

CHAPTER 5

Lrif1 is required for Trim28-mediated repression of *Dux* in mouse embryonic stem cells

Darina Šikrová[1], Román González-Prieto[2], Alfred C. O. Vertegaal[2], Judit Balog[1], Lucia Daxinger[1] and Silvère M. van der Maarel[1]

[1]Department of Human Genetics, Leiden University Medical Center, 2333ZC Leiden, The Netherlands
[2]Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands

Manuscript in preparation

Abstract

Germline mutations in *SMCHD1*, *DNMT3B* and *LRIF1* can cause Facioscapulohumeral muscular dystrophy type 2 (FSHD2). FSHD is an epigenetic skeletal muscle disorder in which incomplete heterochromatinization of the D4Z4 macrosatellite repeat causes spurious expression of the repeat-embedded *DUX4* gene in skeletal muscle, ultimately leading to muscle weakness and wasting. All three proteins play a role in chromatin organization and gene silencing, however, a potential direct functional interplay has not been elucidated yet. Here, we show that siRNA-mediated depletion of Lrif1, but not of the other two FSHD2 genes, in mouse embryonic stem cells leads to upregulation of the 2-cell cleavage stage transcriptional program driven by the transcription factor Dux, which is the mouse functional homologue of human DUX4. Furthermore, we show that Lrif1 interacts with Trim28, a known *Dux* repressor, and that this interaction is independent of Cbx proteins and Smchd1. We uncover that *Dux* upregulation in Lrif1 knock-down mESCs is due to decreased Trim28 occupancy at the *Dux* locus itself. Together, our results provide evidence for a conserved function of Lrif1 in repressing the expression of an early zygotic genome activator both in mouse and human.

Main

To test the potential functional cooperation among FSHD2 gene products, we made use of serum/LIF-cultured E14 mESCs in which the expression of all three FSHD2 genes physiologically coincides, and performed siRNA-mediated knock down of each disease gene product followed by total RNA-seq. In contrast to the *Smchd1* gene, for which only one mRNA isoform is expressed in this culture system, *Dnmt3b* and *Lrif1* genes give rise to at least three different protein-coding isoforms in E14 mESCs (Figure 1A). Interestingly, only two protein-coding isoforms are annotated for human *LRIF1* (referred to as long and short) whereas in mouse a third isoform is produced by an alternative upstream transcriptional start site (Suppl. Figure 1A). This isoform contains an N-terminal extension of 16 aa to the short Lrif1 isoform (Figure 1A). Thus, to ensure targeting of all isoforms, we used a mix of four siRNAs for each gene. Cells were harvested after two consecutive two days-long knock-downs, which resulted in efficient protein (Figure 1A) and mRNA (Figure 1B) depletion, while mRNA and protein levels of the untargeted FSHD2 genes remained unaffected (Figure 1A and 1B). The mRNA levels of three tested pluripotency markers (*Oct4*, *Nanog* and *Sox2*) were largely unaffected upon respective knock-downs, however, we detected mild but significant downregulation of *Sox2* mRNA levels upon Lrif1 knock-down (Figure 1C). Together, this suggests that short-term depletion of the three FSHD2 disease proteins does not impair pluripotency in serum/LIF grown mESCs.

To better understand the roles of these factors in gene regulation, we first performed total RNA-seq. This revealed only subtle gene expression changes in the knock-down conditions (when considering 2-fold expression changes in either direction) as compared to the control condition. The Lrif1 knock-down condition showed the highest number of differentially expressed genes (749 DEGs with p.adj. <0.05), the majority of which were of modest fold changes (Figure 1D). Next, we assessed whether differentially expressed genes were shared between the different knock downs. Despite statistically significant overlaps between differentially upregulated and downregulated genes in paired comparisons of the knock-down conditions, only a limited number of upregulated and downregulated genes (23 and 9, respectively) was common among all three knock-down conditions (Figure 1G and 1H, Suppl. Table 1). This limited overlap prohibits pathway analysis and suggests rather divergent effects of these proteins on the transcriptome in mESCs.

Interestingly, the top differentially upregulated genes in Lrif1 knock-down mESCs (such as the *Zscan4* cluster genes) belong to a class of genes that is specifically expressed at the two cell (2C) cleavage stage of the mouse embryo[1–4] (Figure 1D). The 2C-like cells spontaneously arise in mESC culture accounting for less than 1% of the population[3] and they mimic some of the distinctive features of the 2C-stage embryos (reviewed here[5]). Furthermore, expression analysis of repetitive elements in Lrif1 knock-down mESCs showed a significant increase in transcripts originating from repeats, which are known to be de-repressed in the 2-cell embryo as well as in the 2C-like mESCs population, such as major satellites and MERVL elements (Suppl. Figure 1B).

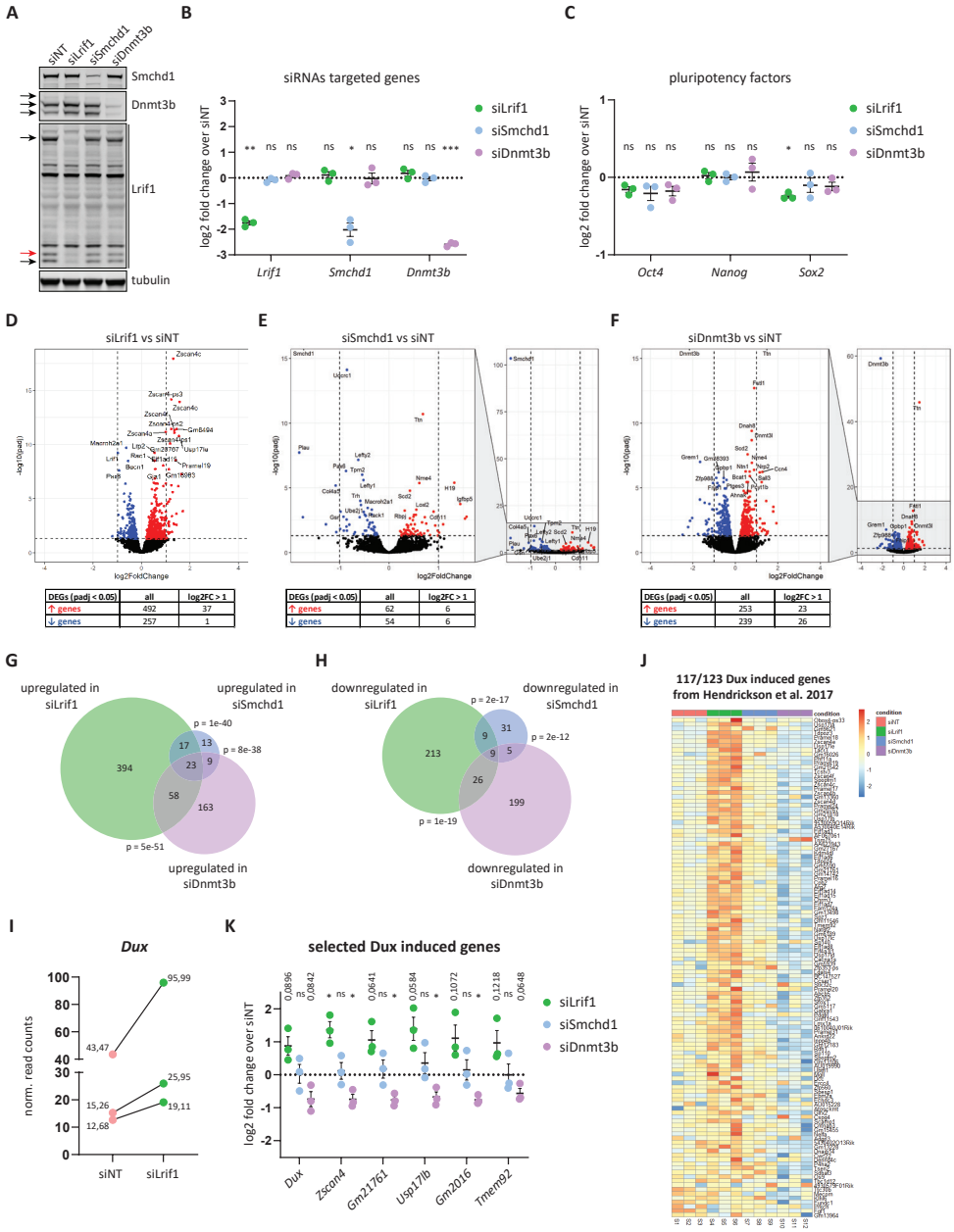


Figure 1. Lrif1 knock-down causes upregulation of Dux sensitive genes. A) Western blot confirmation of successful siRNA-mediated knock-down of three FSHD2 genes (Lrif1, Smchd1 and Dnmt3b) in E14 mESCs. Arrows mark different protein isoforms of Dnmt3b and Lrif1. The red arrow marks extra mouse-specific Lrif1 isoform not identified in human. Tubulin served as loading control. **B)** RT-qPCR confirmation of downregulation of three FSHD2 genes after siRNA-mediated knock-down in E14 mESCs. Expression levels detected in knock-down conditions were normalized to the siNT condition and log2 transformed. Every dot represents an independent biological replicate. Whiskers represent mean ± SEM. Statistical significance was calculate by one-sample t-test (ns: not significant, *:

< 0.05, **: < 0.01, ***: < 0.001). **C)** RT-qPCR of three pluripotency genes (*Oct4*, *Nanog* and *Sox2*) after individual siRNA-mediated knock-down of three FSHD2 genes. Expression levels in knock-down conditions were normalized to the siNT condition and log2 transformed. Every dot represents an independent biological replicate. Whiskers represent mean ± SEM. Statistical significance was calculate by one-sample t-test (ns: not significant, *: < 0.05). **D)** Volcano plot showing gene expression changes following Lrif1 knock-down. Upregulated genes are highlighted in red and downregulated genes are highlighted in blue. Dashed lines indicate a fold change of two (log_2 fold of 1) on the x axis and significance of 0.05 (−log10 p.adj of 1.3) on the y axis. Top 20 differentially expressed genes (DEGs) are labelled. Table summary of DEGs is provided below the plot. **E)** Volcano plot showing gene expression changes following Smchd1 knock-down. Upregulated genes are highlighted in red and downregulated genes are highlighted in blue. Dashed lines indicate a fold change of two (log_2 fold of 1) on the x axis and significance of 0.05 (−log10 p.adj of 1.3) on the y axis. Top 20 differentially expressed genes (DEGs) are labelled. Table summary of DEGs is provided below the plot. **F)** Volcano plot showing gene expression changes following Dnmt3b knock-down. Upregulated genes are highlighted in red and downregulated genes are highlighted in blue. Dashed lines indicate a fold change of two (log_2 fold of 1) on the x axis and significance of 0.05 (−log10 p.adj of 1.3) on the y axis. Top 20 differentially expressed genes (DEGs) are labelled. Table summary of DEGs is provided below the plot. **G)** Overlap of common significantly upregulated genes (p.adj. < 0.05) in all three knock-down conditions. The significance of overlaps was calculated with Fischer's exact test. **H)** Overlap of common significantly downregulated genes (p.adj. < 0.05) in all three knock-down conditions. The significance of overlaps was calculated with Fischer's exact test. **I)** Changes in normalized read counts of Dux transcripts after Lrif1 knock down compared to non-targeting siRNA condition. **J)** Heatmap depicting expression changes of previously reported 117 Dux-induced genes (out of 123 reported by Hendrickson et al.) in different knock-down conditions for which there was a non-zero reads count. **K)** RT-qPCR of Dux and five Dux-sensitive genes after siRNA-mediated knock-down in E14 mESCs. Expression levels detected in knock-down conditions were normalized to siNT condition and log2 transformed. Every dot represents an independent biological replicate. Whiskers represent mean ± SEM. Statistical significance was calculate by one-sample t-test (ns: not significant, *: < 0.05).

Expression of many 2C-like genes and repeats is known to be driven by the transcription factor Dux, which is the functional homologue of primate DUX4. Initially, the *Dux* gene itself was not identified as significantly differentially upregulated in Lrif1 knock-down mESCs. However, plotting the normalized read counts of *Dux* in each Lrif1 knock-down experiment individually showed a modest increase in read numbers in each knock-down experiment (Figure 1I), whereas *Dux* normalized read counts did not increase in the other two knock-down conditions (Suppl. Figure 1C and 1D). Therefore, we closely examined the expression levels of selected genes previously described by Hendrickson et al.[6], which are sensitive to Dux overexpression in mESCs (hereafter referred to as Dux signature genes). This inspection revealed that, in general, the mRNA levels of Dux signature genes increased upon Lrif1 knock-down (Figure 1J). In contrast, Dux signature genes remained unchanged in the Smchd1 knock-down situation (Figure 1J). In addition, Dnmt3b knock-down seemed to result in decreased expression of *Dux* and Dux signature genes (Suppl. Figure 1D and Figure 1J). We validated with RT-qPCR the mRNA expression of five selected Dux signature genes (*Zscan4*, *Gm21761*, *Usp17lb*, *Gm2016*, *Tmem92*) and *Dux* itself, and confirmed their upregulation in Lrif1 knock-down mESCs, albeit not always reaching statistical significance (Figure 1K). We further confirmed reduced and unaffected mRNA levels of these genes in Dnmt3b and Smchd1 knock-down mESCs, respectively (Figure 1K). Consistent with the RNA-seq, the expression changes were subtle. Therefore, these results indicate that Lrif1 confers a mild repression of the Dux driven 2C-like transcriptional program in mESCs under these experimental conditions.

Initially, treatment of mESCs with trichostatin A, an inhibitor of histone deacetylases, was shown to promote the emergence of 2C-like cells in mESC culture, which suggested that the chromatin configuration plays a role in this cellular transition[3]. Several chromatin-regulating factors have since been reported to directly influence *Dux* expression in mESCs [7–11]. We decided to investigate the protein interactome of Lrif1 in mESCs with the aim to identify potential interactors that could explain Lrif1's contribution to regulation of 2C-like cells. To this end, we generated constructs encoding two GFP-tagged Lrif1 isoforms which correspond to the amino acid sequences of the human long and short LRIF1 isoforms. Upon transient transfection of these plasmids in E14 mESCs, we performed GFP-specific pull-down followed by mass spectrometry (MS). The MS analysis identified 54 proteins enriched in GFP-Lrif1s pull-down of which 37 were nuclear (Figure 2A) and 94 proteins enriched in GFP-Lrif1l pull-down of which 44 were nuclear (Figure 2B). The full spectrum of results can be viewed in Suppl. Table 2, however, we focused our further analysis only on nuclear proteins since Lrif1 mainly localizes to this cellular compartment[12].

We found that nuclear proteins enriched in the GFP-Lrif1s pull-down largely overlapped with proteins identified in the GFP-Lrif1l pull-down (Figure 2C). Smchd1 and three Cbx paralogues were among the top four interactors, which is consistent with previous findings[13–15]. We identified Trim28 (tripartite motif-containing protein 28; also known as Kap1) as a common interacting partner of both Lrif1 isoforms Trim28 has previously been shown to be involved in direct repression of the *Dux* locus in mESCs[7,10]. In addition, several studies showed that depletion of Trim28 in mESCs cells leads to an increase in the 2C-like population in Dux-dependent manner[3,7,10]. To address a putative cooperation between Lrif1 and Trim28, we first confirmed the Trim28 interaction with both Lrif1 isoforms by repeating the transfection of GFP-tagged Lrif1 isoforms in mESCs followed by GFP-specific pull down and western blot analysis (Figure 2D). We further validated this interaction by performing reciprocal endogenous Co-IPs from mESC whole cell extract treated with benzonase to rule out possible DNA-mediated interactions using two different Trim28 antibodies and one Lrif1 antibody (Figure 2E and 2F). Detection of endogenous Lrif1s was prohibited by its co-migration in the gel with the antibody light chain that was used for Co-IP which was of the same species origin as the primary antibody used for Lrif1 detection.

Interestingly, we could also pull-down Smchd1 (Figure 2E and 2F) and Cbx3 (Figure 2F) in the Trim28 Co-IPs. Similar to Lrif1, Trim28 contains a conserved Cbx binding motif (PxVxL; where x represents any amino acid), which is essential for transcriptional silencing imposed by Trim28[16]. We speculated that the interaction between Lrif1 and Trim28 might therefore be bridged via Cbx proteins, which are the homologues of human HP1 proteins[17]. To test this hypothesis, we introduced previously characterized mutations (mPVL) in the HP1 binding motif of the human LRIF1 long isoform[15] to our GFP-tagged long and short mouse Lrif1 constructs (Figure 2G). The mPVL mutant carries two amino acid substitutions (V47D/L51E in the short isoform; V567D/L569E in the long isoform) in the conserved HP1 binding motif which abolish the interaction of LRIF1 with the chromoshadow domain of human HP1 proteins. We included in this experiment an additional previously characterized LRIF1 mutant termed m1[15].

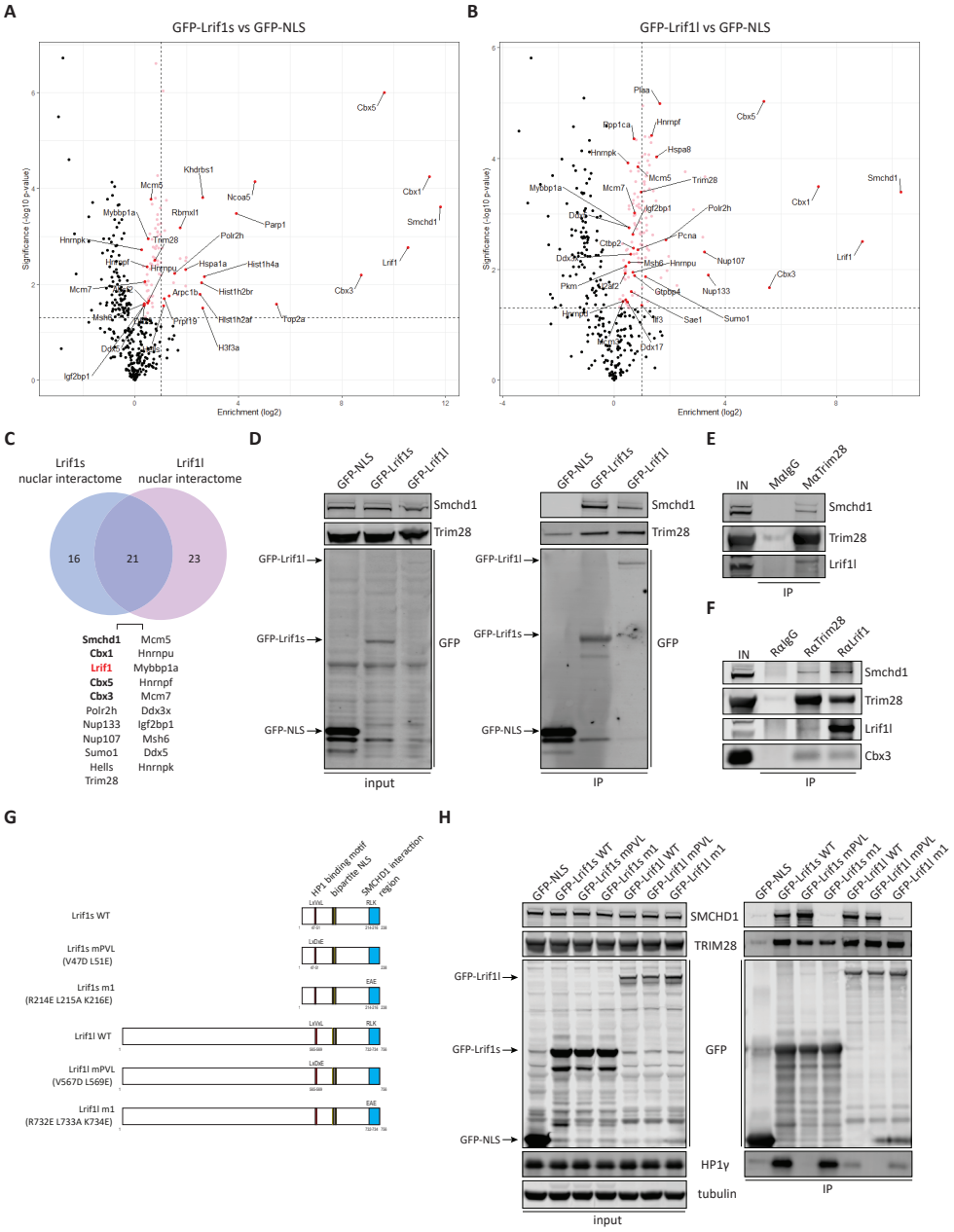


Figure 2. Lrif1 isoforms interact with Trim28. A) Volcano plot showing the differential interactome of the GFP-tagged Lrif1 short isoform over GFP-NLS as identified by GFP pull-down followed by label-free MS. GFP pull-downs were performed in biological triplicate. The enrichment (log2) is plotted on the x axis and the significance (−log10 p-value) is plotted on the y axis. The dashed line indicates a significance of 0.05 (−log10 P value of 1.3) on the y axis. Red dots mark significantly enriched nuclear proteins and pink dots mark significantly enriched non-nuclear proteins. Ribosomal proteins are not shown. All significantly enriched nuclear proteins are labelled. **B)** Volcano plot showing the differential interactome of the GFP-tagged Lrif1 long isoform over GFP-NLS. The same description applies for this plot

as for the plot in A). **C)** Venn diagram of overlapping significantly enriched nuclear proteins between GFP-tagged Lrif1 short and long isoform interactomes. **D)** Western blot confirmation of the Lrif1-Trim28 interaction by GFP pull-down of GFP-NLS, GFP-Lrif1s or GFP-Lrif1l. **E)** Endogenous MαTrim28 Co-IP on benzonase treated mESC whole cell extract. MαIgG was used as a negative control. Only the long isoform of Lrif1 is probed for as the short Lrif1 isoform protein migrates at the hight of IgG light chain. **F)** Reciprocal endogenous RαTrim28 and RαLrif1 Co-IP on benzonase treated mESC whole cell extract. MαIgG was used as a negative control. Only the long isoform of Lrif1 is probed for as the short Lrif1 isoform protein migrates at the hight of IgG light chain. **G)** Schematic representation of WT Lrif1 isoforms and their mutant forms used for GFP Co-IPs to test for TRIM28 interaction. **H)** GFP pull-downs of GFP-NLS, GFP-tagged WT and mutant Lrif1 isoforms in benzonase treated HEK293T whole cell extracts.

This mutant carries three amino acid substitutions in the C-terminal coil-coiled domain (R214/L215A/K216E in the short isoform; R732E/L733A/K734E in the long isoform) of LRIF1, which compromises the interaction with SMCHD1[15]. Since the transfection efficiency of mESCs was suboptimal for these CoIPs and there was a substantial background of Trim28 enrichment in the GFP only pull down from mESCs whole cell lysates, although with less signal than with the GFP-Lrif1 fusion proteins (Figure 2D), we performed the experiment in HEK293T cells. As anticipated due to functional motif conservation[15], both mouse Lrif1 isoforms interacted with the human SMCHD1 and CBX3/HP1$_\gamma$ proteins as well as with human TRIM28 (Figure 2H). Next, we assessed the Trim28 interaction with the mutated forms of Lrif1. As expected, the mPVL mutant abolished the interaction of Lrif1 with human HP1$_\gamma$ which corresponds with mouse Cbx3 and the m1 mutant of Lrif1 abolished the interaction with human SMCHD1. Surprisingly, neither of the mutants affected Lrif1's interaction with TRIM28 (Figure 2H). This suggests that the interaction of Lrif1 with TRIM28 is not mediated via HP1 proteins nor via SMCHD1 or its coiled-coil domain and that another region shared by both Lrif1 isoforms is responsible for this interaction.

Since Trim28 is known to repress *Dux* by directly binding to its genomic locus[7,10] and we uncovered an interaction between Lrif1 and Trim28, we were keen to investigate a potential interplay of Lrif1 and Trim28 at the *Dux* locus. First, we employed siRNA-mediated short-term depletion of either Lrif1 or Trim28 and confirmed their knock down efficiency by western blot (Figure 3A) as well as by RT-qPCR (Figure 3B). Lrif1 knock-down did not affect protein levels of Trim28 or the other two Lrif1 interacting partners (Smchd1 and Cbx3), which are also known to regulate *Dux* expression[8,18]. This result rules out the possibility that the observed increased *Dux* expression in Lrif1 knock-down is due to lower levels of any of these *Dux* repressors. A two day-long knock-down of Trim28 was already sufficient to cause mild downregulation in expression of three examined pluripotency factors (Figure 3C), which is in agreement with its essential role in pluripotency maintenance and self-renewal of mESCs cultured in serum/LIF condition[19]. Despite that, we could detect by RT-qPCR a modest upregulation of *Dux* and five of its signature genes in both knock-down situations (Figure 3D). Next, we wanted to assess if Lrif1 has a direct role in regulating the *Dux* locus by measuring Lrif1 occupancy using chromatin immunoprecipitation (ChIP). Since there is no ChIP-grade antibody for mouse Lrif1 available, we focused on a potential Lrif1-dependent Trim28 binding to the *Dux* locus. As expected, Trim28 knock-down leads to decreased Trim28 enrichment at the *Dux* locus (Figure 3E) as well as at IAPEz elements (Suppl. Figure 2A), which are also innate genomic targets of Trim28-imposed repression in mESCs[19,20].

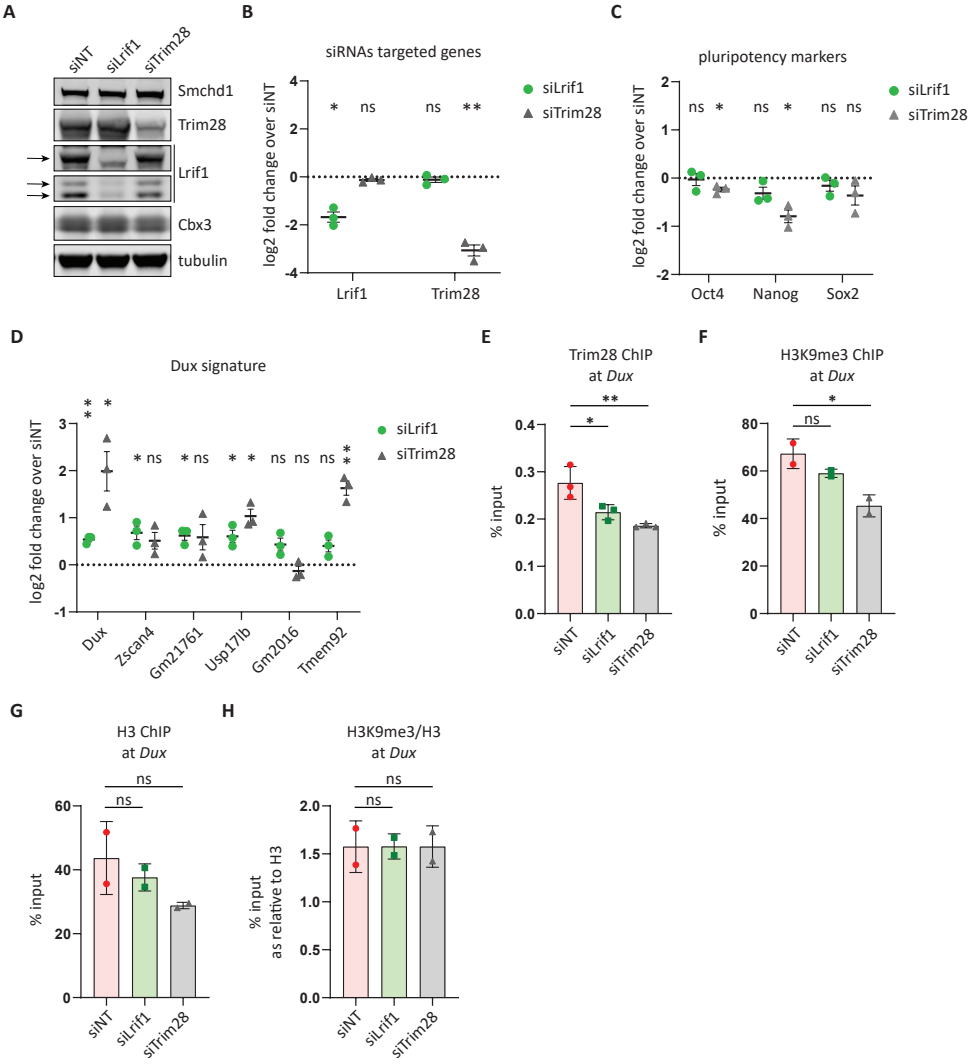


Figure 3. Trim28 requires Lrif1 for its binding to the *Dux* repeat. A) Western blot confirmation of successful siRNA-mediated knock-down of Lrif1 or Trim28 in E14 mESCs. Arrows mark different protein isoforms of Lrif1. Tubulin served as a loading control. Protein levels of two other Lrif1 interactors (Smchd1 and Cbx3) are not changed. **B)** RT-qPCR confirmation of Lrif1 and Trim28 downregulation after siRNA-mediated knock-down in E14 mESCs. Expression levels detected in knock-down conditions were normalized to the siNT condition and log2 transformed. Every dot represents an independent biological replicate. Whiskers represent mean ± SEM. Statistical significance was calculate by one-sample t-test (ns: not significant, *: < 0.05, **: < 0.01). **C)** RT-qPCR of three pluripotency genes (*Oct4*, *Nanog* and *Sox2*) after siRNA-mediated knock-down of Lrif1 or Trim28. Expression levels in knock-down conditions were normalized to the siNT condition and log2 transformed. Every dot represents an independent biological replicate. Whiskers represent mean ± SEM. Statistical significance was calculated by one-sample t-test (ns: not significant, *: < 0.05). **D)** RT-qPCR of Dux and five Dux- sensitive genes after siRNA-mediated knock-down of Lrif1 or Trim28. Expression levels in knock-down conditions were normalized to the siNT condition and log2 transformed. Every dot represents an independent biological replicate. Whiskers represent mean ± SEM. Statistical significance was calculated by one-sample t-test (ns:

not significant, *: < 0.05, **: < 0.01). **E)** Trim28 ChIP-qPCR of the 5' Dux region in E14 mESCs after treatment with respective siRNAs. Bars and whiskers represent mean ± SEM of three independent experiments. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (*: < 0.05, **: < 0.01). **F)** H3K9me3 ChIP-qPCR of the 5' Dux region in E14 mESCs after treatment with respective siRNAs. Bars and whiskers represent mean ± SEM of two independent experiments. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (ns: not significant, *: < 0.05). **G)** H3 ChIP-qPCR of the 5' Dux region in E14 mESCs after treatment with respective siRNAs. Bars and whiskers represent mean ± SEM of two independent experiments. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (ns: not significant). **H)** Ratio of H3K9me3 levels to H3 levels at Dux calculated from enrichment values presented in F) for H3K9me3 and G) for H3. Statistical significance was calculated by one-way ANOVA with Dunnett's post hoc test (ns: not significant).

The Trim28 signal at intron 2 of *Gapdh,* which served as negative control region*,* remained unchanged in both knock-down conditions thus representing only antibody background signal (Suppl. Figure 2B). Interestingly, Lrif1 knock-down itself resulted in reduced Trim28 enrichment at *Dux*, albeit to a lesser degree than observed in the Trim28 knock-down condition (Figure 3E). In contrast, the Trim28 enrichment at IAPez elements remained unaffected in Lrif1 knock-down mESCs (Suppl. Figure 2A). This is in agreement with our siLrif1 RNAseq data, where we did not detect increased expression from this class of repetitive elements (Suppl. Figure 1B). Together this points to an Lrif1-independent regulation of these repeats by Trim28.

Lastly, since the repressive histone modification H3K9me3 is a known canonical marker of Trim28-mediated repression[19], we measured its levels at the *Dux* locus to test if reduced Trim28 binding at this locus upon Lrif1 knock-down leads to a concomitant decrease of this modification. ChIP-qPCR showed that both knock-down conditions resulted in decreased H3K9me3 levels at *Dux* (Figure 3F), however, this was attributable to the lower levels of H3 itself (Figure 3G) as the ratio of the modified H3 to all H3 remained unchanged (Figure 3H). This result is suggestive of an increased chromatin accessibility at this locus and may explain the relatively subtle *Dux* expression changes upon knockdown of Lrif1 or Trim28.

Collectively, our findings identify a functional relationship between Lrif1 and Trim28 and support a conserved function of Lrif1 in the regulation of *Dux*/*DUX4* expression in mammals.

Material and Methods

Cell culture

E14 mouse embryonic stem cells (mESCs) were grown on 0.1% gelatin (Sigma, #G-1890) coated plates on an UV-irradiated feeder layer of MEFs. E14 mESCs were maintained in medium composed of KnockOut™ DMEM (Gibco, #10829018) supplemented with 10% FBS (Biowest, #S1810), 1x MEM Non-Essential Amino Acids Solution (Gibco, #11140050), 2 mM L-Glutamine (Gibco, #25030149), 1 mM Sodium Pyruvate (Gibco, #11360070), 0.1 mM 2-Mercaptoethanol (Gibco, #31350010) and 10^5 U/mL Leukemia Inhibitory Factor (EMD Millipore, #ESG1107). HEK293T cells were maintained in medium composed of Gibco DMEM, High Glucose, Pyruvate (Gibco, #119950) with addition of 10% FBS (Biowest, #S1810) and 1x Penicillin/streptomycin (Gibco, #15140122).

siRNA transfections

mESCs were reverse transfected with siGENOME siRNA SMARTpool (Horizon) at a final concentration of 40 nM using Lipofectamine™ RNAiMAX (Thermo Fisher Scientific, #13778030). siGENOME Non-Targeting Pool #2 was used as a negative control. Two days after the first transfection, cells were either harvested or 1/5th of the cells

were reverse transfected again as per the first transfection. Cells were harvested for subsequent analysis two days post-transfection.

RNA isolation
Cells were harvested in Qiazol (Qiagen, #79306) and total RNA was isolated using RNeasy mini kit (Qiagen, #74101) accompanied with on-column DNase I treatment.

cDNA synthesis followed by RT-qPCR
cDNA synthesis was performed with poly-dT primer using the RevertAid H Minus First Strand cDNA synthesis kit (Thermo Fisher Scientific, #K1621). Gene expression was analysed using iQ™ SYBR® Green Supermix (Biorad, #1708887) on a CFX384 Touch Real-Time PCR Detection System. All primers used for RT-qPCR are listed in Supplementary Table 3. mβ-actin was used as a housekeeping gene.

RNA processing for total RNA-seq
Total RNA was isolated as described above and RNA-seq was outsourced to Macrogen-Europe. Sequencing libraries were prepared with TruSeq Stranded Total RNA with Ribo-Zero Human/Mouse/Rat kit () according to the manufacturer's manual. Samples were sequenced as 100 bp paired-end on a HiSeq instrument.

RNA-seq data analysis
Quality assessment of the raw sequencing reads was done using FastQC v0.11.6. Adapters were removed by Trimmomatic v0.38 with parameters PE -threads 10 -phred33 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:40. The remaining quality-filtered reads were aligned to the mouse reference genome mm10 with the corresponding annotation file from Ensemble using the STAR aligner v2.7.1 with parameters. A read count table was obtained with HTSeq-count v0.9.1 using the GENCODE MV23 annotation with the option "–stranded reverse". The differential expression statistical analysis was done with DESeq2 v1.24.0 (R package). In the design we included correction for known experimental and sequencing batch effects (see Suppl. Table 4). Otherwise, default settings were used except for siSmchd1 analysis, where differential expression analysis was done on pre-filtered data, where 78% of the lowest expressed genes (based on row-wise mean of normalized counts) was filtered out due to an otherwise faulty outcome when independent filtering was used. The final list of differential expressed genes contains genes for which the adjusted p-value (Benjamini-Hochberg correction) is <0.05. RNA-seq plots were generated with ggplot2 package v3.3.3.

Cloning
The open reading frames (ORFs) of Lrif1s and Lrif1l were subcloned first into pEF1a-FB-dCas9-puro (Addgene plasmid #100547) by replacing the Cas9 insert via NheI/XbaI sites. The Lrif1s/l ORFs were amplified using an N-terminal specific primer for either Lrif1s (mLrif1s NheI F: 5'-TCCTTGCTAGCATGGCATCAATAGTAAAAAAGGAAATTC-3') or Lrif1l (mLrif1l NheI F: 5'-TCCTTGCTAGCATGTCTAATAGTCTCCAGAGCG-3') combined with the same reverse primer (mLrif1 XbaI R: 5'-CCTCATCTAGATTATTGTTTTTGGTACATCTTCTTACGC-3'). Resulting plasmids were further modified by replacing the N-terminal FlagAvi tag with EGFP via BstbI/NheI sites. EGFP was amplified from the pEGFP-C1 plasmid using the following primers: EGFP BstbI F: 5'-GATCTTTCGAAAGCCACCATGGTGAGCAAGGG-3' & EGFP NheI R:5'-ACATGCTAGCAAGGATCCTCGAAGCTTGAGCTCGAGATC-3'). Control plasmid expressing only EGFP-NLS was created by replacing FlagAvi-Cas9 insert in pEF1a-FB-dCas9-puro with EGFP-NLS via BstbI/XbaI sites. EGFP-NLS was amplified from the C1-EGFP-NLS plasmid (king gift of Prof. Haico van Attikum; Leiden University Medical Center) with the following primers: EGFP BstbI F: 5'- GATCTTTCGAAAGCCACCATGGTGAGCAAGGG-3' & EGFP-NLS XbaI R: 5'-CCTCATCTAGACTAAACCTTTCTCTTCTTCTTAGGACC-3'). The two C-terminal Lrif1 mutants (denoted in the text as mPVL and m1) were introduced to both the short and the long Lrif1 isoform and were created by overlapping PCR with mutagenic primers carrying the respective mutations (for primers combinations see Supplementary Table S5). The respective pEF1a-EGFP-Lrif1s/l-puro plasmids were used as a template for creating inserts 1 and 2, which were subsequently used in the final merging PCR. The final PCR products were cloned in the pEF1a-EGFP-Lrif1s/l-puro via BamHI/XbaI, thus exchanging the Lrif1s/l WT ORFs with their mutant counterpart.

Western blotting
Samples were lysed with RIPA buffer (0.1% SDS, 1% Igepal CA-630, 150mM NaCl, 0.5% Sodium Deoxycholate, 20mM EDTA) supplemented with 1x Complete™, EDTA-free Protease Inhibitor Cocktail (PI) (Sigma-Aldrich, #11873580001). Samples were incubated on ice for 10 min followed by centrifugation at 15,000g for 15 min at

4°C. Protein concentration was determined with Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, #23225). For western blotting, samples were first mixed with 6X sample buffer (0.375M Tris pH 6.8, 12% SDS, 60% glycerol, 0.6M DTT, 0.06% bromophenol blue) to 1x final concentration, boiled for 10 min at 95°C and resolved on Novex™ NuPAGE™ 4-12% Bis-Tris protein gels (Invitrogen, #NP0321BOX). Post-run gel was transferred to an Immobilon-FL PVDF membrane (Merck, #IPFL00010). The membrane was blocked for 1 h in 4% skim milk in PBS followed by incubation overnight at 4°C with primary antibodies diluted in Immuno Booster solution I (Takara, #T7111A): RαGFP (1:1000, Abcam, #ab290), RαSMCHD1 (1:1000, Abcam #ab31865), RαLRIF1 (1:1000, Proteintech, #26115-1-AP), MαKAP1 (1:1000, Abcam, #ab22553), MαHP1γ clone 42s2 (1:1000, EMD Millipore, #05-690) and Mα-αTubulin (1:4000, Sigma-Aldrich #T6199). The next day, membranes were washed twice with PBS-T (0.01% Tween-20) and incubated with the following secondary antibodies diluted in Immuno Booster solution II (Takara, #T7111A): IRDye® 800CW goat anti-rabbit IgG (1:10,000, Li-cor #P/N 925-32211) and IRDye® 680CW donkey anti-mouse IgG (1:10,000, #P/N 925-68072) for 1h at room temperature. Membranes were washed twice with PBS-T prior scanning on the Odyssey® CLx Imaging System (Li-cor).

Plasmids transfections

For co-immunoprecipitation with GFP-trap beads, $10x10^6$ mESCs were reverse transfected with 5 µg of plasmid DNA with Lipofectamine 3000 (Thermo Fisher Scientific, #L3000008) on a Ø6 cm dish with 0.1% gelatine. Each transfection condition was done in three biological replicates. 30h post-transfection, cells were harvested for downstream co-immunoprecipitation with GFP-trap beads. For GFP co-immunoprecipitation in HEK293T cells, $2.5x10^6$ cells were seeded one day prior plasmid transfection on a Ø10 cm dish. The next day, cells were transfected with 6 µg of plasmid DNA with polyethylenimine (PEI) in 1:3 volume ratio. Cells were harvested 30h post-transfection.

GFP-Trap co-immunoprecipitation for mass spectrometry

Transfected mESCs were washed 2x with ice-cold PBS and lysed on the dish with 600 µl of NP40 lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1% NP-40) supplemented with 1x PI, 20 mM NaF and 20 mM NEM. Whole cell lysates were incubated at 4°C, for 15 min while rotating, then spun down for 14,000g, 10 min at 4°C. 5% volume of supernatant was saved as input, mixed with 6xSB to a final 1X concentration and boiled for 10 min at 95°C. The remainder of the supernatant was added to 20 µl pre-washed GFP-Trap agarose beads (Chromotek, #gta-20) and incubated for 1.5 h at 4°C while rotating. Beads were subsequently washed 2x with NP40 lysis buffer followed by 3x wash with NP40 lysis buffer without NP40 and the final three washes were done with freshly prepared 50 mM ammonium bicarbonate (ABC). After the last wash, 10% of the beads was used for protein elution in 2xSB by boiling for 15 min at 95°C while shaking to check for IP efficiency by western blot. The rest of the beads were incubated overnight with 2.5 µg of sequencing grade trypsin (Promega, #V5111) (dissolved in 50 mM ABC) at 37°C while shaking. The next day, digested peptides were filtered through a pre-washed 0.45 µm filter (EMD Millipore, #UFC40LH25) followed by acidification through addition of trifluoroacetic acid (TFA) to a 2% final concentration. Peptide solutions were loaded on a custom-made Stage Tip as containing a disk made of tC18 cartridge (Waters, #WAT036820) as described previously [21]. Stage Tips were washed twice with 0.1% formic acid, and peptides were eluted with 2x 25 µl of 32.5% acetonitrile in 0.1 % formic acid. Eluates were vacuum dried with a SpeedVac RC10.10 and kept at -80°C.

Mass spectrometry data acquisition

Mass spectrometry data was acquired essentially as described in Gonzalez-Prieto et al. [22]. In brief, a Liquid Chromatography gradient was performed on an EASY-nLC 1000 system (Proxeon, Odense, Denmark) connected to a Q-Exactive Orbitrap (Thermo Fisher Scientific, Germany) through a nano-electrospray ion source. The Q-Exactive was coupled to a 20 cm analytical column with an inner-diameter of 75 µm, in-house packed with 1.9 µm C18-AQ beads (Reprospher-DE, Pur, Dr. Maish, Ammerbuch-Entringen, Germany). For each sample, two different acquisition methods were performed as technical repeats. The chromatography gradient length was 70 minutes from 2% to 30% acetonitrile in followed by 5 minutes gradient to 95% acetonitrile in 0.1% formic acid prior to column re-equilibration at a flow rate of 200 nL/minute. The mass spectrometer was operated in data-dependent acquisition (DDA) mode. The first technical repeat was performed with a top-10 method. The maximum MS1 and MS2 injection times were 250 ms and 60 ms, respectively. For the second technical repeat, a Top5 method was used with MS1 and MS2 injection times being 250 ms and 256 ms, respectively. In both technical repeats, full-scan MS spectra were acquired in a range from 300 to 1600 m/z at a target value of 3×10^6 and a resolution of 70,000 and the Higher-Collisional Dissociation (HCD) tandem mass spectra (MS/MS) were recorded at a target value of 1×10^5 with a resolution of 17,500. Minimum AGC target was set to 1×10^4 and the normalized collision energy (NCE) was set to 25. The isolation window was 2.2

m/z wide. The precursor ion masses of scanned ions were dynamically excluded (DE) from MS/MS analysis for 20 sec. Ions with charge 1, and greater than 6 were excluded from triggering MS2 analysis.

Mass spectrometry data analysis

LC-MS/MS Raw files were analyzed using MaxQuant software (v1.6.14) according to Tyanova et al.[23] using default settings with the following modifications. Maximum number of mis-cleavages by trypsin/p was set to 3. Variable modifications included Oxidation (M), Acetyl (Protein N-terminus) and Phospho (STY) with a maximum number per peptide of 3. Carbamidomethyl (C) was deactivated as fixed modification. Label-free Quantification was enabled without the Fast LFQ algorithm. We performed the search against an in silico digested UniProt reference proteome for Homo sapiens (19th Sep 2019). The match-between-runs feature was enabled with a 0.7 min match time window and 20 min alignment time window. Protein quantification included all the peptides. MaxQuant proteingroups.txt file output was further analyzed using the Perseus computational platform (v1.6.14) as described by Tyanova et al.[23]. Potential contaminants and reverse peptides were removed, the matrix was log2 transformed and proteins not identified in 3 out of 3 replicates for at least one condition were also removed. Missing values were randomly imputed from normal distribution width of 0.3 and a downshift of 1.8. Statistical conditions between the groups were calculated by t-test with a permutation based FDR of 0.05 and an S0 of 0.1. Statistical tables were exported and data was further processed in Microsoft Excel 365 for comprehensive visualization.

GFP-Trap co-immunoprecipitation with Lrif1 mutants in HEK293T cells

Transfected HEK293T cells were washed 2x with ice-cold PBS and lysed on the dish with 1 ml of NP40 lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1% NP-40) supplemented with 1x PI, 20 mM NaF, 20 mM NEM, 2 mM $MgCl_2$ and 250U Benzonase (EMD Millipore, #E1014). Whole cell lysates were incubated at 4°C, for 1 h while rotating, then spun down for 14,000g, 10 min at 4°C. 1 mg of whole cell lysate was added to 20 µl pre-washed GFP-Trap agarose beads and incubated for 1.5 h at 4°C while rotating. Beads were washed 4x with NP40 lysis buffer and proteins were eluted by boiling the beads in 2xSB at 95°C for 15 min while shaking. Input samples represent 2% of material used for IP.

Endogenous co-immunoprecipitation in mESCs

mESCs were washed 2x with ice-cold PBS and lysed on the dish with EBC lysis buffers (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5% NP-40, 2 mM $MgCl_2$) supplemented with 1x PI, 20 mM NaF and 20 mM NEM. Whole cell lysates were incubated at 4°C, for 15 min while rotating, then spun down for 14,000g, 10 min at 4°C. 500 µg of the whole cell lysate was added to 20 µl of antibody pre-linked Dynabeads and incubated overnight at 4°C while rotating. Protein A Dynabeads (Thermo Fisher Scientific, #10002D) were used for conjugation of the antibodies of rabbit origin and protein G Dynabeads (Thermo Fisher Scientific,#10003D) were used for Co-IP with the antibodies of mouse origin. The following antibodies were used for endogenous Co-IPs: RαKap1 (Abcam, #ab10483), RαLRIF1 (Proteintech, #26115-1-AP), RαIgG (Cell Signalling, #2729S), MαKap1 (Abcam, #ab22553) and MαIgG (Merck, #12-371). The next day, beads were washed 4x with EBC lysis buffer and proteins were eluted by boiling the beads in 2xSB at 95°C for 15 min while shaking.

Chromatin immunoprecipitation followed by qPCR

Feeder MEFs were removed by pre-plating the trypsinized cell suspension 2x for 20 min on gelatinized culture plates. The supernatant was collected and washed once in 1x warm PBS followed by crosslinking with 1% formaldehyde in PBS for 10 min at room temperature while tumbling. The reaction was quenched by adding glycine to a final concentration of 125 mM. Crosslinked cells were washed twice with PBS and the cell pellet was either stored at -80°C or proceeded to chromatin isolation. Cell pellets were resuspended in ice-cold ChIP buffer (1.5 ml lysis buffer/10 x 10^6 cells) (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5 % Igepal CA-630, 1% Triton X-100) supplemented with Complete' Protease Inhibitor Cocktail tablet (Sigma-Aldrich, #11697498001). After 10 min incubation on ice, samples were spun down at 8,000 g for 2 min at 4°C. The pellets were resuspended for the second time in ChIP buffer, incubated for 5 min on ice and spun down again. The final nuclear pellets were resuspended in ChIP buffer and sonicated at the highest power output for 15 cycles (1 cycle: 30 sec ON/30 sec OFF) using a Bioruptor instrument (Diagenode). For ChIP, chromatin was first pre-cleared with BSA-blocked protein A Sepharose beads (GE Healthcare, #17-5280-21) by rotating for 30-60 min at 4°C. For histone ChIP, 6 µg of chromatin was used and for Trim28 ChIP, 30 µg of chromatin was used in a final volume of 500 µl. 50 µl (10%) of each chromatin was kept as input sample for later normalization. ChIP was carried out by rotation at 4°C with following primary antibodies: RαTrim28 (Abcam, #ab10483), RαH3 (Abcam, ab1791), RαH3K9me3 (Active Motif,

#39161) or rabbit polyclonal IgG (Abcam, #ab37415), which served as a negative control. The second day, 20 µl of protein A Sepharose beads pre-blocked with BSA were added to all samples and incubated for 2 h at 4°C while rotating. Afterwards, beads were washed as follows: once with low salt wash buffer (1 % Triton X-100, 0.1 % SDS, 2 mM EDTA, 20 mM Tris-HCl, 150 mM NaCl), high salt wash buffer (1 % Triton X-100, 0.1 % SDS, 2 mM EDTA, 20 mM Tris-HCl, 500 mM NaCl), LiCl wash buffer (250 mM LiCl, 1% Igepal CA-630, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl) and twice with TE wash buffer (10 mM Tris-HCl, 1 mM EDTA). For DNA extraction, 10% (w/v) of Chelex 100 resin was added to the beads and boiled at 95°C for 15 min while vigorously shaking. Supernatant was used for qPCR analysis at 60°C using following primers for the *Dux* locus: Dux ChIP F2 5'-CTAGCGACTTGCCCTCCTTCTG-3' and Dux ChIP R2: 5'-ATTCAGAGGGGCTGGAGCAG-3'; *en masse* IAPEz[10]: IAPEz fwd 5'- ACGGGAACACTTCATTACCACC-3' and IAPEz rev 5'- TTGAGAAGGATTCAACTGCGTG-3'; Gapdh int2[24]: Gapdh_int2 F 5'- ATCCTGTAGGCCAGGTGATG-3' and Gapdh_int2 R 5'- AGGCTCAAGGGCTTTTAAGG-3'.

Acknowledgements

We would like to thank members of van der Maarel group for critical reading of the manuscript. DS, JB and SMVDM are members of the European Reference Network for Rare Neuromuscular Diseases [ERN EURO-NMD] and/or members of the Netherlands Neuromuscular Center (NL-NMD). This work was supported by funds from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) grant number R01AR045203 and the Prinses Beatrix Spierfonds grant number W.OR15-26.

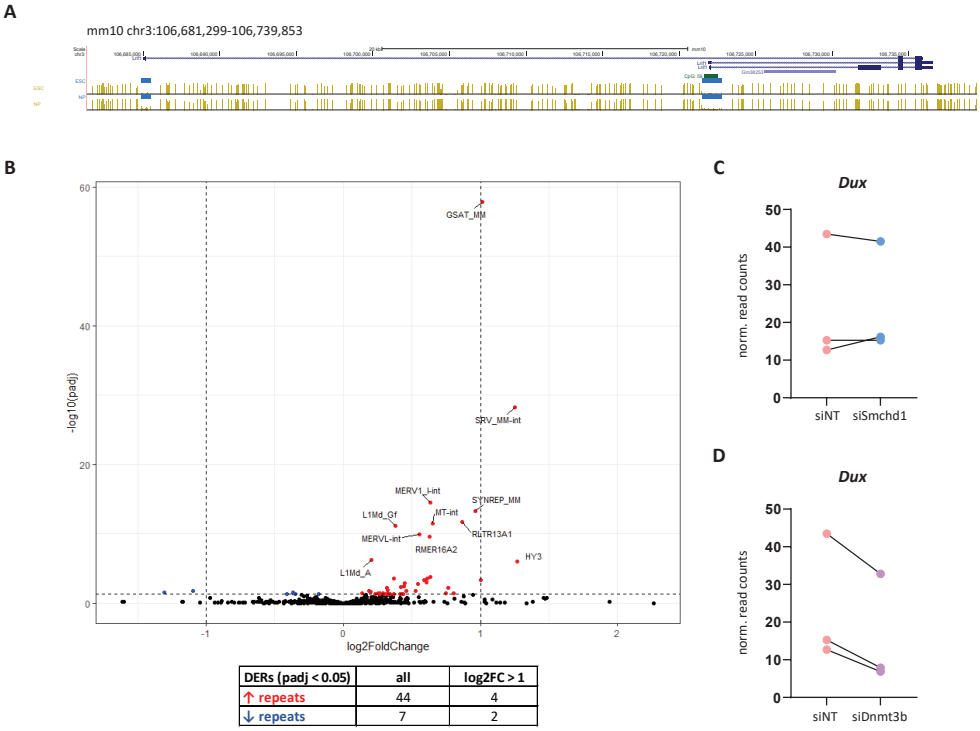
Conflict of interest

Authors declare no conflict of interest.

References

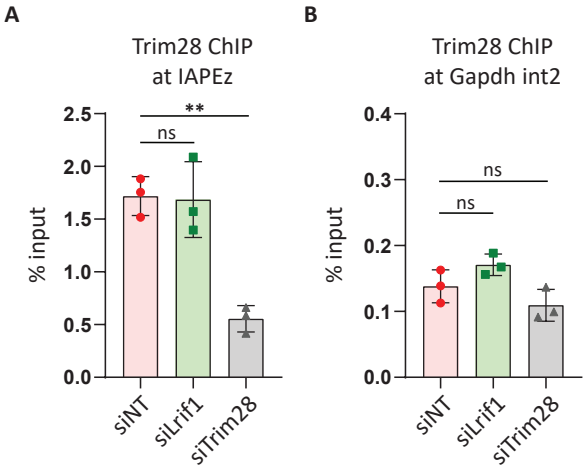
1. Akiyama, T. *et al.* Transient bursts of Zscan4 expression are accompanied by the rapid derepression of heterochromatin in mouse embryonic stem cells. *DNA Res.* **22**, 307–318 (2015).
2. Eckersley-Maslin, M. A. A. *et al.* MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Rep.* **17**, 179–192 (2016).
3. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
4. Zhang, W. *et al.* Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Res.* **47**, 8485–8501 (2019).
5. Genet, M. & Torres-Padilla, M.-E. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* **147**, (2020).
6. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
7. Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391-405.e19 (2018).
8. Huang, Z. *et al.* The chromosomal protein SMCHD1 regulates DNA methylation and the 2c-like state of embryonic stem cells by antagonizing TET proteins. *Sci. Adv.* **7**, eabb9149 (2021).
9. Cossec, J. C. *et al.* SUMO Safeguards Somatic and Pluripotent Cell Identities by Enforcing Distinct Chromatin States. *Cell Stem Cell* **23**, 742-757.e8 (2018).
10. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
11. Eckersley-Maslin, M. *et al.* Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* **33**, 194–208 (2019).
12. Li, H. J., Haque, Z. K., Chen, A. & Mendelsohn, M. RIF-1, a novel nuclear receptor corepressor that associates with the nuclear matrix. *J. Cell. Biochem.* **102**, 1021–1035 (2007).
13. Akram, S. *et al.* LRIF1 interacts with HP1α to coordinate accurate chromosome segregation during mitosis. *J. Mol. Cell Biol.* (2018) doi:10.1093/jmcb/mjy040.
14. Brideau, N. J. *et al.* Independent Mechanisms Target SMCHD1 to Trimethylated Histone H3 Lysine 9-Modified Chromatin and the Inactive X Chromosome. *Mol. Cell. Biol.* **35**, 4053–68 (2015).
15. Nozawa, R.-S. *et al.* Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* **20**, 566–573 (2013).
16. Stoll, G. A. *et al.* Structure of KAP1 tripartite motif identifies molecular interfaces required for retroelement silencing. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15042–15051 (2019).
17. Lomberk, G., Wallrath, L. & Urrutia, R. The Heterochromatin Protein 1 family. *Genome Biol.* **7**, 228 (2006).
18. Maksakova, I. A. *et al.* Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics and Chromatin* **6**, 15 (2013).
19. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237–240 (2010).
20. Rowe, H. M. *et al.* TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res.* **23**, 452–461 (2013).
21. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
22. González-Prieto, R. *et al.* Global non-covalent SUMO interaction networks reveal SUMO-dependent stabilization of the non-homologous end joining complex. *Cell Rep.* **34**, 108691 (2021).
23. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
24. Matsui, T. *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nat. 2010 4647290* **464**, 927–931 (2010).

Supplementary Information



DERs (padj < 0.05)	all	log2FC > 1
↑ repeats	44	4
↓ repeats	7	2

Suppl. Figure 1. Lrif1 knock-down causes upregulation of 2C-specific repeats. A) Snapshot of *Lrif1* locus from UCSC genome browser (mm10) showing three *Lrif1* isoforms. Whole-genome bisulfite sequencing tracks showing mouse embryonic stem cell (ESC) and neuronal (NP) methylomes over *Lrif1* locus from Stadler et al.[25] (available from DNA methylation hub for mm10 in UCSC genome browser) . Regional hypomethylation over two different *Lrif1* transcriptional start sites is marked with blue boxes. **B)** Volcano plot showing expression changes of repeats following Lrif1 knock-down. Upregulated repeats are highlighted in red and downregulated repeats are highlighted in blue. Dashed lines indicate a fold change of two (log_2 fold of 1) on the x axis and significance of 0.05 (–log10 p.adj of 1.3) on the y axis. The top 10 differentially expressed repeats (DERs) are labelled. Table summary of DERs is provided below the plot. **C)** Normalized read counts of *Dux* transcripts after Smchd1 knock- down compared to non-targeting siRNA condition. **D)** Normalized read counts of *Dux* transcripts after Dnmt3b knock down compared to non-targeting siRNA condition.



Suppl. Figure 2. Lrif1 does not affect Trim28 binding to IAPEz. Trim28 ChIP-qPCR of **A)** IAPEz elements, **B)** intron 2 of *Gapdh* in E14 mESCs after treatment with respective siRNAs. Bars and whiskers represent mean ± SEM of three independent experiments. Statistical significance was calculate by one-way ANOVA with Dunnett's post hoc test (ns: not significant; *: < 0.05, **: < 0.01).

Suppl. Table 1. Commonly identified differentially up- and downregulated genes in all three knock-down conditions related to Figure 1 G & H.

Common differentially upregulated genes

ensembl_gene_id	mgi_symbol	chr	start	end	strand
ENSMUSG00000039202	Abhd2	7	78922947	79015256	1
ENSMUSG00000069833	Ahnak	19	8966648	9054278	1
ENSMUSG00000032826	Ank2	3	126715261	127292999	-1
ENSMUSG00000029338	Antxr2	5	98030642	98178902	-1
ENSMUSG00000031511	Arhgef7	8	11777721	11885219	1
ENSMUSG00000024501	Dpysl3	18	43454049	43571351	-1
ENSMUSG00000026131	Dst	1	33947306	34347742	1
ENSMUSG00000003518	Dusp3	11	101861969	101877839	-1
ENSMUSG00000025278	Flnb	14	14518185	14651816	-1
ENSMUSG00000022816	Fstl1	16	37597235	37656876	1
ENSMUSG00000025241	Fyco1	9	123618565	123680964	-1
ENSMUSG00000020176	Grb10	11	11880508	11988683	-1
ENSMUSG00000027007	Itprid2	2	79465696	79503310	1
ENSMUSG00000026478	Lamc1	1	153094668	153208532	-1
ENSMUSG00000031207	Msn	X	95139648	95212158	1
ENSMUSG00000024177	Nme4	17	26310708	26314576	-1
ENSMUSG00000063972	Nr6a1	2	38613382	38817700	-1
ENSMUSG00000039191	Rbpj	5	53623494	53814704	1
ENSMUSG00000037071	Scd1	19	44382894	44396318	-1
ENSMUSG00000025203	Scd2	19	44282113	44295303	1
ENSMUSG00000061186	Sfmbt2	2	10375321	10600064	1
ENSMUSG00000020422	Tns3	11	8381652	8614681	-1
ENSMUSG00000051747	Ttn	2	76534324	76812891	-1

Common differentially downregulated genes

ensembl_gene_id	mgi_symbol	chr	start	end	strand
ENSMUSG00000003309	Ap1m2	9	21205571	21223633	-1
ENSMUSG00000028218	Cibar1	4	12153409	12172015	-1
ENSMUSG00000027552	E2f5	3	14643701	14671369	1
ENSMUSG00000015937	Macroh2a1	13	56221432	56284174	-1
ENSMUSG00000004891	Nes	3	87878385	87887758	1
ENSMUSG00000040204	Pclaf	9	65797519	65810548	1
ENSMUSG00000028134	Ptbp2	3	119512391	119578115	-1
ENSMUSG00000032487	Ptgs2	1	149975782	149983978	1
ENSMUSG00000028464	Tpm2	4	43514711	43523765	-1

Suppl. Table 2. Statistical analysis of the proteins identified by mass spectrometry in the different GFP-tagged Lrif1 isoforms co-immunoprecipitation assays related to Figure 2 A & B.

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	Gene names	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)
Q6P5D8	Smchd1	+	3,62	0,00	11,80	+	3,40	0,00	10,32
P83917	Cbx1	+	4,25	0,00	11,38	+	3,49	0,00	7,36
Q8CDD9	Lrif1	+	2,77	0,00	10,55	+	2,50	0,00	8,93
Q61686	Cbx5	+	6,00	0,00	9,63	+	5,03	0,00	5,39
Q9DCC5	Cbx3	+	2,19	0,00	8,75	+	1,66	0,00	5,59
Q01320	Top2a	+	1,59	0,00	5,47		0,21	0,59	-0,91
Q91W39	Ncoa5	+	4,14	0,00	4,63		NaN	1,00	0,00
Q921K2	Parp1	+	3,47	0,00	3,92	+	1,02	0,01	-2,28
P62806	Hist1h4a	+	2,16	0,00	2,67	+	1,55	0,00	-2,29
Q60749	Khdrbs1	+	3,81	0,00	2,62	+	1,15	0,01	-0,66
E0CZ27	H3f3a	+	1,50	0,00	2,61		0,51	0,14	-1,16
Q8CBB6	Hist1h2br	+	2,03	0,00	2,56	+	1,35	0,01	-1,91
Q8CGP5	Hist1h2af	+	1,79	0,00	2,51	+	1,31	0,01	-2,12
A0A0R4J0I9	Lrp1	+	2,56	0,00	1,98		NaN	1,00	0,00
A2ARV4	Lrp2	+	1,97	0,00	1,95		0,54	0,14	0,63
Q61696	Hspa1a	+	2,31	0,00	1,95		0,81	0,05	0,62
P62334	Psmc6	+	2,50	0,00	1,94		0,36	0,31	0,70
A0A2I3BRL8	Rbmxl1	+	3,18	0,00	1,74	+	2,26	0,00	-0,56
Q91VI7	Rnh1	+	2,54	0,00	1,74	+	3,67	0,00	3,28
P17427	Ap2a2	+	1,78	0,00	1,62	+	1,85	0,00	1,68
Q923G2	Polr2h	+	2,22	0,00	1,52	+	2,53	0,00	1,88
Q8R0G9	Nup133	+	0,93	0,03	1,51	+	1,90	0,00	3,39
Q9ERD7	Tubb3	+	2,04	0,00	1,48	+	3,76	0,00	2,56
Q8BH74	Nup107	+	1,13	0,02	1,46	+	2,31	0,00	3,25
Q9DAE2	Rbmxl2	+	0,86	0,03	1,44		0,52	0,14	0,83
P63166	Sumo1	+	0,86	0,04	1,31	+	1,86	0,00	1,14
Q9WV32	Arpc1b	+	1,76	0,00	1,30		NaN	1,00	0,00
Q9D883	U2af1		0,64	0,09	1,26	+	0,75	0,04	1,45
P14115	Rpl27a	+	3,75	0,00	1,20	+	3,37	0,00	0,74
Q6ZWZ4	Rpl36	+	3,12	0,00	1,14	+	3,22	0,00	1,04
Q99KP6	Prpf19	+	1,70	0,00	1,13		0,19	0,71	0,24
Q60848-2	Hells	+	1,55	0,01	1,10	+	1,28	0,01	0,76
P62242	Rps8	+	6,03	0,00	1,09	+	4,95	0,00	1,04
P47963	Rpl13	+	2,76	0,00	1,03	+	4,08	0,00	1,12
P61514	Rpl37a	+	1,11	0,02	1,00	+	1,05	0,01	0,93
Q9R0Q7	Ptges3	+	2,19	0,00	0,99	+	3,08	0,00	1,90
P63017	Hspa8	+	3,27	0,00	0,99	+	4,03	0,00	1,53
Q9CR57	Rpl14	+	2,46	0,00	0,98	+	3,12	0,00	1,14
A0A0G2JDW7	Rps27	+	3,59	0,00	0,97	+	2,83	0,00	0,75
Q8BP67	Rpl24	+	3,05	0,00	0,96	+	3,53	0,00	1,10
F8WHL2	Copa		0,80	0,06	0,96	+	0,79	0,04	0,90
P25444	Rps2	+	2,22	0,00	0,93	+	2,67	0,00	0,97

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	Gene names	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)
Q9CZM2	Rpl15	+	2,75	0,00	0,92	+	3,10	0,00	0,96
P14148	Rpl7	+	3,25	0,00	0,89	+	2,93	0,00	1,15
O55142	Rpl35a	+	3,09	0,00	0,88	+	2,62	0,00	1,39
P27659	Rpl3	+	2,42	0,00	0,88	+	2,64	0,00	0,93
P20029	Hspa5	+	4,27	0,00	0,88	+	3,15	0,00	0,54
P62702	Rps4x	+	2,50	0,00	0,87	+	2,63	0,00	0,97
P62849-2	Rps24	+	3,45	0,00	0,87	+	2,85	0,00	0,94
A0A1D5RLW5	Rpl18a	+	3,01	0,00	0,87	+	3,59	0,00	1,08
P12970	Rpl7a	+	3,09	0,00	0,87	+	4,39	0,00	1,28
P19253	Rpl13a	+	3,35	0,00	0,84	+	4,06	0,00	1,26
P62855	Rps26	+	2,80	0,00	0,83	+	2,38	0,00	1,19
P68369	Tuba1a		0,34	0,38	0,83	+	0,81	0,03	1,72
P15864	Hist1h1c	+	1,04	0,03	0,82		0,44	0,36	0,24
P62889	Rpl30	+	1,07	0,03	0,81	+	2,19	0,00	0,99
Q9D8E6	Rpl4	+	6,61	0,00	0,80	+	3,74	0,00	1,20
Q99MN1	Kars	+	1,18	0,02	0,80	+	3,07	0,00	2,86
P62751	Rpl23a	+	3,79	0,00	0,79	+	3,96	0,00	1,21
A0A3B2WDD2	Rpl10a	+	2,74	0,00	0,78	+	3,91	0,00	0,90
P99027	Rplp2	+	3,17	0,00	0,78	+	3,10	0,00	0,91
I7HLV2	Rpl10	+	2,55	0,00	0,76	+	3,15	0,00	0,83
Q9JKX6	Nudt5		0,73	0,08	0,76	+	2,23	0,00	1,97
Q62318	Trim28	+	2,51	0,00	0,76	+	3,39	0,00	0,97
Q9CQM8	Rpl21	+	2,75	0,00	0,76	+	3,17	0,00	0,89
P14869	Rplp0	+	2,50	0,00	0,74	+	2,66	0,00	1,07
Q3U4X8	Lig1		0,44	0,25	0,71	+	1,15	0,01	1,41
Q99ME9	Gtpbp4		0,59	0,15	0,71	+	1,89	0,00	0,73
P47911	Rpl6	+	2,78	0,00	0,69	+	3,58	0,00	1,10
Q99L45	Eif2s2	+	2,57	0,00	0,69	+	3,46	0,00	1,27
D3Z2H7	Ctnnd1		0,27	0,51	0,69	+	2,58	0,00	3,19
D6RH49	Rps27l		0,36	0,36	0,69	+	1,98	0,00	1,90
P61358	Rpl27	+	2,38	0,00	0,68	+	3,77	0,00	0,88
P61255	Rpl26	+	1,53	0,01	0,68	+	2,22	0,00	1,14
P35980	Rpl18	+	1,76	0,01	0,67	+	2,35	0,00	0,88
D3YTQ9	Rps15	+	1,61	0,01	0,66	+	2,12	0,00	0,87
P62267	Rps23	+	1,74	0,01	0,65	+	2,43	0,00	1,01
P62918	Rpl8	+	2,94	0,00	0,65	+	3,88	0,00	1,26
P62900	Rpl31	+	3,67	0,00	0,65	+	1,59	0,00	1,57
P35979	Rpl12	+	2,43	0,00	0,64	+	3,12	0,00	0,90
A0A1B0GRR3	Rps11	+	1,20	0,03	0,63	+	3,58	0,00	0,78
P41105	Rpl28	+	1,99	0,01	0,62	+	2,70	0,00	1,17
P49718	Mcm5	+	3,77	0,00	0,62	+	3,85	0,00	0,86
P62192	Psmc1		0,68	0,11	0,61	+	1,42	0,01	1,21
Q8BVQ9	Psmc2		0,94	0,06	0,59	+	2,95	0,00	1,35
H3BKN0	Nsun2	+	2,44	0,00	0,59	+	3,68	0,00	1,15

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	Gene names	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)
Q6ZWX6	Eif2s1	+	3,05	0,00	0,58	+	4,39	0,00	1,12
P63087	Ppp1cc		0,33	0,43	0,57	+	0,83	0,03	1,27
Q6ZWN5	Rps9	+	2,32	0,00	0,57	+	3,41	0,00	1,09
Q9D1R9	Rpl34	+	1,24	0,03	0,57	+	2,11	0,00	0,95
P62301	Rps13	+	1,75	0,01	0,56	+	4,04	0,00	1,01
Q8VEK3-2	Hnrnpu	+	1,66	0,01	0,54	+	1,94	0,00	0,68
Q8C2Q3	Rbm14		0,33	0,44	0,52	+	1,24	0,01	1,29
P97351	Rps3a	+	2,07	0,01	0,51	+	2,73	0,00	0,82
Q6ZWZ7	Rpl17	+	1,70	0,01	0,50	+	2,20	0,00	0,88
Q5M8M8	Rpl29		0,65	0,14	0,50	+	1,23	0,01	0,86
A0A1L1SQA8	Rps25	+	1,88	0,01	0,50	+	2,34	0,01	0,43
Q7TPV4	Mybbp1a	+	2,95	0,00	0,50	+	2,75	0,00	0,54
Q99LE6	Abcf2	+	1,60	0,02	0,48	+	2,46	0,00	0,95
A2A547	Rpl19	+	1,41	0,02	0,47	+	2,66	0,00	1,03
Q9Z2X1	Hnrnpf	+	2,36	0,01	0,47	+	4,41	0,00	1,36
Q8BJY1	Psmd5	+	1,27	0,03	0,46	+	3,23	0,00	1,51
Q6ZWV7	Rpl35	+	1,86	0,01	0,46	+	2,34	0,00	1,25
A0A494BAX5	Nars	+	2,21	0,01	0,45	+	2,79	0,00	1,18
P62754	Rps6	+	1,98	0,01	0,43	+	2,96	0,00	0,72
P62264	Rps14	+	1,20	0,04	0,43	+	3,18	0,00	0,69
P83882	Rpl36a		0,99	0,06	0,42	+	1,53	0,01	0,85
F6YVP7	Gm10260	+	1,32	0,03	0,41	+	3,25	0,00	0,70
A0A0A6YW67	Gm8797		0,47	0,30	0,39	+	2,26	0,00	-1,50
A0A0H2UH27	Fxr1		0,90	0,09	0,39	+	1,58	0,01	0,73
P62245	Rps15a	+	2,02	0,02	0,39		0,54	0,40	0,15
Q61881	Mcm7	+	2,05	0,02	0,38	+	3,02	0,00	0,74
Q62167	Ddx3x	+	1,28	0,05	0,37	+	2,27	0,00	0,60
A0A087WPL5	Dhx9	+	1,56	0,03	0,37		1,18	0,06	0,26
P62960	Ybx1		0,41	0,36	0,37	+	1,15	0,01	1,09
O88477	Igf2bp1	+	1,57	0,03	0,35	+	2,63	0,00	0,67
P54276	Msh6	+	1,60	0,03	0,35	+	2,12	0,01	0,54
P55302	Lrpap1		0,60	0,22	0,34	+	1,61	0,01	-0,64
Q8BTS0	Ddx5	+	1,55	0,03	0,33	+	2,75	0,00	0,55
P05213	Tuba1b	+	2,38	0,02	0,32	+	4,29	0,00	1,33
P62911	Rpl32		1,28	0,06	0,30	+	2,42	0,00	1,18
Q9CZX8	Rps19		0,53	0,31	0,27	+	2,44	0,00	1,02
P68372	Tubb4b		1,00	0,12	0,27	+	3,30	0,00	0,85
Q60668-3	Hnrnpd		0,98	0,12	0,27	+	1,45	0,01	0,43
Q3UKJ7	Smu1		0,31	0,53	0,27	+	1,22	0,03	0,38
P61979-2	Hnrnpk	+	2,72	0,02	0,26	+	3,92	0,00	0,50
Q3U741	Ddx17		0,74	0,21	0,25	+	1,41	0,01	0,48
P17918	Pcna		0,61	0,28	0,24	+	2,35	0,00	0,85
Q9JJI8	Rpl38		1,03	0,14	0,23	+	2,27	0,00	0,72
P16460	Ass1		1,10	0,13	0,22	+	3,09	0,00	0,55

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	Gene names	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)
Q3KQM4	U2af2		0,76	0,23	0,22	+	1,93	0,01	0,41
P14131	Rps16		0,39	0,51	0,19	+	1,34	0,01	0,50
A0A140LIZ5	Psmc4		0,46	0,45	0,18	+	2,29	0,00	0,81
Q61553	Fscn1		1,21	0,20	0,16	+	2,17	0,01	0,35
A2AGN7	Psmc3		0,57	0,42	0,16	+	1,42	0,01	0,44
Q9CWF2	Tubb2b		0,49	0,48	0,15	+	3,05	0,00	1,11
P51881	Slc25a5		1,16	0,26	0,14	+	1,97	0,01	0,43
Q8K1K2	Psmc5		0,96	0,40	0,11	+	3,77	0,00	0,60
P09405	Ncl		0,53	0,55	0,11	+	0,98	0,03	0,58
P52480	Pkm		0,55	0,55	0,11	+	2,05	0,01	0,41
Q8C0C7	Farsa		0,36	0,67	0,09	+	2,49	0,00	0,64
Q9R0N0	Galk1		0,42	0,67	0,08	+	2,99	0,00	0,76
P27612	Plaa		0,52	0,66	0,07	+	4,99	0,00	1,65
P35486	Pdha1		0,13	0,84	0,06	+	1,15	0,03	0,41
P29595	Nedd8		0,08	0,89	0,05	+	1,68	0,01	-0,53
Q91YZ2	Ctbp2		0,14	0,85	0,05	+	2,38	0,00	0,69
G5E902	Slc25a3		0,20	0,84	0,04	+	2,01	0,01	0,40
Q9R1T2-2	Sae1		0,04	0,95	0,02	+	1,60	0,01	0,61
P80315	Cct4		0,05	0,97	0,01	+	1,85	0,01	0,35
P99024	Tubb5		0,03	0,98	0,00	+	2,80	0,00	0,76
Q45VK5	Ilf3		NaN	1,00	0,00	+	1,36	0,01	1,00
D3Z7K0	Otub1		NaN	1,00	0,00	+	1,31	0,01	0,53
P70404	Idh3g		NaN	1,00	0,00	+	1,57	0,00	1,62
Q9CQM5	Txndc17		NaN	1,00	0,00	+	1,71	0,00	2,26
A0A087WPE4	Tceb1		0,02	0,97	-0,02	+	1,13	0,01	0,70
P60843	Eif4a1		0,23	0,87	-0,03	+	2,47	0,01	0,30
P51410	Rpl9		0,21	0,84	-0,04	+	2,05	0,01	-0,50
P09103	P4hb		0,17	0,84	-0,04	+	1,15	0,01	-0,61
P20152	Vim		0,02	0,97	-0,05	+	1,18	0,01	-1,53
E9Q242	Adsl		0,22	0,79	-0,06	+	2,36	0,00	0,49
E9QAI5	Cad		0,81	0,63	-0,07	+	4,34	0,00	0,79
E9QN08	Eef1d		0,25	0,73	-0,09	+	1,86	0,01	0,52
Q922D8	Mthfd1		0,23	0,73	-0,10	+	1,11	0,03	-0,43
P80317	Cct6a		0,50	0,56	-0,11	+	1,97	0,01	0,38
P80318	Cct3		0,52	0,54	-0,12	+	1,81	0,01	0,40
Q01853	Vcp		0,85	0,37	-0,13	+	2,95	0,00	-1,01
P62137	Ppp1ca		0,34	0,58	-0,16	+	4,36	0,00	0,71
D3YZX3	Gnb2		0,11	0,82	-0,16	+	1,67	0,00	1,01
P25206	Mcm3		0,42	0,51	-0,16	+	1,42	0,03	0,33
A2AM74	Kif17		1,58	0,12	-0,18	+	0,84	0,04	-0,77
P68134	Acta1		0,15	0,77	-0,19	+	1,48	0,00	-2,21
Q9EST5	Anp32b		0,09	0,84	-0,19	+	0,87	0,02	1,67
A0A2R8W6Y5	Larp4		0,78	0,24	-0,20	+	1,38	0,01	0,45
P38647	Hspa9		0,95	0,16	-0,23	+	2,32	0,01	-0,45

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	Gene names	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)
G3V004	Calu		0,30	0,55	-0,24	+	1,16	0,02	0,55
Q8VDW0	Ddx39a		1,02	0,13	-0,24	+	1,54	0,03	-0,30
P63330	Ppp2ca		1,61	0,06	-0,25	+	2,96	0,00	-0,49
E9Q8F0	Rbm39		1,08	0,11	-0,25	+	1,23	0,02	-0,46
P19096	Fasn		0,83	0,17	-0,26	+	1,53	0,02	0,33
P97315	Csrp1		0,68	0,22	-0,26	+	1,58	0,01	-0,70
Q4VBE8	Wdr18	+	1,87	0,03	-0,27		0,06	0,94	0,04
P18760	Cfl1		1,46	0,06	-0,27	+	2,35	0,01	-0,42
Q8K3F7	Tdh		1,04	0,10	-0,28	+	1,37	0,02	-0,38
Q792Z1	Try10		1,36	0,06	-0,29	+	1,01	0,05	-0,38
P52624	Upp1		0,90	0,13	-0,29	+	0,94	0,03	0,59
P80316	Cct5	+	2,08	0,02	-0,30		1,17	0,14	0,18
A0A0A0MQA5	Tuba4a		0,23	0,63	-0,30	+	3,41	0,00	1,41
Q61990	Pcbp2	+	1,56	0,03	-0,32		1,77	0,07	-0,18
P48962	Slc25a4	+	1,45	0,04	-0,33		0,33	0,74	-0,07
A0A075B6B4	Trav6-4	+	1,60	0,03	-0,33	+	1,53	0,01	-0,65
P08249	Mdh2		0,68	0,17	-0,36	+	1,36	0,01	-0,68
P63085	Mapk1		0,48	0,30	-0,36	+	2,05	0,00	-1,46
P80314	Cct2	+	1,76	0,02	-0,38	+	4,59	0,01	-0,27
P60335	Pcbp1	+	1,26	0,04	-0,39		0,56	0,37	-0,17
P63325	Rps10	+	1,51	0,03	-0,40	+	1,65	0,01	-0,38
P17742	Ppia		1,01	0,07	-0,40	+	2,24	0,00	-0,87
Q61024	Asns	+	1,86	0,02	-0,41	+	2,69	0,00	-0,61
P45376	Akr1b1	+	1,70	0,02	-0,41		0,80	0,06	-0,50
P56480	Atp5b	+	1,40	0,03	-0,42	+	4,12	0,00	-0,89
Q03265	Atp5a1	+	3,06	0,00	-0,42	+	3,74	0,00	-0,83
P26443	Glud1	+	2,17	0,01	-0,43	+	3,80	0,00	-1,04
A0A087WS46	Eef1b2	+	1,42	0,03	-0,44		1,46	0,05	0,23
P50247	Ahcy	+	1,83	0,01	-0,45		0,78	0,13	-0,29
Q8K2B3	Sdha	+	1,50	0,02	-0,46	+	1,12	0,02	-0,50
Q9DB20	Atp5o	+	1,46	0,02	-0,46	+	1,83	0,00	-0,84
Q9D8W5	Psmd12		0,84	0,09	-0,47	+	1,52	0,01	-0,98
P19324	Serpinh1	+	1,68	0,02	-0,47	+	1,93	0,00	-0,73
Q497W9	Dhx15	+	1,34	0,03	-0,49	+	1,79	0,01	-0,63
P84078	Arf1	+	1,52	0,02	-0,50	+	4,32	0,00	-0,90
A0A0A0MQM0	Eif5a	+	1,79	0,01	-0,50	+	2,14	0,01	-0,46
P29341	Pabpc1	+	2,95	0,00	-0,52	+	2,94	0,00	-0,62
P27773	Pdia3	+	3,35	0,00	-0,52	+	5,09	0,00	-1,09
E9PZF0	Gm20390	+	1,87	0,01	-0,53	+	3,05	0,00	-1,06
Q8BWY3	Etf1	+	1,34	0,02	-0,54	+	2,35	0,00	-0,98
Q9CPY7	Lap3	+	1,94	0,01	-0,54	+	3,38	0,00	-1,16
Q61171	Prdx2	+	1,66	0,01	-0,54	+	2,92	0,00	-1,16
P46935	Nedd4		0,51	0,22	-0,55	+	1,39	0,01	-0,99
P07901	Hsp90aa1	+	1,71	0,01	-0,55		0,62	0,19	-0,29

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	Gene names	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)	Significant (FDR=0.05 S0=0.1)	-Log p-value	q-value	Difference (log2)
Q9D0R8	Lsm12	+	1,58	0,01	-0,56	+	1,81	0,01	-0,50
O88569-3	Hnrnpa2b1		0,55	0,19	-0,56	+	1,90	0,00	-1,37
P06151	Ldha	+	1,90	0,01	-0,57	+	1,22	0,01	-1,09
Q8BG32	Psmd11	+	2,42	0,00	-0,58	+	3,06	0,00	-0,77
Q9DCL9	Paics	+	2,08	0,00	-0,59	+	1,71	0,01	-0,57
Q922R8	Pdia6	+	1,79	0,01	-0,60	+	2,57	0,00	-0,92
Q9CZ13	Uqcrc1	+	1,47	0,02	-0,60		0,41	0,50	0,15
A0A0J9YUD8	Hmgb1	+	1,57	0,01	-0,62		0,03	0,96	-0,06
P63038	Hspd1	+	2,61	0,00	-0,62	+	3,09	0,00	-1,02
P62827	Ran	+	3,28	0,00	-0,62	+	3,55	0,00	-0,98
P68040	Gnb2l1	+	3,66	0,00	-0,63	+	4,13	0,00	-1,17
Q61598-2	Gdi2	+	2,15	0,00	-0,63	+	2,40	0,00	-1,06
Q9CXW3	Cacybp	+	1,49	0,01	-0,64	+	2,28	0,00	-1,15
Q64433	Hspe1	+	1,37	0,02	-0,65	+	2,23	0,00	-0,94
P58252	Eef2	+	2,17	0,00	-0,66	+	2,85	0,00	-1,02
Q9CQR2	Rps21	+	3,64	0,00	-0,67	+	1,13	0,01	-1,69
Q7TNC4-2	Luc7l2		0,90	0,06	-0,68	+	0,95	0,03	-0,66
P40142	Tkt	+	2,00	0,00	-0,69	+	3,37	0,00	-1,12
P10126	Eef1a1	+	3,19	0,00	-0,70	+	3,63	0,00	-0,88
Q61937	Npm1	+	1,63	0,01	-0,70		0,69	0,10	-0,46
A0A0U1RNT6	Mat2a	+	2,24	0,00	-0,71		0,89	0,08	-0,33
P17751	Tpi1	+	2,81	0,00	-0,71	+	2,59	0,00	-1,08
P63028	Tpt1	+	2,48	0,00	-0,72	+	2,75	0,00	-1,09
A0A1D5RLS2	Nudt21	+	1,85	0,00	-0,73	+	2,48	0,00	-0,97
A6ZI44	Aldoa	+	2,31	0,00	-0,73	+	2,47	0,00	-1,07
Q62446	Fkbp3	+	1,47	0,01	-0,75	+	2,23	0,00	-1,28
P17182	Eno1	+	2,81	0,00	-0,75	+	2,56	0,00	-1,12
Q8K274	Fn3krp		0,83	0,06	-0,76	+	1,08	0,01	-0,92
Q5F2E7	Nufip2	+	1,68	0,01	-0,76	+	1,66	0,01	-0,77
P30681	Hmgb2	+	1,33	0,02	-0,77		0,42	0,19	-1,54
P14685	Psmd3	+	1,57	0,01	-0,79	+	2,08	0,00	-1,14
P11440	Cdk1		0,48	0,21	-0,79	+	2,25	0,00	-0,64
S4R1W1	Gm3839	+	3,35	0,00	-0,80	+	2,67	0,00	-1,00
B1AXW5	Prdx1	+	4,03	0,00	-0,80	+	2,67	0,00	-1,19
P57784	Snrpa1	+	2,26	0,00	-0,82	+	2,57	0,00	-1,04
Q3U2G2	Hspa4	+	0,98	0,03	-0,82	+	0,82	0,03	-0,95
Q9D0M3	Cyc1	+	1,01	0,03	-0,83		0,81	0,06	-0,47
E9Q5Q0	Atxn2l	+	3,02	0,00	-0,83	+	3,32	0,00	-0,84
Q3UL36	Arglu1	+	2,05	0,00	-0,85	+	2,25	0,00	-1,04
A0A0N4SV32	Serbp1	+	2,69	0,00	-0,86	+	2,59	0,00	-0,96
Q9CPN9	22100 10C04Rik	+	1,48	0,01	-0,86	+	1,12	0,04	-0,39
P47738	Aldh2	+	0,87	0,05	-0,87	+	1,00	0,03	-0,56
O89086	Rbm3	+	1,26	0,02	-0,88	+	1,48	0,01	-0,94

GENE		Lrif1 short vs Control				Lrif1 long vs Control			
Uniprot ID	**Gene names**	**Significant (FDR=0.05 S0=0.1)**	**-Log p-value**	**q-value**	**Difference (log2)**	**Significant (FDR=0.05 S0=0.1)**	**-Log p-value**	**q-value**	**Difference (log2)**
Q60817	Naca	+	1,44	0,01	-0,88	+	2,72	0,00	-1,52
G3UXT7	Fus	+	1,60	0,01	-0,88	+	1,63	0,01	-0,80
Q9CZN7-2	Shmt2	+	1,62	0,01	-0,89	+	4,57	0,00	-1,47
Q9CZU6	Cs	+	3,86	0,00	-0,89	+	2,29	0,00	-1,28
P14206	Rpsa	+	4,14	0,00	-0,89	+	3,02	0,00	-1,31
Q60864	Stip1	+	1,95	0,00	-0,89	+	2,70	0,00	-1,39
Q5EBP8	Hnrnpa1	+	0,88	0,05	-0,89	+	1,27	0,01	-1,08
P63260	Actg1	+	1,78	0,00	-0,90	+	3,87	0,00	-1,46
P24369	Ppib	+	2,14	0,00	-0,90	+	2,22	0,00	-1,42
P61982	Ywhag	+	2,88	0,00	-0,91	+	3,38	0,00	-1,08
Q9DBJ1	Pgam1	+	2,58	0,00	-0,92	+	2,29	0,00	-1,17
P54823	Ddx6		0,47	0,22	-0,94	+	1,09	0,03	-0,53
Q5XJY5	Arcn1	+	1,06	0,02	-0,97	+	1,53	0,01	0,66
Q8VIJ6	Sfpq	+	1,96	0,00	-0,99	+	2,54	0,00	-0,95
Q8BK67	Rcc2	+	2,37	0,00	-0,99	+	2,05	0,00	-1,26
A2AL12	Hnrnpa3	+	2,96	0,00	-1,01	+	2,04	0,00	-1,06
Q8QZY1	Eif3l	+	1,08	0,02	-1,01	+	1,31	0,01	-1,21
P07356	Anxa2	+	2,71	0,00	-1,07	+	3,19	0,00	-1,69
Q80X90	Flnb	+	3,77	0,00	-1,07	+	2,87	0,00	-1,13
Q99K48	Nono	+	1,62	0,00	-1,10	+	2,52	0,00	-0,64
B1AZS9	Prdx4	+	1,92	0,00	-1,11	+	3,78	0,00	-1,39
D3Z0Y2	Prdx6	+	1,56	0,01	-1,12	+	1,88	0,00	-1,59
Q9CWJ9	Atic	+	1,01	0,03	-1,17	+	0,96	0,02	-0,89
Q61792	Lasp1	+	2,95	0,00	-1,22	+	3,46	0,00	-1,34
Q8BGJ5	Ptbp1	+	2,65	0,00	-1,26	+	2,61	0,00	-1,52
Q99KI0	Aco2	+	2,10	0,00	-1,31	+	2,58	0,00	-1,17
A0A1L1SV25	Actn4	+	1,60	0,00	-1,35	+	3,69	0,00	-2,21
Q9JMD0-4	Znf207	+	1,09	0,02	-1,36	+	2,36	0,00	-1,95
O35685	Nudc	+	3,71	0,00	-1,65	+	3,99	0,00	-1,88
A2BE93	Set	+	1,46	0,00	-1,69		0,08	0,90	-0,15
P10852	Slc3a2	+	3,40	0,00	-1,79	+	2,95	0,00	-2,38
Q99LX0	Park7	+	2,89	0,00	-1,79		0,15	0,85	-0,07
P14733	Lmnb1	+	1,21	0,01	-1,80	+	1,89	0,00	-1,83
P26043	Rdx	+	2,22	0,00	-2,29	+	2,44	0,00	-2,36
P55821	Stmn2	+	3,06	0,00	-2,35	+	3,19	0,00	-3,10
F8WIT2	Anxa6	+	4,59	0,00	-2,55	+	5,81	0,00	-2,98
Q07076	Anxa7	+	6,72	0,00	-2,80	+	3,53	0,00	-2,65
Q6PIX5-2	Rhbdf1		0,66	0,07	-2,86	+	2,15	0,00	-1,20
P21107-2	Tpm3	+	5,50	0,00	-2,97	+	4,50	0,00	-3,43

Suppl. Table 3. Primers used for qRT-PCR.

Primer name	Primer sequence 5'->3'
mβ-actin_RT_F	GGCTGTATTCCCCTCCATCG
mβ-actin_RT_R	CCAGTTGGTAACAATGCCATGT
mLRIF1l+s qPCR F	AAGATGCAAACATTGTGGTG
mLRIF1l+s qPCR R	CCATCTTCATGGTTTCCGC
Smchd1 ex44_F	AAGCCCTTTGGAAATCCAGT
Smchd1 ex46_R	TGGGGCAGTGTGTGATTTTA
mDnmt3b_RT_Ex16-17_F	GGAAGAATTTGAGCCACCCA
mDnmt3b_RT_Ex18_R	GACTTCGGAGGCAATGTACTT
endo-mOct4-F	TAGGTGAGCCGTCTTTCCAC
endo-mOct4-R	GCTTAGCCAGGTTCGAGGAT
endo-mSox2-F	AGGGCTGGGAGAAAGAAGAG
endo-mSox2-R	CCGCGATTGTTGTGATTAGT
endo-mNanog-F	CTCAAGTCCTGAGGCTGACA
endo-mNanog-R	TGAAACCTGTCCTTGAGTGC
mTrim28-1241-F2	CTGGTACGAACTCCACAGGT
mTrim28-1439-R2	CCACTTACCTCTCCCTCACC
mDux_1 F	ACTTCTAGCCCCAGCGACTC
mDux_1 R	CCATGCTGCCAGGATTTCTA
Gm21761 F	GATCCCTGAGGGTAAGTCCTCC
Gm21761 R	TGCTTCCTATCCAGCTCTTGAGG
Usp17lb F	CTTCCCAGAAGATCCAGCC
Usp17lb R	CTGTGCTTTCCATTGGCAG
Gm2016 F	TACTCACCAGGTCAATGCAG
Gm2016 R	AGGAAGGTGTAGTCTCCCT
Tmem92 F	GTAAGCTTCAATGAGACTGCA
Tmem92 R	GCAGCATTCCTTGACACAG
mZscan4e-358-F	TTGAAGCCTCCTGTCATGGT
mZscan4e-515-R	TGTGTGGTGTCTACTGGCAT

Suppl. Table 4. Information about batch effects due to separate knock-down experiments and re-sequencing of siDnmt3b_1 sample.

Sample Name	Experiment ID	Sequencing run ID
siNT_1	E1	SR1
siNT_2	E2	SR1
siNT_3	E3	SR1
siLrif1_1	E1	SR1
siLrif1_2	E2	SR1
siLrif1_3	E3	SR1
siSmchd1_1	E1	SR1
siSmchd1_2	E2	SR1
siSmchd1_3	E3	SR1
siDnmt3b_1	E1	SR2
siDnmt3b_2	E2	SR1
siDnmt3b_3	E3	SR1

Suppl. Table 5. Cloning primers for creating mPVL and m1 mutants in Lrif1l/s ORF.

Mutant Name	Insert	Forward primer (5'->3')	Reverse primer (5'->3')
Lrif1l mPVL	Insert 1	CATGGTCCTGCTGGAGTTCGTG	GATGGTCAGGAATTCGAGTTTCA-CAGTCTCTCAAATCTTTAGTGAG
	Insert 2	CTCACTAAAGATTTGAGAGACTGT-GAAACTCGAATTCCTGACCATC	ACAGGGATTTCTTGTCTCCC
	Insert 1 + Insert 2	CATGGTCCTGCTGGAGTTCGTG	ACAGGGATTTCTTGTCTCCC
Lrif1l m1	Insert 1	CATGGTCCTGCTGGAGTTCGTG	CTTTTTCTCTTAAAATTTGCT-CAGCTTCTCTTATTTTTTCATCTCTGATG
	Insert 2	CATCAGAGATGAAAAAATAA-GAGAAGCTGAGCAAATTTTAA-GAGAAAAAG	ACAGGGATTTCTTGTCTCCC
	Insert 1 + Insert 2	CATGGTCCTGCTGGAGTTCGTG	ACAGGGATTTCTTGTCTCCC
Lrif1s mPVL	Insert 1	CATGGTCCTGCTGGAGTTCGTG	GATGGTCAGGAATTCGAGTTTCA-CAGTCTCTCAAATCTTTAGTGAG
	Insert 2	CTCACTAAAGATTTGAGAGACTGT-GAAACTCGAATTCCTGACCATC	ACAGGGATTTCTTGTCTCCC
	Insert 1 + Insert 2	CATGGTCCTGCTGGAGTTCGTG	ACAGGGATTTCTTGTCTCCC
Lrif1s m1	Insert 1	CATGGTCCTGCTGGAGTTCGTG	CTTTTTCTCTTAAAATTTGCT-CAGCTTCTCTTATTTTTTCATCTCTGATG
	Insert 2	CATCAGAGATGAAAAAATAA-GAGAAGCTGAGCAAATTTTAA-GAGAAAAAG	ACAGGGATTTCTTGTCTCCC
	Insert 1 + Insert 2	CATGGTCCTGCTGGAGTTCGTG	ACAGGGATTTCTTGTCTCCC

CHAPTER 6

General Discussion

Previous studies uncovered the identity of genetic and epigenetic factors contributing to FSHD being either a contraction of the 4qA-linked D4Z4 repeat to a size of 1-10 units (FSHD1) or mutations in D4Z4 chromatin regulators combined with an intermediate-sized 4qA-linked D4Z4 repeat (FSHD2). Both situations lead to a disruption of the heterochromatic structure of the D4Z4 macrosatellite repeat in somatic cells. The current scientific consensus in the FSHD field is that both forms of the disease, albeit mechanistically distinct, converge at the level of expression of the D4Z4 repeat-encoded *DUX4* gene in skeletal muscle. Although we still lack a thorough understanding of the pathological pathways triggered by DUX4, the long-awaited identification of the FSHD disease gene over a decade ago helped us to shift our focus from exploring merely symptomatic or generic treatments for FSHD to developing specific molecular therapies aiming at interfering with *DUX4* expression in skeletal muscle. More detailed knowledge about the minimal genetic and epigenetic requirements for stable *DUX4* expression in muscle cells could thus translate into more potent and longer-lasting therapeutic strategies.

For a long time, our knowledge about FSHD was largely based on population and family studies looking at inter-individual differences associated with the disease. This led to the identification of the disease locus (D4Z4) as well as two of its *trans* modifiers, namely SMCHD1 and DNMT3B. Especially, identifying rare FSHD cases with "non-standard" genetic and/or epigenetic characteristics, such as FSHD2 cases, can further help us to untangle the molecular mechanisms underlying D4Z4 dysregulation in FSHD. In addition, recent advances in the development of genome modifying tools allow us to start directly testing the relevance of these (epi)genetic observations collected from population and family studies in a more controlled manner and distinguish which observed features are only associated with the disease and which are causally related. These complementary research approaches were also used in the work presented in this thesis, which contributes to both the genetic and epigenetic understanding of FSHD.

So far, a single 4qA-specific single nucleotide polymorphism creating a polyadenylation signal (PAS) for *DUX4* in somatic cells offered a straightforward genetic explanation for the unique linkage of FSHD to a *DUX4*-expressing D4Z4 repeat. For that reason, this PAS has been considered an attractive therapeutic target. In **chapter 2**, we capitalize on this long-standing view regarding the essentiality of the non-canonical *DUX4* PAS for the production of polyadenylated *DUX4* transcript from the 4qA repeat in FSHD myocytes by developing a genetic therapy targeting its sequence motif. Although we observe the desired effect of *DUX4* downregulation, we also uncover a more complex genetic basis for FSHD as the data suggest that a combinatorial effect of multiple 4qA-specific sequence polymorphisms *in cis* to the *DUX4* PAS SNP contribute to *DUX4* expression and disease presentation.

Genetically, the FSHD-associated partial loss of chromatin-mediated *DUX4* repression has been explained by either reduced D4Z4 copy number or by germline mutations in *SMCHD1* or *DNMT3B*. However, some individuals with a clinical presentation of FSHD remain genetically undiagnosed. In **chapter 3**, we expand the hereditary basis of FSHD by identifying an

individual presenting with clinical and molecular features characteristic for FSHD who carries a homozygous loss-of-function mutation in the *LRIF1* gene. Following up on this discovery in **chapters 4 and 5**, we provide an initial framework for understanding the role of LRIF1 in D4Z4 repression by investigating the consequences of its loss in human somatic cells and in mouse embryonic stem cells, respectively, and we propose that LRIF1 most likely influences the establishment of the D4Z4 chromatin structure as we show that it does not play a role in the somatic maintenance of this structure.

Targeting *cis* modifier(s) for genetic therapy in FSHD

The fact that the *DUX4* open reading frame is contained within a single exon, which has been partially or fully multiplied throughout the primate genome, [1] creates an obstacle for employing a straightforward *DUX4* knock-out strategy using CRISPR/Cas9. Targeting D4Z4 sequences directly might lead to genome-wide double stranded breaks and such widespread collateral DNA damage from Cas9 activity might cause undesirable genomic instability and could be therefore more harmful than beneficial [2]. For this and other reasons, many studies looked into alternative ways to achieve *DUX4* repression either by (1) using antisense oligonucleotides [3–10], miRNAs [11,12] or recombinant U7 small nuclear RNA (snRNA) [13] to manipulate its post-transcriptional fate, (2) trying to prevent its transcription through re-establishing a repressive D4Z4 chromatin environment with the use of a modified CRISPR/Cas9 system fused to diverse repressor proteins [14–16], or (3) by engaging the endogenous RNAi pathway [17]. However, the DNA editing toolkit has expanded in the meantime from the initial simple Cas9 nuclease to DNA editing solutions that do not rely on double strand DNA breaks such as base editors [18,19] and prime editors [20]. Furthermore, the downregulation of gene expression can be achieved not only by introducing premature stop codons in its open reading frame but also by mutating conserved regulatory *cis* elements important for proper pre-mRNA processing such as splice sites [21] or PASs [22]. Indeed, antisense oligonucleotide-mediated steric hindrance of either splice sites or the PAS in *DUX4* pre-mRNA was shown to lead to its efficient knock-down both *in vitro* and *in vivo* [6,8–10,13].

Consequences of targeting *DUX4* polyadenylation signal

As *DUX4* transcripts expressed in FSHD skeletal muscle cells utilize the PAS that lies in exon 3 located immediately distal to the 4qA-linked D4Z4 macrosatellite repeat structure, in **chapter 2**, we explored the potential of *DUX4* PAS mutagenesis as a genetic therapy for FSHD. Given the adenine-rich nature of canonical PAS motifs, AATAAA and ATTAAA, the latter representing the *DUX4* PAS, we decided to use a previously developed adenine base editing system that can convert A:T base pairs of choice into G:C base pairs as long as the PAM sequence is appropriately spaced [18]. First, we tested two different ABE versions which were based on SpCas9 available at that time, namely SpABE7.10 [18] and SpABEmax [23]. With both base editors, we could achieve editing of the *DUX4* PAS in the haploid model cell line HAP1 confirming that the locus is targetable by this system. Also in our hands, ABEmax showed superiority over ABE7.10 in its editing efficiency as previously published [23]. Next, we carried

out *DUX4* PAS editing in three independent FSHD immortalized myogenic cell lines. In all of them, we could obtain successfully edited clonal lines carrying diverse editing outcomes that impair the *DUX4* PAS. A surprising observation was that there were relatively large differences in *DUX4* expression levels together with its targets within each group of unedited clones as well as edited clones derived from the same parental cell culture. Interestingly, the highest inter-clonal expression variability (over three orders of magnitude) was detected among clones from the FSHD1 cell line which carries a 7 unit-long 4qA D4Z4 repeat, while this variability was much less prominent (only one order of magnitude) between clones derived from an FSHD1 cell line with 3 unit-long 4qA D4Z4 repeat and clones from an FSHD2 cell line with a heterozygous *SMCHD1* mutation combined with 11 unit-long 4qA D4Z4 repeat. The *DUX4* expression level of individual clones seems to be mitotically stable as examining *DUX4* expression after further clonal outgrowth of low or high *DUX4* expressing clones was similar to the parental clone. The nature of this clonal variability remains to be investigated but could be due to subtle clone-specific epigenetic differences in the D4Z4 locus or might relate to the immortalization process of the cell lines as different clones might carry different integration sites for the immortalization transgenes (*hTERT* and *CDK4*), which could influence their expression. The latter might be especially relevant since the length of 4q telomere was previously shown to modulate *DUX4* expression [24], thus clonal lines with different amounts of hTERT, an enzyme which is responsible for post-replicative lengthening of telomeres, could result in different telomere lengths in the clones that might contribute to observed *DUX4* expression differences. Nevertheless, editing of the *DUX4* PAS did yield lower *DUX4* expression levels which correlated with a reduction in steady state mRNA levels of its direct transcriptional target genes suggesting successful knock-down also on protein level. However, in contrast to our expectation based on the current genetic explanation for this disease, i.e. 4qA-specific *DUX4* expression due to a functional PAS being present only in the 4qA background but not in the 10qA background, we did not achieve complete abrogation of polyadenylated *DUX4* transcript production by editing any or all of the three distal adenines of the PAS motif. Examining the cleavage and polyadenylation sites of the *DUX4* transcripts produced in the edited clones did reveal that the majority of them ended at a different position than *DUX4* transcripts from unedited cells. One could argue that our mutagenesis was focused on different nucleotides than the SNP that differs between 4qA and 10qA and that editing any of the three distal adenines of the PAS into guanines could thus be less detrimental than the third nucleotide position of PAS motif changing from T to C, which defines the 4qA/10qA SNP. However, we have also derived two clones after editing which carried a partial or complete deletion of the PAS sequence and yet we detected polyadenylated *DUX4* transcripts in these clones. Furthermore, two recent studies also attempted to abolish the *DUX4* PAS on DNA level in immortalized FSHD1 myoblasts [16,25]. One study used a standard CRISPR/Cas9 nuclease system combined with a pair of sgRNAs flanking the *DUX4* PAS region to completely excise it. In agreement with our results, this study also showed reduced levels of *DUX4* mRNA as well as two of its target genes (*ZSCAN4* and *TRIM43*) in the edited cells compared to non-edited FSHD1 myogenic cells. The other study aimed to disrupt the *DUX4* PAS by inserting a sequence that can be recognized by miR-1. miR-1 is a miRNA that is naturally expressed in skeletal muscle and binding of miR-1

to its cognate site within mRNAs interferes with their translation (reviewed here: Safa et al., 2020). By extension, the authors hypothesized that any residually produced *DUX4* mRNA bearing this sequence would be further inhibited from DUX4 protein production. However, the authors managed to derive only a single clone with the expected insertion (out of 227 clones screened clones) and were not able to examine its effect as the clonal line ceased to proliferate. However, as a by-product of editing, they also obtained one clone in which the *DUX4* PAS was deleted altogether and showed, consistent with our observation, that this leads to decreased but not fully absent *DUX4* mRNA levels.

Regarding the feasibility of genome editing approaches for genetic therapy in FSHD, each of them poses different challenges that relate to specificity, efficiency and *in vivo* delivery. Both aforementioned published studies relied on creating double strand breaks either with CRISPR/Cas9 nuclease or with TALENs. Since we do not know the "uniqueness" of the pLAM region in the genome, creating double stranded breaks might lead to similar undesired increased mutagenesis as with targeting the D4Z4 repeat units directly. However, if one was to resort to excising the *DUX4* PAS sequence, TALENs might be a more attractive option than CRISPR/Cas9 due to TALENS' more stringent target site recognition (up to 36 bp long) and because they can be designed to target virtually any DNA sequence [27], whereas Cas9 targeting requires the presence of its cognate PAM site and usually relies on the recognition of an additional 19 to 22 bp depending on the Cas9 species [28]. This represents a challenge also in our approach since we rely on Cas9-mediated recognition of the target site. Furthermore, the sgRNA design for base editing is even more restricted as the adenines to be edited need to be within a certain distance from the PAM site [18]. Multiple groups are trying to address these limitations by either further engineering Cas9 variants with broadened PAM site compatibility [29,30], by modifying the deaminase enzyme to widen its editing window and improve its catalytic properties [31] or by reorganizing the 3D architecture of the whole base editing complex [32].

How much editing is enough?

One of the outstanding questions is how many nuclei would need to be edited to achieve therapeutic benefit. Some clues can be derived from studies of mosaic FSHD individuals when post-zygotic contractions of the D4Z4 repeat result in a mixture of normal-sized and FSHD1-sized alleles within one individual [33]. These cases tend to present with a later disease onset and a milder progression of the disease than non-mosaic cases with comparable contracted repeat sizes [34]. This seems to depend on the residual repeat size and the proportion of cells carrying the contracted allele [33]. Further substantiating this dilution effect, an *in vitro* study testing whether fusing FSHD1 myoblasts with healthy myoblasts could rescue the myogenic differentiation defect and DUX4-related expression phenotype suggested the requirement of at least 50% of healthy nuclei to be mixed with FSHD1 nuclei to form a hybrid myotube for the near-complete phenotype correction [35]. However, this percentage might still depend on the capacity of the FSHD nucleus to express *DUX4*; thus individuals with shorter D4Z4 repeats might require a larger proportion of unaffected nuclei to suppress the pathogenic effects of

DUX4. There are many as yet unknown factors that might influence *DUX4* expression in skeletal muscle and therefore it is difficult to predict how much editing is necessary *in vivo*. But, since myogenic cells with the edited *DUX4* PAS can still express *DUX4*, albeit at lower levels, the number of edited nuclei *in vivo* might need to be higher than the 50% suggested by tissue culture experiments to achieve the desired effect. Therefore, it will be of utmost importance to test adenine base editing strategy of *DUX4* PAS in a skeletal muscle tissue context to assess its translatability.

Considerations for *DUX4* PAS adenine base editing *in vivo*

In our study, we used an original SpCas9-based editing system which left us with a single sgRNA for targeting the *DUX4* PAS as this PAM site was the only one that fulfilled the base editing design criteria. We also tested two different base editing systems using SaCas9 or CjCas9 orthologues, however, we did not observe *DUX4* PAS base editing with those. Assuming equal editing efficiency, these would be preferred over SpCas9-based system as their size would allow for their intact packaging into AAV vectors (maximum packaging capacity being 5 kb), which are currently considered the gold standard for delivering gene therapies *in vivo* [36]. The SpABE *in vivo* delivery problem can be partially solved by employing a dual trans-splicing adeno-associated virus (AAV) approach that relies on first splitting the construct into two halves, which are then delivered by separate AAVs followed by their *in vivo* reconstitution [37]. The caveat of this approach is the requirement of transducing the cells with both independent AAV particles and *in vivo* protein re-assembly efficiency. Since the target tissue in FSHD is skeletal muscle, the first concern might not be as relevant, as myofibers are syncytia containing hundreds of nuclei sharing their cytoplasmic space [38]. Therefore, many more nuclei could potentially receive the reconstituted base editor for their subsequent editing by infecting a single myofiber as opposed to the need of infecting the comparable number of individual mononuclear cells. Few groups have already tested the intein split system for delivering SpABE to skeletal muscle tissue and reported variable A to G editing efficiencies ranging from 4 to 30% [39–41]. Further improvements in skeletal muscle trophism of AAVs [42], use of muscle-specific promoters [43,44] as well as ABE optimization [31] might improve the *in vivo* therapeutic potential of adenine base editing.

Currently, many preclinical studies are conducted in mouse models for respective diseases. This represents a challenge in the FSHD field given that the D4Z4 macrosatellite repeat is primate-specific [45]. For this reason, different mouse models expressing the *DUX4* transgene were developed, each based on different considerations with respect to the design of the transgene construct [46–50]. Two FSHD mouse models seem to be particularly well suited to test the adenine base editing strategy *in vivo*, namely the iDUX4pA [51] and the FLExDUX4 model [50]. Both mouse models have integrated in their genome the human *DUX4* gene structure including its 3'UTR region in which *DUX4* expression relies on its native PAS sequence. Both also allow for tunable *DUX4* expression enabling modeling of variable disease severity. One missing feature in these two mouse models, which is of relevance to FSHD, is the repeat structure and epigenetic context of the endogenous human locus. The so-called D4Z4-2.5

mouse model was generated with this aspect in mind when an EcoRI fragment cloned from an FSHD1 individual containing a 2-unit long 4qA D4Z4 array was randomly integrated in the mouse genome [49]. While this mouse model does recapitulate chromatin features associated with the contracted D4Z4 repeat in FSHD1, *DUX4* expression is very low in skeletal muscles and mice do not develop any skeletal muscle phenotype. Furthermore, the D4Z4 transgene was integrated at least four times in tandem, which would make the editing evaluation of the *DUX4* PAS target site more cumbersome due to its multiplication. In contrast, the aforementioned FLExDUX4 model has already been utilized by different research groups for testing antisense oligonucleotides approaches [3–5,7] as well as AAV-mediated delivery of a CRISPR/Cas9 repressor system [14] to reduce *DUX4* expression *in vivo*. Therefore, using this mouse model for testing also our *DUX4* PAS base editing strategy would allow for the comparison of its effectiveness with other therapeutic approaches which are being currently investigated.

Another concern with adenine base editing, as with any CRISPR/Cas9-derived platform, is the potential off-target effects. In our study, we have started to address this concern by performing an *in silico* prediction of potential off-target DNA sites based on their sequence homology to the used sgRNA followed by a PCR-targeted Illumina sequencing investigation of ten of the top-scoring off-target sites out of 227 predicted ones by the CRISPOR prediction tool. We could detect adenine to guanine editing at three out of ten examined sites, albeit with much lower frequencies as for the target site. Nevertheless, this showed that the used sgRNA can guide the ABE to other genomic sites. Therefore, a more thorough evaluation of the sgRNA-dependent DNA editing is required. While the new T2T genome assembly will be informative in predicting novel off-target sites, ideally, an unbiased approach should be pursued, which would experimentally assess potential off-target sites genome-wide such as the recently developed EndoV-seq method [52]. Apart from DNA off-target editing, ABEs can also induce A to I editing in cellular RNAs [53–56]. This might become especially problematic when AAVs were to be used for the ABE delivery as they sustain long-term expression *in vivo* that could result in cumulative transcriptome changes over time especially in such a low turnover tissue such as skeletal muscle. Therefore, safer ABE variants with reduced RNA editing activity might be needed before their introduction to the clinic [53]. Furthermore, since our initial evaluation of off-target editing was conducted in HAP1 cells, a better model more closely representing the target tissue should be used to evaluate the safety of this approach. For this, a myogenic model in the form of a 2D cell culture, 3D muscle bundle or muscle xenograft derived from cells of a healthy individual with a permissive 4qA D4Z4 allele would be suitable. Any expression changes observed in these models would be attributable to the off-target effect of editing rather than *DUX4* PAS editing as the locus is in that case silent. Furthermore, derivation of such model from an FSHD1 mosaic individual would allow obtaining cells representing both the disease as well as healthy state creating genetically-matched settings for their comparison [57,58].

Finding other *cis* modifiers of *DUX4* expression

Auxiliary *cis* sequence elements have been shown to influence the efficiency of PAS usage [59]. One such element has been also reported downstream of the 4qA *DUX4* PAS and its targeting with antisense oligonucleotides leads to reduced gene expression [60]. However, its characterization and functional testing has only been conducted by using a reporter construct that was transfected in HEK293T cells. Therefore, it remains to be addressed if this sequence also plays a role in *DUX4* transcript processing from the endogenous locus in FSHD myogenic cells. Furthermore, it is not likely that the combination of the *DUX4* PAS with the aforementioned *cis* element fully explains the genetic basis of FSHD since residual *DUX4* expression can be detected after deleting this whole downstream region altogether [16]. Unfortunately, the cleavage and polyadenylation site of residual *DUX4* transcripts has not been assessed after this intervention. Nevertheless, other 4qA-specific polymorphisms which differ from 10qA alleles most likely influence *DUX4* expression. Identifying these other *cis* modifiers could provide us with alternative genetic targets or can be used in combination with targeting the *DUX4* PAS thus further enhancing the *DUX4* knock-down potential. To identify these, population genetics strategies can be employed and custom *in vitro* genetic cellular models can be studied. Firstly, the 4qA and 10qA D4Z4 repeats are known to undergo inter-chromosomal rearrangements creating hybrid alleles [61,62]. Recently, two individuals presenting with FSHD were identified, who have a contracted hybrid D4Z4 repeat that ends with a 4qA type repeat on chromosome 10 from which *DUX4* is expressed in myogenic cell cultures [63]. Such genetic rearrangements are rare, but with time more individuals might be identified with different rearrangement breakpoints which could narrow down the minimal 4qA polymorphisms which are important for *DUX4* expression. Secondly, genome editing tools allow us to speed up this discovery process by creating different genetic situations ourselves by either forced *in vitro* rearrangements between 4qA and 10qA with CRISPR/Cas9 [64] or by converting each 4qA polymorphism into a 10qA sequence at a time with the use of base editors or prime editors. Doing this in a transcriptionally permissive 4qA D4Z4 chromatin environment in FSHD cells would also permit immediate assessment of the effect of each SNP on *DUX4* expression. A reciprocal approach can be also employed, i.e. converting SNPs in the 10qA D4Z4 allele into 4qA-like to assess their role in the gain of *DUX4* expression. For example, a cytidine in 10qA *DUX4* PAS (AT**C**AAA) can be converted into thymine as present in 4qA *DUX4* PAS (AT**T**AAA) by employing BE4-Gam cytidine base editor [65] at least in HEK293T cells (Figure 1). Doing this in a myogenic cell line derived from a control individual who carries either contracted 10qA D4Z4 repeat or an FSHD-causative *SMCHD1* mutation in combination with 10qA allele of intermediate size, scenarios which both provide a chromatin susceptible state for *DUX4* expression, would provide further insight into necessity of having a functional PAS for sustainable *DUX4* expression vs other differences within 10qA D4Z4 that might hinder stable *DUX4* transcription.

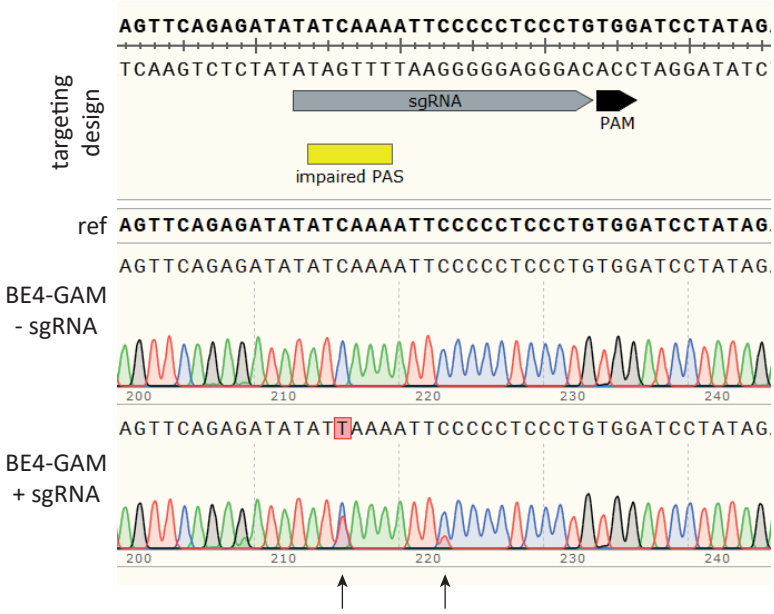


Figure 1. Cytidine base editing approach for converting the 10qA *DUX4* PAS SNP into a 4qA *DUX4* PAS SNP. Snapgene view of a section of the pLAM region in the 10qA D4Z4 allele with outlined the SpCas9 PAM site (black), the sgRNA (grey) and the impaired *DUX4* PAS sequence (yellow). Sanger sequencing alignment showing successful editing of cytidine into thymine within 10qA *DUX4* PAS in HEK293T cells which received both BE4-Gam and the sgRNA. The mutated nucleotide is highlighted in red and with an arrow (C -> T). Notice also bystander editing which occurred on one of the downstream cytidines marked with an arrow.

Functional characterization of a newly identified FSHD gene – *LRIF1*

Identification of *LRIF1* as a new FSHD disease gene

In **chapter 3**, we describe a male individual presenting with clinical symptoms of FSHD and having profound hypomethylation of both 4q and 10q D4Z4 repeats which is reminiscent of FSHD2. This individual was identified in a screen for FSHD2 cases in a Japanese FSHD cohort with unknown aetiology [66]. Of the 20 patients having D4Z4 hypomethylation and a permissive allele, a mutation in *SMCHD1* was identified by Sanger sequencing in 13 of them. Candidate Sanger sequencing of *LRIF1,* among other genes*,* in the remaining 7 unexplained cases revealed one patient with a homozygous 4 nt frame-shift duplication (c.869_872dup) in exon 2. Exon 2 of *LRIF1* is differentially spliced resulting in the production of two different mRNA and protein isoforms (referred to as long and short depending on whether exon 2 is included). Western blot analysis of the proband's fibroblasts confirmed the selective loss of the long isoform of LRIF1 due to a premature stop codon (p.Trp291Ter). The proband has a 13 unit-long 4qA D4Z4 repeat which in combination with its hypomethylation makes him susceptible to *DUX4* expression in skeletal muscle. As *DUX4* expression is considered

to be the root cause of FSHD, its detection in proband's cells would provide molecular confirmation of his clinical diagnosis. Only primary dermal fibroblasts were available from this individual which required their MYOD1-forced trans-differentiation into myogenic cells [67] to test *DUX4* expression, which was indeed detected together with selected DUX4 target genes. Furthermore, immortalized fibroblasts of the proband showed increased levels of H3K4me2 and H3K27me3 and decreased levels of H3K9me3 at D4Z4, which is consistent with the previously described chromatin profile of D4Z4 in FSHD2 individuals with mutations in *SMCHD1* [68,69]. Interestingly, both LRIF1 isoforms are known interaction partners of SMCHD1 in somatic cells [70]. We showed that LRIF1 binds to D4Z4 in unaffected myogenic cells during proliferation (myoblasts) and after differentiation (myotubes). However, we were unable to assess if both LRIF1 isoforms bind to this region and could only show pan-LRIF1 enrichment at D4Z4 due to the lack of isoform specificity of commercially available antibodies. This pan-LRIF1 enrichment at D4Z4 was together with SMCHD1 reduced in the proband, which suggested that recruitment of SMCHD1 to D4Z4 is either dependent on LRIF1 or on D4Z4 chromatin marks which have changed due to the homozygous germline *LRIF1* mutation.

In contrast to *DNMT3B* and *SMCHD1*, whose heterozygous loss-of-function mutations are sufficient to cause D4Z4 hypomethylation, from this family it seems that a homozygous loss of at least the long LRIF1 isoform is required to result in D4Z4 hypomethylation as the mother, who is a heterozygous carrier of this mutation, has normal D4Z4 methylation levels (60% vs 15% in the proband). This might also suggest that either there is no functional redundancy of LRIF1 isoforms in respect to D4Z4 repression as the short isoform of LRIF1 is still expressed in the proband or that the combined amount of the C-terminal portion both LRIF1 isoforms is necessary for D4Z4 repression, in which case the proband's situation could be interpreted as haploinsufficiency of the LRIF1 C-terminus. So far, all LRIF1 functional domains have been described to reside in this C-terminal part of the protein (including HP1 binding motif, nuclear localization signal and SMCHD1 interaction region), whereas the function, if any, of the N-terminally extended region specific to the LRIF1 long isoform is unknown. Therefore, the proposed C-terminal haploinsufficiency explanation is appealing. However, we have observed increased amounts of the short LRIF1 isoform in the proband's cells by western blot and the same phenomenon was also observed upon siRNA-mediated knock-down of the long isoform of LRIF1 in control, FSHD1 and FSHD2 myoblasts. In **chapter 4**, we show that this increase in the short isoform is due to a direct autoregulatory loop of the long LRIF1 isoform acting on the *LRIF1* locus, where it acts as a transcriptional repressor. Loss of the long LRIF1 isoform thus leads to transcriptional upregulation of the *LRIF1* locus, ultimately resulting in higher expression of the short LRIF1 isoform. Therefore, the LRIF1 C-terminus haploinsufficiency scenario in the proband seems an unlikely explanation for FSHD and rather suggests functional divergence between the two isoforms with a critical and unique role of the long LRIF1 isoform in D4Z4 repression.

Further evidence for a new inheritance pattern leading to FSHD

Interestingly, another FSHD2 family with a potentially *LRIF1* damaging variant was uncovered from the aforementioned screen (Figure 2, unpublished results; collaboration with Prof. Nishino, National Institute of Neuroscience, Japan). In this family, a heterozygous 2 nt deletion (c.2148_2149del) resulting in a premature stop codon (p.His718PhefsTer4) in exon 4 of *LRIF1* was detected, thus affecting both LRIF1 isoforms. In addition, a heterozygous 1 nt substitution in *DNMT3B* gene (c.1229G>A) was also detected in this family leading to an in-frame missense variant (p.Arg410Gln). This DNMT3B variant has been reported previously in dbSNP (rs772079891), ExAC and GnomAD databases, although its allelic frequency is rather rare (<0.0001%). However, despite it being predicted by *in silico* prediction tool PolyPhen-2 to be possibly damaging, it has not been reported in the ClinVar database suggesting that the variant by itself is non-pathogenic. Indeed, in this family only the combination of LRIF1 and DNMT3B variants together with a 7 unit-long 4qA D4Z4 repeat resulted in FSHD and repeat hypomethylation. Arg410 in DNMT3B is predicted to be citrullinated and DNMT3B was previously identified as a substrate for peptidylarginine deiminase 4 (PAD4)-mediated citrullination [71]. Earlier, citrullination of DNMT3A by PAD4 was shown to positively influence its protein stability, therefore, one could hypothesize that citrullination of DNMT3B might work in a similar fashion and its loss could lead to lower amounts of functional DNMT3B protein, which only in combination with other predisposing factors. i.e. a FSHD-sized repeat and a heterozygous pathogenic variant in LRIF1 causes sufficient *DUX4* de-repression in skeletal muscle to cause disease. On the other hand, since the *LRIF1* variant found in this family is located in the last exon, it is unlikely that it causes haploinsufficiency of both LRIF1 isoforms. Instead it could rather result in the production of truncated isoforms missing the last C-terminal 48 aa. Coincidentally, the SMCHD1 interaction region maps to this very C-terminal end of LRIF1, therefore, the resulting protein isoforms might act in a dominant negative manner as they still contain the nuclear localization signal and the HP1 binding motif. In this case, they might compete with WT LRIF1 isoforms for HP1 binding but fail to recruit or interact with SMCHD1 for chromatin compaction. Nevertheless, this mutation alone is not sufficient to cause D4Z4 hypomethylation (mother case) and it remains to be investigated if *LRIF1* mutation carriers in this family indeed produce a mixture of full-length and truncated LRIF1 isoforms. Thus, the possible synergistic effect of these two LRIF1 and DNMT3B variants on the D4Z4 chromatin structure might be sufficient to cause *DUX4* expression in the affected siblings.

Since our publication [72], no other FSHD cases caused by *LRIF1* mutations have been reported. However, a grant submitted to the FSHD Global Research organization by Prof. Rosella Tupler (University of Massachusetts Medical School, Worcester, USA) reports two sisters diagnosed with a severe form of FSHD who were born from healthy parents (https://fshdglobal.org/grants/grant-26/). According to the freely available grant summary, both sisters carry the same homozygous mutation in *LRIF1* causing the loss of one of the two LRIF1 isoforms. Although it is not specified if the missing isoform is the long one, a mutation that would result in the specific loss of the short isoform while not affecting the long isoform is highly

improbable. Hypothetically, only mutations strengthening the acceptor splice site of exon 2 leading to its constitutive splicing or a specific in-frame mutation abrogating the methionine start codon of the short LRIF1 isoform in exon 3 could potentially result in the loss of short isoform. However, in the latter case, the resulting mutation (either missense or methionine deletion) would be also present in the long isoform with an unknown effect on the protein. Therefore it is safe to speculate that the reported mutation, like in our case, leads to the loss of the long isoform. As both daughters are homozygous carriers, it can be deduced that the parents are heterozygous carriers of this mutation and that at least one of them has to carry a permissive 4qA D4Z4 repeat which was inherited by both daughters. Since neither parent is affected, this again indicates that the heterozygous loss of the long LRIF1 isoform in combination with a 4qA D4Z4 repeat is insufficient to cause FSHD. It will be interesting to gain more information about this family in regards to what is exactly the disease-causing *LRIF1* mutation, the D4Z4 methylation pattern of the family members as well as the 4q D4Z4 sizes and their haplotypes. Altogether, this and our data add to the modes of inheritance in FSHD, now including monogenic autosomal dominant (FSHD1 (OMIM: 158900); contracted permissive 4qA D4Z4 repeat), digenic autosomal dominant (FSHD2 (OMIM: 158901); heterozygous mutations in either *SMCHD1* or *DNMT3B* in combination with permissive 4qA D4Z4 repeat) and digenic autosomal recessive (FSHD3? (OMIM: 619477); recessive mutations in *LRIF1* in combination with permissive 4qA D4Z4 repeat). Furthermore, the necessity of biallelic mutations in *LRIF1* specifically leading to the loss of the long LRIF1 isoform might explain why FSHD cases due to *LRIF1* mutations are rare.

Regulation of D4Z4 repression by SMCHD1 and LRIF1 in somatic cells

The so far explained contraction-independent FSHD cases have been attributed to germline mutations in three factors, namely DNMT3B, SMCHD1 and LRIF1, resulting in heterochromatin erosion of the D4Z4 repeat in somatic cells. Of these three factors, only SMCHD1 and LRIF1 are significantly expressed in soma and were shown to modulate somatic *DUX4* expression by modifying their protein levels [69,72–74], which is suggestive of a role in D4Z4 repression also in somatic cells. Therefore, we decided to create isogenic myogenic cell models to investigate the effect of somatic loss of either SMCHD1 or LRIF1 on D4Z4 chromatin. Since both FSHD-associated LRIF1 and SMCHD1 mutations are considered loss of function (or sometimes dominant negative in case of SMCHD1), by reasoning that heterozygous mutations may only yield subtle changes in gene expression, we decided to create independent homozygous knock-outs of both genes. We generated three different knock-out situations: 1) full SMCHD1 knock-out ($SMCHD1^{KO}$), 2) selective long LRIF1 isoform knock-out ($LRIF1L^{KO}$) or 3) full LRIF1 knock-out ($LRIF1LS^{KO}$) in two different control myogenic cell lines carrying permissive 4qA D4Z4 alleles of different sizes (32- and 13-units long). All three somatic KO conditions were viable and we have not observed major differences in proliferation between the different conditions (unpublished observation), although $SMCHD1^{KO}$ cells exhibited enhanced myogenic differentiation. The role of SMCHD1 in myogenesis should thus be further studied as based on this observation it seems to behave as an inhibitor of this process. In addition, it was previously shown that protein levels of

SMCHD1 naturally decrease during myogenic differentiation [69] again suggesting that lower levels of SMCHD1 are favouring differentiation. This should be kept in mind when considering FSHD therapies involving overexpression of SMCHD1, as higher SMCHD1 levels might cause delayed myogenic differentiation and interfere with muscle regeneration.

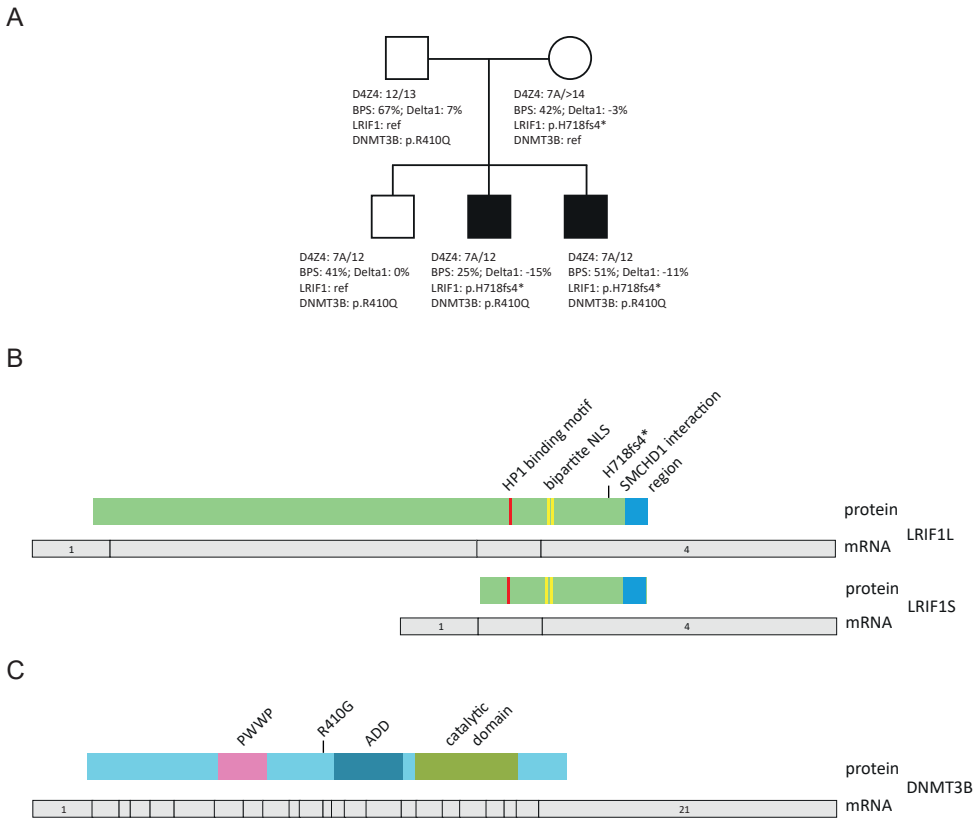


Figure 2. FSHD potentially caused by combined heterozygous *LRIF1* and *DNMT3B* variants. A) Pedigree of the family with information about D4Z4 sizing and its allelic type (A/B) if known, D4Z4 methylation (as determined by bisulfite PCR of DR1 region (BPS) and Delta1 score, which represents a difference in methylation between experimentally observed methylation level and predicted methylation level based on control individuals), and *DNMT3B* and *LRIF1* variants. **B)** Schematic representation of the two LRIF1 mRNA and protein isoforms with the indicated position of identified LRIF1 variant (H718fs4). **C)** Schematic representation of the full DNMT3B mRNA and protein isoform together with known functional domains. Note that DNMT3B can also undergo differential splicing leading to at least three different protein isoforms. The DNMT3B variant (R410G) identified in the FSHD family is indicated.

Interestingly, all KO situations resulted in mild *DUX4* transcriptional de-repression which was insufficient for robust activation of examined DUX4 targets as typically seen in FSHD1 or FSHD2 muscle cell cultures. These transcriptional changes were in concordance with an unchanged D4Z4 chromatin structure, i.e. unaltered DNA methylation levels as well as absence of changes in histone modifications (H3K4me2, H3K9me3, H3K27me3), which are deregulated in FSHD somatic cells. We further investigated the interdependency of SMCHD1

and LRIF1 binding to D4Z4 and found that LRIF1 binding to D4Z4 is SMCHD1-dependent but that SMCHD1 binding is independent of either LRIF1 isoform. The combined loss of SMCHD1 and LRIF1 from D4Z4 in SMCHD1^{KO} cells thus might explain the more pronounced *DUX4* de-repression in these cells as compared to either LRIF1^{KO} situation. Interestingly, despite the overall LRIF1 enrichment at D4Z4 not changing in LRIF1L^{KO} cells, we could detect *DUX4* de-repression suggesting that even in somatic cells the function of both LRIF1 isoforms is non-redundant in respect to D4Z4 repression.

In addition to SMCHD1-dependent LRIF1 recruitment, other chromatin factors must be responsible for its D4Z4 association since selectively increasing SMCHD1 levels at the D4Z4 repeat does not result in increased LRIF1 binding as shown by studying the FSHD2 cell line in which we corrected the *SMCHD1* mutation. This can perhaps be explained by the persistent de-repressed chromatin state in these somatic cells even after *SMCHD1* correction as DNA methylation levels or histone modification patterns were not rescued. Therefore, one can hypothesize that stable LRIF1 association with D4Z4 is apart from SMCHD1 either directly or indirectly dependent also on the normal repressive chromatin structure. In agreement with this, we also found reduced enrichment of SMCHD1 and LRIF1 at D4Z4 in fibroblasts which were derived from individuals with germline mutations in *DNMT3B* as well as in HCT116 cells in which both *DNMT1* and *DNMT3B* were knocked out resulting in D4Z4 hypomethylation. This suggests that SMCHD1 and LRIF1 binding to D4Z4 is either directly or indirectly influenced by DNA methylation. In both cell models, DNA hypomethylation leads to a decrease in H3K9me3. Previously, it was shown that reducing H3K9me3 levels at D4Z4 either by chaetocin treatment or SUV39H1 knock-down results in SMCHD1 dissociation from D4Z4 [1]. However, H3K9me3 alone likely cannot act as the primary targeting mechanism of SMCHD1 and LRIF1 to D4Z4 since neither of them is known to directly recognize this mark. Therefore, intermediate factors must specify their targeting. In respect to that, LRIF1 was shown to interact with HP1 proteins via its HP1 recognition motif [70], thus its association with D4Z4 could be mediated by both SMCHD1 and one of the HP1 homologues. Particularly, HP1$_{\gamma}$ is enriched at the D4Z4 repeat in control cells [1] and we could detect decreased HP1$_{\gamma}$ levels in HCT116 DNMT double knock-out cells (unpublished observations). The HP1-dependent LRIF1 stability at D4Z4 could be tested by knocking down or knocking out of either individual HP1 homologues or their combination since they can act redundantly depending on the genomic context [75]. However, the hypothesis for SMCHD1 recruitment to D4Z4 in somatic cells is more challenging. It would be interesting to test whether SMCHD1 association with HP1 is strictly LRIF1-dependent or whether there is an additional independent mechanism that mediates their interaction which would explain the observed H3K9me3-dependent SMCHD1 association with D4Z4. Nevertheless, how these two factors exactly mediate D4Z4 repression remains elusive. They might aid in further chromatin compaction such as in the case of the inactive X chromosome [70], or antagonize the binding of activating factors as shown by competition of Smchd1 and Ctcf at the protocadherin gene cluster in mouse neural stem cells [76], or they might help in tethering D4Z4 to the silent nuclear compartment such as lamina-associated domains (LADs). Both SMCHD1 and LRIF1 have been recently identified as components of the LADs microproteome [77].

Implicating Lrif1 in repression of the *Dux* repeat in mESCs

In **chapter 5**, we initially wanted to investigate the possible relationship between the trio of FSHD genes – *SMCHD1*, *DNMT3B* and *LRIF1*. For this, we used mouse ES cells cultured in serum condition in which all three genes are expressed. First, we compared the transcriptomes of mESCs in which we knocked down each factor individually to identify common differentially expressed genes, which would have suggested that all three proteins co-regulate the same genomic regions. Surprisingly, we found very little albeit significant overlap between the three knock-down conditions. Therefore, we continued with the most remarkable finding, which is the upregulation of 2C-specific genes as well as repeats in Lrif1 knock-down cells. This we attributed to the upregulation of *Dux,* which is a known activator of the 2C transcriptional program [78,79]. A recent study identified Smchd1 as a direct *Dux* repressor in mESCs and interactor of Tet proteins [80]. The authors proposed a model in which the interaction of Smchd1 with Tet proteins results in local shielding of the *Dux* locus from Tet-mediated DNA demethylation thus protecting it from its re-activation. Unexpectedly, we did not detect upregulation of *Dux* or 2C-specific genes in Smchd1 knock-down cells. In addition, as opposed to our knock-out studies in human immortalized myoblasts, where knocking out SMCHD1 outperformed LRIF1 knock-outs in terms of the number of differentially expressed genes, in mESCs we observed the opposite. This could be due to the transient depletion strategy as either the knock-down efficiency or its duration might not have been sufficient for complete de-repression of Smchd1-repressed loci. Therefore, creating knock-out situations for these genes would be important to confirm the observed transcriptional phenotypes of the respective knock-down situation. Furthermore, since the knock-down of Lrif1 was performed by using a mixture of four different siRNAs leading to depletion of all Lrif1 isoforms (three in mouse as opposed to two in human), it would be interesting to see if Lrif1 isoform-specific knock-outs would elicit the same transcriptional response of *Dux* as is the case for human *DUX4* in somatic cells. Lastly, since mESCs naturally fluctuate between pluripotent and 2C-like states *in vitro* [81], it will be important to assess if the Lrif1 knock-down influences this fluctuation equilibrium in favor of 2C-like cells which would explain the detection of their transcriptomic signature in our bulk RNA-seq data. To test this hypothesis, a fluorescent reporter specifically labelling the 2C-like cell population [81,82] could be used to quantify the shift in ESC vs 2C-like cells population by FACS in response to Lrif1 depletion.

In our quest to explain the 2C-like transcriptional signature upon Lrif1 knockdown, we identified Trim28 (also known as Kap1) as a novel interacting partner of both Lrif1 isoforms in mESCs. Trim28 has been previously reported to act as a negative regulator of conversion of mESCs into 2C-like cells [81] via a mechanism that involves direct repression of *Dux* locus [78,83]. We observed reduced binding of Trim28 to *Dux* in Lrif1-depleted mESCs suggesting a direct or indirect involvement of Lrif1 in Trim28-mediated *Dux* repression. In agreement with this, we also observed reduced H3 levels at the *Dux* locus which could be attributed to chromatin remodeling of the locus or increased chromatin accessibility. Trim28 contains several functional domains that facilitate protein-protein interactions including an N-terminal RING-B-box-coiled-coil (RBCC) domain, which mediates binding to hnRNPK [84]

and KRAB-ZFPs [85]; a PxVxL motif for interaction with HP1 homologues [86] and a PHD finger-bromodomain important for interaction with Setdb1 [87] and NuRD complex [88]. We showed that the interaction of Lrif1 with TRIM28 is not bridged by HP1 proteins, which are common interacting partners of both proteins, and this interaction also does not rely on Lrif1's C-terminal alpha helix that mediates the interaction with SMCHD1. Therefore, it will be crucial to determine first if the Lrif1 interaction with Trim28 is direct by performing *in vitro* GST pull down assay with purified proteins. In case of a direct interaction, a crosslinking-MS/MS might help to narrow down the Lrif1 domain that facilitates interaction with Trim28. In addition, the PHD finger domain of Trim28 also acts as SUMO E3 ligase for sumoylation of its adjacent bromodomain [89,90] and sumoylation of Trim28 is necessary for its interaction with Setdb1 and the NuRD complex to silence ERV elements in mESCs [91]. Since we used Sumo-protecting IP conditions by adding N-ethylmaleimide (an unspecific chemical inhibitor of de-sumoylation) into our lysis and IP buffers, it should be determined if the interaction of Lrif1 with Trim28 is also Sumo-dependent.

Recently, the *Dux* repeat has been shown to localize to perinucleolar heterochromatin (PNH) in mESCs, which was dependent on nucleolar integrity mediated by rRNA biogenesis [92]. Interestingly, earlier work already reported a link between the *Dux* repression and rRNA synthesis [83]. Both of these processes were dependent on LINE1-mediated recruitment of nucleolin (NCL) together with Trim28 to *Dux* and rDNA [83]. Indeed, disruption of rRNA synthesis leads to dissociation of the Ncl/Trim28 complex from PNH and increased conversion of mESCs into 2C-like cells due to a failure in *Dux* repression [92]. The involvement of nucleoli in *Dux* repression is intriguing as the nuclear architecture undergoes rapid reorganization during early embryogenesis and nucleoli become structurally mature during this process [93]. Therefore, one could hypothesize that rRNA biogenesis-induced nucleolar organization may play a role in the rapid shut down of the transcriptional burst of *Dux* during the transition from the 2C to 4C cleavage stage. Interestingly, we detected significant albeit weak enrichment of Ncl (log2 FC of 0.6) specifically in the Lrif1 long isoform interactome mass spec data. Therefore, if this interaction is confirmed, it would be interesting to study the role of Lrif1 in PNH regulation and its link to *Dux* repression.

How to dissect the function of the two LRIF1 isoforms

Since only loss of one LRIF1 isoform is associated with FSHD, this suggests that both isoforms have different functions. The experiments described below, in part already performed, could help to shed light on the function and significance of each of the two LRIF1 isoforms.

First, we determined that somatic loss of either the long isoform or both LRIF1 isoforms does not result in severe genome-wide transcriptional consequences of polyadenylated transcripts, which our RNA-seq analysis was restricted to (**chapter 4**). Knocking out the long LRIF1 isoform in immortalized control myoblasts yields only 10 upregulated genes and one downregulated gene, while a full LRIF1 knock-out leads to 58 upregulated and 21 downregulated genes. This suggests either an additive effect of losing both isoforms or

additional loci regulated specifically by the short LRIF1 isoform. Of note, we also investigated the possibility of using an adenine base editor for mutagenesis of the LRIF1 short isoform start codon (ATG) into a phenylalanine codon (ACG) to interfere with its translation, thus mimicking a knock-out situation. Such approach has been recently published as an alternative method for gene silencing [94]. The resulting phenylalanine is like methionine a hydrophobic amino acid, therefore the effect of this missense mutation in the long isoform might be neutral. Preliminary tests in HEK293T cells were encouraging as the targeted adenine could be converted into guanine on DNA level (Figure 3). However, it still needs to be investigated if such substitution indeed leads to reduced protein production of the short LRIF1 isoform.

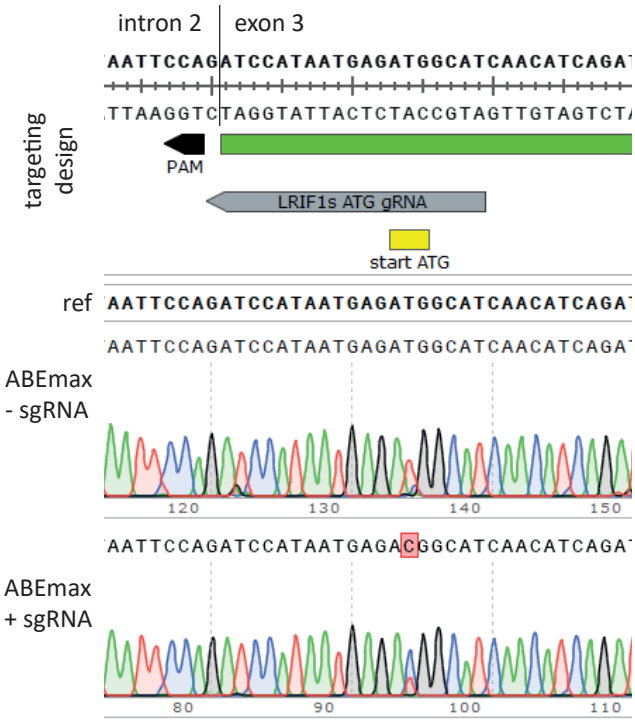


Figure 3. Adenine base editing approach for mutagenesis of the LRIF1 short isoform start codon. Snapgene view of the LRIF1 intron2/exon3 junction with the outlined SpCas9 PAM site (black), sgRNA (grey) and start codon of the LRIF1 short isoform (yellow) (top). Sanger sequencing alignment showing successful editing in HEK293T cells which received both SpABEmax and the sgRNA. The mutated nucleotide is highlighted in red (T -> C).

Second, knowing the genome-wide binding sites of the two isoforms might also help in elucidation of their function and if there are genomic sites which they co-regulate in contrast to the *LRIF1* promoter. Since the enrichment of LRIF1 at D4Z4 was only assessed by using a pan-LRIF1 antibody, it will be important to investigate if both isoforms bind to D4Z4 in control cells and whether they compete for the same binding sites. To circumvent the limitation of available commercial LRIF1 antibodies, we explored CMV promoter-

driven individual overexpression of either N-terminally 3xFlag-tagged or GFP-tagged LRIF1 isoforms by their lentiviral-mediated integration in control muscle cells. While we could achieve expression of the tagged short LRIF1 isoform construct, we were unable to obtain transduced cells that express the tagged long LRIF1 isoform and this was independent of the used tag (unpublished observations). We encountered the same problem when we introduced this lentiviral construct into LRIF1LS^{KO} cells reasoning that re-introduction of the tagged form might be tolerated in these cells. However, it seems that the expression driven by the CMV promoter is too strong and either a weaker promoter should be chosen that would mimic more endogenous-like LRIF1 long isoform expression or an inducible promoter system could be used to control the expression of tagged LRIF1. Nevertheless, this suggests that cells expressing higher amounts of the long LRIF1 isoform for a longer time might be under negative selection pressure. It would be interesting to determine which portion of the N-terminal extended region of the long isoform is responsible for this phenotype.

Third, we determined the Lrif1 isoform specific interactomes by transient overexpression of the GFP-tagged forms in mESCs (**chapter 5**). In contrast to previous identification of only four interacting partners (SMCHD1, HP1α, HP1β and HP1γ) of human LRIF1 in T-REx-293 cells (a HEK293T cell line derivative) [70], our analysis revealed a relatively large number of nuclear proteins (37 proteins enriched in Lrif1s IP and 44 proteins enriched in Lrif1l IP), some of which were specific to individual isoforms. But also in our case, the most enriched proteins were Smchd1 and the three HP1 homologues. This might suggest that these proteins are the core interacting partners of both LRIF1 isoforms, while other identified proteins in our dataset might represent additional interacting partners which might modify the targeting or function of LRIF1. However, the confirmation of their endogenous interaction is pending. We did not identify Dnmt3b as an interacting partner of either Lrif1 isoform. This does not necessarily mean that Dnmt3b does not associate with Lrif1 as the negative result might be also due to the conditions under which we performed our IP. It was shown that different chromatin factors require different salt and enzymatic conditions for their chromatin release [95] as chromatin complexes can be partitioned in distinct biochemical environments [96]. Indeed, different HP1γ-interacting proteins were identified when different nuclear protein extraction methods were used. Interestingly, Dnmt3b was co-purified with HP1γ only when using higher salt concentration (300 mM) in combination with MNase digestion [95], which differed from our IP conditions (only 150 mM salt and no MNase digestion). Therefore, it would be intriguing to test different protein extraction conditions for testing the Lrif1 and Dnmt3b interaction.

Fourth, although the 3D protein structure of LRIF1 has not been experimentally determined yet, the recent development of a novel machine learning approach termed AlphaFold allows for more reliable protein structure prediction from its amino acid sequence [97]. Knowing at least the approximate 3D structure of LRIF1 could facilitate our understanding of the function of its N-terminus. In the AlphaFold database [98], the LRIF1 protein seems rather disorganized with only small local structured domains such as the C-terminal alpha helix which is important for SMCHD1 interaction [70] (Figure 4A, B). In addition, two β-strand

structures are predicted in the long LRIF1 isoform-specific portion of the protein (Figure 4A, B). Interestingly, N-terminally truncated LRIF1 which lacks the first 224 aa (a region that also contains the two aforementioned β-strand structures) showed enhanced binding to HP1α in a yeast two-hybrid assay as compared to its full-length counterpart [70]. It would be of interest to create different deletion mutants of these domains as they might modify the function of the long isoform or its affinity to HP1 decorated genomic regions or might underlie its toxicity when overexpressed. The latter observation is especially intriguing considering the autoregulatory negative feedback loop of the *LRIF1* locus by the long LRIF1 isoform. This further suggests that cells developed a buffering mechanism that ensures only certain levels of the LRIF1 long isoform to be produced by regulating the transcription of the locus as well as by differential splicing. Based on these observations, overexpression of the long LRIF1 isoform would be an unlikely candidate for FSHD therapy as opposed to the previously considered SMCHD1 overexpression.

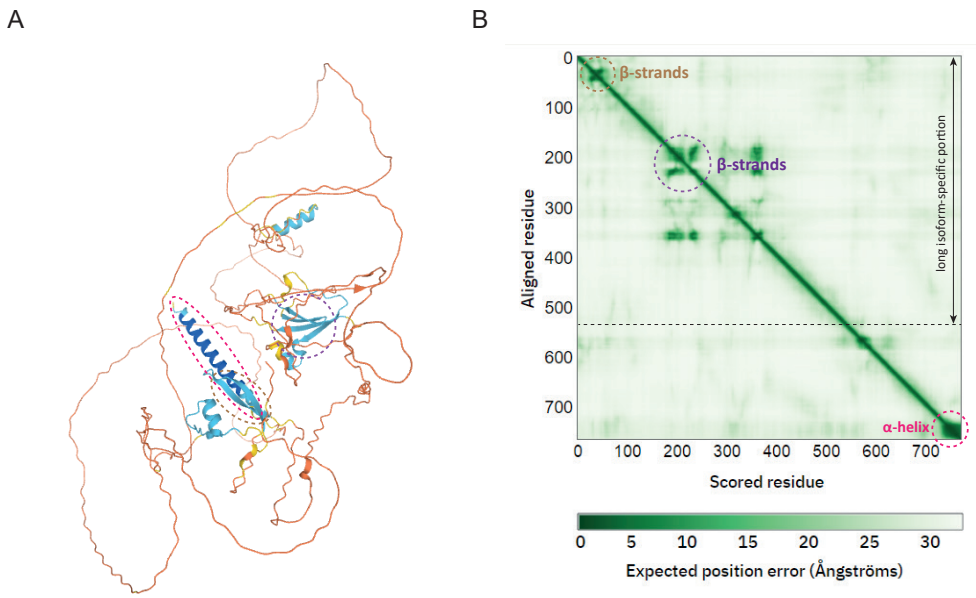


Figure 4. 3D model of the LRIF1 protein structure. A) AlphaFold prediction of the 3D structure of the long LRIF1 isoform. Colors of the domains denote the confidence of modelled structure as predicted by the algorithm (dark blue > light blue > yellow > orange from the most to the least confident prediction). **B)** Heatmap of predicted aligned error of AlphaFold model**,** which aids in assessing inter-domain accuracy. Three domains are circled and corresponding regions in the 3D are marked with the same colors.

Conclusion

The work in this thesis focused on functional studies of one *cis* (*DUX4* PAS) and one *trans* (LRIF1) modifier of *DUX4* expression, both being involved in FSHD pathogenesis. We provided evidence that fine-tuning of the unified genetic model of FSHD is required

to explain *DUX4* expression from the epigenetically de-repressed D4Z4 repeat on a 4qA chromosomal background in skeletal muscles. Furthermore, we showed that similar to DNMT3B, also SMCHD1 and LRIF1 probably contribute to the establishment of the D4Z4 chromatin structure rather than being important for its maintenance in somatic cells. Therefore, creating an *Lrif1* loss-of-function mouse model could provide insight into its function during early development and how its mutations can lead to FSHD. These studies have contributed to our understanding of the molecular mechanisms of *DUX4* regulation and may guide the development of molecular therapies for FSHD.

References

1. Zeng, W. *et al.* Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. *Hum. Mutat.* **35**, 998–1010 (2014).
2. Burgio, G. & Teboul, L. Anticipating and Identifying Collateral Damage in Genome Editing. *Trends Genet.* **36**, 905–914 (2020).
3. Lim, K. R. Q. *et al.* DUX4 Transcript Knockdown with Antisense 2'-O-Methoxyethyl Gapmers for the Treatment of Facioscapulohumeral Muscular Dystrophy. *Mol. Ther.* **29**, 848–858 (2021).
4. Bouwman, L. F. *et al.* Systemic delivery of a DUX4-targeting antisense oligonucleotide to treat facioscapulohumeral muscular dystrophy. *Mol. Ther. - Nucleic Acids* **26**, 813–827 (2021).
5. Lu-Nguyen, N., Malerba, A., Herath, S., Dickson, G. & Popplewell, L. Systemic antisense therapeutics inhibiting DUX4 expression ameliorates FSHD-like pathology in an FSHD mouse model. *Hum. Mol. Genet.* **30**, 1398–1412 (2021).
6. Vanderplanck, C. *et al.* The FSHD Atrophic Myotube Phenotype Is Caused by DUX4 Expression. *PLoS One* **6**, e26820 (2011).
7. Lim, K. R. Q. *et al.* Inhibition of DUX4 expression with antisense LNA gapmers as a therapy for facioscapulohumeral muscular dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 16509–16515 (2020).
8. Chen, J. C. *et al.* Morpholino-mediated Knockdown of DUX4 Toward Facioscapulohumeral Muscular Dystrophy Therapeutics. *Mol. Ther.* **24**, 1405–1411 (2016).
9. Ansseau, E. *et al.* Antisense Oligonucleotides Used to Target the DUX4 mRNA as Therapeutic Approaches in FaciosScapuloHumeral Muscular Dystrophy (FSHD). *Genes (Basel).* **8**, (2017).
10. Marsollier, A.-C. *et al.* Antisense targeting of 3' end elements involved in *DUX4* mRNA processing is an efficient therapeutic strategy for facioscapulohumeral dystrophy: a new gene-silencing approach. *Hum. Mol. Genet.* **25**, 1468–1478 (2016).
11. Wallace, L. M. *et al.* Pre-clinical Safety and Off-Target Studies to Support Translation of AAV-Mediated RNAi Therapy for FSHD. *Mol. Ther. - Methods Clin. Dev.* **8**, 121–130 (2018).
12. Wallace, L. M. *et al.* RNA Interference Inhibits DUX4-induced Muscle Toxicity In Vivo: Implications for a Targeted FSHD Therapy. *Mol. Ther.* **20**, 1417–1423 (2012).
13. Rashnonejad, A., Amini-Chermahini, G., Taylor, N. K., Wein, N. & Harper, S. Q. Designed U7 snRNAs inhibit DUX4 expression and improve FSHD-associated outcomes in DUX4 overexpressing cells and FSHD patient myotubes. *Mol. Ther. - Nucleic Acids* **23**, 476–486 (2021).
14. Himeda, C. L., Jones, T. I. & Jones, P. L. Targeted epigenetic repression by CRISPR/dSaCas9 suppresses pathogenic DUX4-fl expression in FSHD. *Mol. Ther. - Methods Clin. Dev.* **20**, 298–311 (2021).
15. Himeda, C. L., Jones, T. I. & Jones, P. L. CRISPR/dCas9-mediated Transcriptional Inhibition Ameliorates the Epigenetic Dysregulation at D4Z4 and Represses DUX4-fl in FSH Muscular Dystrophy. *Mol. Ther.* **24**, 527–535 (2016).
16. Das, S. & Chadwick, B. P. CRISPR mediated targeting of DUX4 distal regulatory element represses DUX4 target genes dysregulated in Facioscapulohumeral muscular dystrophy. *Sci. Rep.* **11**, 12598 (2021).
17. Lim, J.-W. *et al.* DICER/AGO-dependent epigenetic silencing of D4Z4 repeats enhanced by exogenous siRNA suggests mechanisms and therapies for FSHD. *Hum. Mol. Genet.* **24**, 4817–4828 (2015).
18. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
19. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
20. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nat. 2019 5767785* **576**, 149–157 (2019).
21. Kluesner, M. G. *et al.* CRISPR-Cas9 cytidine and adenosine base editing of splice-sites mediates highly-efficient disruption of proteins in primary and immortalized cells. *Nat. Commun. 2021 121* **12**, 1–12 (2021).
22. Chen, M. *et al.* Systematic evaluation of the effect of polyadenylation signal variants on the expression of disease-associated genes. *Genome Res.* **31**, 890–899 (2021).

23. Koblan, L. W. *et al.* Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
24. Stadler, G. *et al.* Telomere position effect regulates DUX4 in human facioscapulohumeral muscular dystrophy. *Nat. Struct. Mol. Biol. 2013 206* **20**, 671–678 (2013).
25. Joubert, R., Mariot, V., Charpentier, M., Concordet, J. P. & Dumonceaux, J. Gene Editing Targeting the DUX4 Polyadenylation Signal: A Therapy for FSHD? *J. Pers. Med.* **11**, 7 (2020).
26. Safa, A. *et al.* miR-1: A comprehensive review of its role in normal development and diverse disorders. *Biomed. Pharmacother.* **132**, 110903 (2020).
27. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol. 2010 292* **29**, 143–148 (2010).
28. Gasiunas, G. *et al.* A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat. Commun. 2020 111* **11**, 1–10 (2020).
29. Miller, S. M. *et al.* Continuous evolution of SpCas9 variants compatible with non-G PAMs. *Nat. Biotechnol. 2020 384* **38**, 471–481 (2020).
30. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).
31. Richter, M. F. *et al.* Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* **38**, 883–891 (2020).
32. Huang, T. P. *et al.* Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat. Biotechnol.* **37**, 626–631 (2019).
33. Van Der Maarel, S. M. *et al.* De Novo Facioscapulohumeral Muscular Dystrophy: Frequent Somatic Mosaicism, Sex-Dependent Phenotype, and the Role of Mitotic Transchromosomal Repeat Interaction between Chromosomes 4 and 10. *Am. J. Hum. Genet.* **66**, 26–35 (2000).
34. Qiu, L. *et al.* Clinical and genetic features of somatic mosaicism in facioscapulohumeral dystrophy. *J. Med. Genet.* **57**, 777–785 (2020).
35. Dib, C. *et al.* Correction of the FSHD myoblast differentiation defect by fusion with healthy myoblasts. *J. Cell. Physiol.* **231**, 62–71 (2016).
36. Wang, D., Tai, P. W. L. & Gao, G. Adeno-associated virus vector as a platform for gene therapy delivery. *Nat. Rev. Drug Discov. 2019 185* **18**, 358–378 (2019).
37. Duan, D., Yue, Y. & Engelhardt, J. F. Expanding AAV Packaging Capacity with Trans-splicing or Overlapping Vectors: A Quantitative Comparison. *Mol. Ther.* **4**, 383–391 (2001).
38. Bruusgaard, J. C., Liestøl, K., Ekmark, M., Kollstad, K. & Gundersen, K. Number and spatial distribution of nuclei in the muscle fibres of normal mice studied in vivo. *J. Physiol.* **551**, 467–478 (2003).
39. Ryu, S.-M. *et al.* Adenine base editing in mouse embryos and an adult mouse model of Duchenne muscular dystrophy. *Nat. Biotechnol. 2018 366* **36**, 536 (2018).
40. Koblan, L. W. *et al.* In vivo base editing rescues Hutchinson–Gilford progeria syndrome in mice. *Nature* **589**, 608–614 (2021).
41. Levy, J. M. *et al.* Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nat. Biomed. Eng.* **4**, 97–110 (2020).
42. Tabebordbar, M. *et al.* Directed evolution of a family of AAV capsid variants enabling potent muscle-directed gene delivery across species. *Cell* **184**, 4919-4938.e22 (2021).
43. Sarcar, S. *et al.* Next-generation muscle-directed gene therapy by in silico vector design. *Nat. Commun. 2019 101* **10**, 1–16 (2019).
44. Piekarowicz, K. *et al.* A Muscle Hybrid Promoter as a Novel Tool for Gene Therapy. *Mol. Ther. - Methods Clin. Dev.* **15**, 157–169 (2019).
45. Clapp, J. *et al.* Evolutionary Conservation of a Coding Function for D4Z4, the Tandem DNA Repeat Mutated in Facioscapulohumeral Muscular Dystrophy. *Am. J. Hum. Genet.* **81**, 264–279 (2007).
46. Bosnakovski, D. *et al.* Low level DUX4 expression disrupts myogenesis through deregulation of myogenic gene expression. *Sci. Rep.* **8**, 1–12 (2018).

47. Giesige, C. R. *et al.* AAV-mediated follistatin gene therapy improves functional outcomes in the TIC-DUX4 mouse model of FSHD. *JCI Insight* **3**, (2018).
48. Wallace, L. M. *et al.* DUX4, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. *Ann. Neurol.* **69**, 540–552 (2011).
49. Krom, Y. D. *et al.* Intrinsic Epigenetic Regulation of the D4Z4 Macrosatellite Repeat in a Transgenic Mouse Model for FSHD. *PLoS Genet.* **9**, e1003415 (2013).
50. Jones, T. & Jones, P. L. A cre-inducible DUX4 transgenic mouse model for investigating facioscapulohumeral muscular dystrophy. *PLoS One* **13**, e0192657 (2018).
51. Bosnakovski, D. *et al.* Muscle pathology from stochastic low level DUX4 expression in an FSHD mouse model. *Nat. Commun. 2017 81* **8**, 1–9 (2017).
52. Liang, P. *et al.* Genome-wide profiling of adenine base editor specificity by EndoV-seq. *Nat. Commun.* **10**, 67 (2019).
53. Li, J. *et al.* Structure-guided engineering of adenine base editor with minimized RNA off-targeting activity. *Nat. Commun. 2021 121* **12**, 1–8 (2021).
54. Grünewald, J. *et al. CRISPR DNA base editors with reduced RNA off-target and self-editing activities*. *Nature Biotechnology* vol. 37 1041–1048 (Nature Publishing Group, 2019).
55. Zhou, C. *et al.* Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* (2019) doi:10.1038/s41586-019-1314-0.
56. Rees, H. A., Wilson, C., Doman, J. L. & Liu, D. R. Analysis and minimization of cellular RNA editing by DNA adenine base editors. *Sci. Adv.* **5**, (2019).
57. Krom, Y. D. *et al.* Generation of isogenic D4Z4 contracted and noncontracted immortal muscle cell clones from a mosaic patient: A cellular model for FSHD. *Am. J. Pathol.* **181**, 1387–1401 (2012).
58. van der Wal, E. *et al.* Generation of genetically matched hiPSC lines from two mosaic facioscapulohumeral dystrophy type 1 patients. *Stem Cell Res.* **40**, 101560 (2019).
59. Chen, F. & Wilusz, J. Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs. *Nucleic Acids Res.* **26**, 2891–2898 (1998).
60. Peart, N. & Wagner, E. J. A distal auxiliary element facilitates cleavage and polyadenylation of Dux4 mRNA in the pathogenic haplotype of FSHD. *Hum. Genet.* **136**, 1291–1301 (2017).
61. Lemmers, R. J. L. F. *et al.* Worldwide Population Analysis of the 4q and 10q Subtelomeres Identifies Only Four Discrete Interchromosomal Sequence Transfers in Human Evolution. *Am. J. Hum. Genet.* **86**, 364–377 (2010).
62. Lemmers, R. J. L. F. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–3 (2010).
63. Lemmers, R. J. L. F. *et al.* Chromosome 10q-linked FSHD identifies *DUX4* as principal disease gene. *J. Med. Genet.* jmedgenet-2020-107041 (2021) doi:10.1136/jmedgenet-2020-107041.
64. Choi, P. S. & Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas technology. *Nat. Commun. 2014 51* **5**, 1–6 (2014).
65. Komor, A. C. *et al.* Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci. Adv.* **3**, (2017).
66. Hamanaka, K. *et al.* Clinical, muscle pathological, and genetic features of Japanese facioscapulohumeral muscular dystrophy 2 (FSHD2) patients with SMCHD1 mutations. *Neuromuscul. Disord.* **26**, 300–308 (2016).
67. Choi, J. *et al.* MyoD converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc. Natl. Acad. Sci.* **87**, 7988–7992 (1990).
68. Balog, J. *et al.* Monosomy 18p is a risk factor for facioscapulohumeral dystrophy. *J. Med. Genet.* **55**, 469–478 (2018).
69. Balog, J. *et al.* Increased DUX4 expression during muscle differentiation correlates with decreased SMCHD1 protein levels at D4Z4. *Epigenetics* **10**, 1133–1142 (2015).
70. Nozawa, R.-S. *et al.* Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* **20**, 566–573 (2013).

71. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nat. 2014 5077490* **507**, 104–108 (2014).
72. Hamanaka, K. *et al.* Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology* **94**, e2441–e2447 (2020).
73. Lemmers, R. J. L. F. *et al.* Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat. Genet.* **44**, 1370–1374 (2012).
74. Goossens, R. *et al.* Intronic SMCHD1 variants in FSHD: Testing the potential for CRISPR-Cas9 genome editing. *J. Med. Genet.* **56**, 828–837 (2019).
75. Bosch-Presegué, L. *et al.* Mammalian HP1 Isoforms Have Specific Roles in Heterochromatin Structure and Organization. *Cell Rep.* **21**, 2048–2057 (2017).
76. Chen, K. *et al.* Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3535-44 (2015).
77. Wong, X. *et al.* Mapping the micro-proteome of the nuclear lamina and lamina-associated domains. *Life Sci. Alliance* **4**, (2021).
78. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
79. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
80. Huang, Z. *et al.* The chromosomal protein SMCHD1 regulates DNA methylation and the 2c-like state of embryonic stem cells by antagonizing TET proteins. *Sci. Adv.* **7**, eabb9149 (2021).
81. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
82. Rodriguez-Terrones, D. *et al.* A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119 (2018).
83. Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391-405.e19 (2018).
84. Thompson, P. J. *et al.* hnRNP K Coordinates Transcriptional Silencing by SETDB1 in Embryonic Stem Cells. *PLOS Genet.* **11**, e1004933 (2015).
85. Peng, H. *et al.* Reconstitution of the KRAB-KAP-1 repressor complex: a model system for defining the molecular anatomy of RING-B box-coiled-coil domain-mediated protein-protein interactions. *J. Mol. Biol.* **295**, 1139–1162 (2000).
86. Ryan, R. F. *et al.* KAP-1 Corepressor Protein Interacts and Colocalizes with Heterochromatic and Euchromatic HP1 Proteins: a Potential Role for Krüppel-Associated Box-Zinc Finger Proteins in Heterochromatin-Mediated Gene Silencing. *Mol. Cell. Biol.* **19**, 4366–4378 (1999).
87. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16**, 919–932 (2002).
88. Schultz, D. C., Friedman, J. R. & Rauscher, F. J. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2α subunit of NuRD. *Genes Dev.* **15**, 428–443 (2001).
89. Ivanov, A. V. *et al.* PHD Domain-Mediated E3 Ligase Activity Directs Intramolecular Sumoylation of an Adjacent Bromodomain Required for Gene Silencing. *Mol. Cell* **28**, 823–837 (2007).
90. Zeng, L. *et al.* Structural insights into human KAP1 PHD finger–bromodomain and its role in gene silencing. *Nat. Struct. Mol. Biol. 2008 156* **15**, 626–633 (2008).
91. Yang, B. X. *et al.* Systematic Identification of Factors for Provirus Silencing in Embryonic Stem Cells. *Cell* **163**, 230–245 (2015).
92. Yu, H. *et al.* rRNA biogenesis regulates mouse 2C-like state by 3D structure reorganization of peri-nucleolar heterochromatin. *Nat. Commun. 2021 121* **12**, 1–21 (2021).

93. Borsos, M. & Torres-Padilla, M. E. Building up the nucleus: nuclear organization in the establishment of totipotency and pluripotency during mammalian development. *Genes Dev.* **30**, 611–621 (2016).

94. Wang, X. *et al.* Efficient Gene Silencing by Adenine Base Editor-Mediated Start Codon Mutation. *Mol. Ther.* **28**, 431–440 (2020).

95. Zaidan, N. Z. *et al.* Compartmentalization of HP1 Proteins in Pluripotency Acquisition and Maintenance. *Stem Cell Reports* **10**, 627–641 (2018).

96. Becker, J. S. *et al.* Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol. Cell* **68**, 1023-1037.e15 (2017).

97. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nat. 2021 5967873* **596**, 583–589 (2021).

98. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* (2021) doi:10.1093/NAR/GKAB1061.

APPENDIX

English Summary

Nederlandse Samenvatting

List of Publications

Curriculum Vitae

Acknowledgements

English Summary

Proper spatiotemporal regulation of gene expression is crucial for organismal development and functioning. Misexpression of genes outside of their natural cellular or tissue context can have profound consequences. One such example is the misexpression of the *DUX4* gene. DUX4 belongs to a class of pioneer transcription factors which can initiate entire gene network changes. In this manner, DUX4 stimulates the process of zygotic genome activation during the 4-cell cleavage stage in humans by triggering the expression of appropriate genes and repeats. Its expression is, therefore, restricted to a very narrow time window during the embryonic cleavage stage. After this period, *DUX4* expression is attenuated for the rest of one's life in the majority of somatic cells/tissues. However, this silencing process is incomplete in some individuals leading to aberrant *DUX4* expression in skeletal muscle triggering, amongst others, a similar embryonic transcriptional program. This has pathological consequences for the muscles and results in a specific type of muscular dystrophy known as facioscapulohumeral muscular dystrophy (FSHD).

The mechanism of *DUX4* repression in somatic cells is governed by the organization of its endogenous locus. The DUX4 open reading frame is repeated usually between ten and hundred times and forms a macrosatellite repeat structure called D4Z4. A partial loss of somatic D4Z4 repression can result either from *in cis* genetic changes (shortening of the repeat to less than 10 units) or *in trans* genetic changes (mutations in D4Z4 chromatin repressors), conditions that result in misexpression of *DUX4* in skeletal muscle. A D4Z4 repeat contraction is found in the majority of FSHD cases (FSHD1), while mutations in D4Z4 chromatin repressors (FSHD2) is rather rare (<5% of FSHD cases). Most often, FSHD2 individuals have heterozygous mutations in the *SMCHD1* gene. In addition, mutations in *DNMT3B* have been also associated with FSHD2, although they are found less frequently than *SMCHD1* mutations. Loss of function mutations in both genes lead to the loss of heterochromatinization of D4Z4 repeat which is marked by CpG hypomethylation and changes in the histone modifications profile. Apart from a transcriptionally permissive chromatin environment, *DUX4* expression also requires a nearby polyadenylation signal (PAS) for its proper post-transcriptional mRNA processing. This PAS sequence lies adjacent to the D4Z4 repeat and is present only at specific 4q subtelomeric variant known as 4qA. In chapter 1, we provide a comprehensive introduction of epigenetics, repeat biology and of our current understanding of FSHD pathogenesis.

In chapter 2, we tested if elimination of the functional *DUX4* PAS found in 4qA alleles would provide a new opportunity to impair *DUX4* expression and thus could be utilized as a genetic therapeutic target in FSHD. For this, we capitalized on the adenine-rich nature of the 4qA *DUX4* PAS sequence (ATTAAA) and used an adenine base editor for the mutagenesis of its three 3' adenines into guanines. We showed that this approach is feasible in immortalized myoblasts derived from three FSHD-affected individuals and that successfully edited cells do indeed produce less polyadenylated *DUX4* transcript. This also translated to reduced expression of DUX4 transcriptional target genes. Furthermore, we showed that mutagenesis

of the *DUX4* PAS by this approach leads to a switch in the cleavage and polyadenylation sites used by residual *DUX4* mRNA species further corroborating a functional defect of the mutated PAS. In conclusion, we showed that adenine base editing of a gene's PAS can be used to achieve efficient gene silencing and that ~25% of all PASs with either AATAAA or ATTAAA motifs are potential candidates for such downregulation strategy.

In chapter 3, we expand the list of FSHD disease genes by identifying an individual with a clinical presentation consistent with FSHD who is a carrier of a homozygous out-of-frame mutation (c.869_872dup) in the *LRIF1* gene. The identified mutation leads to the loss of only one of the two LRIF1 isoforms. Similar to FSHD2 individuals with either *SMCHD1* or *DNMT3B* mutations, cells of this individual also showed pronounced hypomethylation of the D4Z4 repeats and known FSHD-associated changes in D4Z4 histone modifications. Furthermore, after transdifferentiating the proband's fibroblasts into myogenic cells, we detected expression of *DUX4* together with some of the tested DUX4 target genes confirming that the observed epigenetic changes at the locus have transcriptional consequences. Interestingly, LRIF1 was already shown to interact with SMCHD1 suggesting that both proteins might cooperate in D4Z4 silencing. Indeed, we observed that LRIF1 is enriched at D4Z4 in primary myogenic cells from unaffected individuals and that the amount of SMCHD1 is reduced in proband's cells. Furthermore, downregulating endogenous expression of the long LRIF1 isoform in control, FSHD1 as well as FSHD2 myogenic cells further upregulates expression of *DUX4* thus showing modulatory role of LRIF1 in D4Z4 repression in somatic cells.

In chapter 4, we further explored the function of LRIF1 together with its interacting partner SMCHD1 in somatic D4Z4 silencing. We generated knock-out situations for both genes using CRISPR/Cas9 in two independent control immortalized myocytes. However, knock-out of either factor showed only minor transcriptional consequences for *DUX4.* We further showed that this can be explained by the lack of changes in D4Z4 chromatin conformation as investigated by measuring the DNA methylation levels and levels of specific histone modifications at the D4Z4 repeat, which are known to be affected in FSHD2 cells, while remaining unaffected in somatic knock-out cells. We also established the hierarchy of LRIF1 and SMCHD1 recruitment to D4Z4 in somatic cells, with SMCHD1 mediating the recruitment of LRIF1 to D4Z4 but not vice versa. Furthermore, binding of both proteins to D4Z4 is affected in cells derived from FSHD2- or ICF1-affected individuals which carry either monoallelic or biallelic *DNMT3B* mutations and have a compromised D4Z4 chromatin structure. Indeed, in these cells the D4Z4 repeat shows typical FSHD2-related chromatin changes (i.e. DNA hypomethylation, decreased H3K9me3 and increased H3K4me2 and H3K27me3 levels), suggesting that SMCHD1 and LRIF1 recruitment to D4Z4 repeat is sensitive to one or multiple chromatin changes imposed by insufficient DNMT3B-mediated methylation deposition in early development. Furthermore, the fact that a mutation in any of three genes (*SMCHD1*, *DNMT3B*, and *LRIF1*) leads to the same epigenetic D4Z4 characteristics in somatic cells suggests their co-dependency during the establishment of the epigenetic state of D4Z4. Lastly, we also uncovered an autoregulatory feedback loop for the *LRIF1* locus imposed

specifically by its long isoform, when together with SMCHD1 it binds to *LRIF1* promoter and by unknown mechanism represses transcription of the locus.

In chapter 5, we switched from human somatic cells to mouse embryonic stem cells (mESCs) to study the function of Lrif1 in the context of early development. Transient depletion of Lrif1 in mESCs leads to upregulation of a 2-cell/4-cell-like transcriptional program characterized by the expression of genes and repetitive elements driven by the mouse homologue of DUX4, Dux. Investigating the protein interactome of the two Lrif1 isoforms in mESCs further revealed their interaction with Trim28, a known repressor of the *Dux* locus in mESCs. Mechanistically, Trim28 mediates the deposition of H3K9me3 at *Dux* to sustain its silencing and the loss of Lrif1 resulted in reduced occupancy of Trim28 at *Dux*, which was accompanied by reduced H3K9me3 levels. However, the latter effect could be contributed to the reduced H3 levels at *Dux* which suggests broader nucleosome depletion from this locus leading to *Dux* de-repression. These findings elucidate the conserved function of LRIF1 in silencing the genomic locus of a pioneer transcription factor involved in zygotic genome activation.

Finally, in chapter 6, we discussed our findings in chapters 2-5 in a broader context and provide suggestions for future studies.

Nederlandse Samenvatting

Een correcte tijdruimtelijke regulering van genexpressie is essentieel voor een goede ontwikkeling en voor het goed functioneren van een organisme. De misexpressie van genen buiten hun natuurlijke cellulaire of weefsel-context kan ernstige gevolgen hebben voor het organisme. De misexpressie van DUX4 is een goed voorbeeld hiervan. DUX4 behoort tot de pionier transcriptiefactoren, een klasse van transcriptiefactoren die de expressie van gehele gen-netwerken kan aansturen. In deze hoedanigheid stimuleert DUX4 de zogenaamde zygotische genoomactivatie tijdens de eerste klievingsdelingen van de bevruchte eicel door de expressie van een specifieke set van genen en gerepeteerde sequenties te activeren. De expressie van *DUX4* is slechts beperkt tot een kort tijdsinterval gedurende deze klievingsdelingen. Daarna wordt *DUX4* expressie onderdrukt in de meeste somatische weefsels en cellen gedurende de rest van het leven. In sommige mensen is deze onderdrukking van DUX4 incompleet hetgeen leidt tot *DUX4* misexpressie in skeletspieren en de activatie van DUX4-gevoelige genen die normaal alleen in de vroege embryogenese worden geactiveerd door DUX4. De activatie van dit DUX4 programma in de spieren veroorzaakt de spierziekte facioscapulohumerale spierdystrofie (FSHD).

Het mechanisme waarmee de expressie van *DUX4* wordt onderdrukt in somatische cellen is grotendeels afhankelijk van de organisatie van het endogene *DUX4* locus zelf. Het DUX4 open leesraam is 10 tot 100x achter elkaar gerepeteerd aanwezig waardoor het een zogenaamde macrosatelliet repeat vormt die we de D4Z4 repeat noemen. De gedeeltelijke derepressie van de D4Z4 repeat in somatische cellen wordt enerzijds veroorzaakt door veranderingen aan de D4Z4 repeat zelf (een verkorting van de repeat tot 1-10 eenheden) of door mutaties in genen die coderen voor D4Z4 chromatinefactoren die bijdragen aan een repressieve D4Z4 chromatine structuur. In beide gevallen leiden de veranderingen aan de D4Z4 chromatinestructuur tot de aanwezigheid van DUX4 in de spier. FSHD wordt meestal veroorzaakt door het eerste mechanisme, een verkorting van de D4Z4 repeat (FSHD1; >95%), terwijl mutaties in D4Z4 chromatinefactoren zeldzamer zijn (FSHD2; <5%). Vaak hebben FSHD2 patiënten een heterozygote mutatie in de D4Z4 chromatinefactor SMCHD1, maar er zijn ook enkele patiënten beschreven met heterozygote mutaties in de chromatinefactor DNMT3B. In beide gevallen leiden deze mutaties tot een gedeeltelijk verlies van de D4Z4 heterochromatine structuur in somatische cellen, gemarkeerd door CpG hypomethylatie en veranderingen in het histon-modificatieprofiel van D4Z4. Naast deze gedeeltelijke opening van de D4Z4 chromatinestructuur is ook de aanwezigheid van een polymorf *DUX4* polyadenyleringssignaal (PAS) essentieel voor *DUX4* expressie in somatische cellen. Deze PAS bevindt zich direct achter de D4Z4 repeat en is alleen aanwezig op een specifieke genetische achtergrond van chromosoom 4 die we 4qA noemen. In hoofdstuk 1 geven we een uitgebreide introductie in epigenetica, repeat biologie en onze kennis over FSHD.

In hoofdstuk 2 hebben we onderzocht of het verwijderen van de DUX4 PAS uit 4qA allelen een nieuwe kans biedt om DUX4 expressie in de spieren te voorkomen en dus of dit principe kansen biedt voor gentherapie in FSHD. Om dit te doen maakten we gebruik van het feit

dat de DUX4 PAS sequentie (ATTAAA) veel adenines bevat die kunnen worden omgezet in guanines met gebruik van een adenine base editor. We hebben de haalbaarheid van deze aanpak aangetoond in geïmmortaliseerde spiercellijnen van drie FSHD patiënten waarin wij succesvol de *DUX4* PAS konden editen en daarmee de expressie van *DUX4* en zijn netwerkgenen grotendeels konden voorkomen. Daarnaast lieten we zien dat nog aanwezige *DUX4* mRNA producten gebruik maakten van andere, nabijgelegen, cleavage en polyadenyleringssignalen in het genoom hetgeen bevestigde dat de oorspronkelijke *DUX4* PAS niet meer herkend wordt. In een bredere, genoomwijde context kwamen we tot de conclusie dat ~25% van alle AATAAA of ATTAAA polyadenyleringssignalen in ons genoom in principe gevoelig zijn voor deze gentherapie technologie.

In hoofdstuk 3 breiden we de lijst van FSHD ziektegenen verder uit met de beschrijving van een FSHD patiënt met een homozygote open leesraam-verstorende mutatie in *LRIF1* (c.869_872dup). Deze mutatie leidt tot het verlies van één van de twee LRIF1 isovormen (de lange isovorm). Net als FSHD2 patiënten met mutaties in SMCHD1 of DNMT3B laten cellen van deze patiënt sterke D4Z4 hypomethylatie zien alsmede FSHD-bekende veranderingen in het D4Z4 histon-modificatieprofiel. Transdifferentiatie van huidcellen van deze patiënt in spiercellen gaf bewijs voor de misexpressie van DUX4 en zijn netwerkgenen hetgeen bevestigde dat de waargenomen veranderingen in de D4Z4 chromatinestructuur kunnen leiden tot DUX4 expressie in de spier. LRIF1 was eerder geïdentificeerd als partner van SMCHD1 hetgeen suggereert dat beide wellicht samenwerken in het voorkomen van *DUX4* expressie in somatische cellen. Wij konden inderdaad aantonen dat LRIF1 aan de D4Z4 repeat bindt in gezonde spiercellen en dat er minder SMCHD1 aanwezig is op D4Z4 in de huidcellen van de patiënt. Tenslotte toonden we aan dat het verminderen van de lange isovorm van LRIF1 in spiercellen van controle, FSHD1 en FSHD2 individuen leidt tot *DUX4* opregulatie, hetgeen de modulerende rol van LRIF1 op *DUX4* expressie bevestigt.

In hoofdstuk 4 hebben we de rol van LRIF1 en zijn partner SMCHD1 in *DUX4* onderdrukking in somatische cellen verder onderzocht. Met behulp van CRISPR/Cas9 hebben we knockout condities gemaakt voor beide genen in twee onafhankelijke spiercellijnen. Echter, knockout voor elk van deze factoren leidde slechts tot minimale transcriptionele veranderingen van *DUX4*. We toonden aan dat dit kan worden verklaard door een gebrek aan FSHD-specifieke veranderingen aan de D4Z4 chromatinestructuur voor wat betreft DNA methylatie en histonmodificaties. We hebben ook vastgesteld dat SMCHD1 de binding van LRIF1 aan D4Z4 faciliteert in somatische cellen maar dat dit omgekeerd niet het geval is. Bovendien is de binding van beide eiwitten aan D4Z4 verminderd in cellen van ICF1 of FSHD2 patiënten met mutaties in DNMT3B waarin de D4Z4 chromatinestructuur gecompromitteerd is met FSHD2-herkenbare veranderingen (DNA hypomethylatie, verlies van H3K9me3 en toename in H3K4me2 en H3K27me3). Dit suggereert dat de binding van SMCHD1 en LRIF1 aan de D4Z4 repeat gevoelig is voor D4Z4 chromatinestructuur veranderingen die voortvloeien uit verminderde D4Z4 DNA methylatie door DNMT3B gedurende de vroege embryogenese. De constatering dat een mutatie in elk van de drie FSHD2 genen (*SMCHD1*, *DNMT3B*, en *LRIF1*)

tot dezelfde veranderingen in de D4Z4 chromatinestructuur leiden suggereert bovendien een onderlinge afhankelijkheid van deze drie chromatinefactoren tijdens de aanleg van de D4Z4 chromatine structuur. Tenslotte vonden we bewijs voor een autoregulatoire feedback loop voor het LRIF1 locus waarin de lange LRIF1 isovorm samen met SMCHD1 bindt aan de *LRIF1* promoter om expressie van het locus te onderdrukken.

In hoofdstuk 5 verwisselden we humane somatische cellen voor muis embryonale stamcellen (mESCs) om de rol van Lrif1 te kunnen bestuderen in de vroege embryogenese. Transiente verlaging van Lrif1 expressieniveaus in mESCs veroorzaakt een opregulatie van het 2-cel-/4-cel-achtige transcriptionele programma dat wordt gekarakteriseerd door de expressie van genen en repetitieve sequenties die worden geactiveerd dor de muishomoloog van DUX4, Dux. Door het bestuderen van het eiwit-interactoom van beide Lrif1 isovormen konden we een interactie met Trim28, een bekende onderdrukker van het *Dux* locus, vaststellen. Trim28 faciliteert de aanmaak van H3K9me3 op het Dux locus om transcriptie te onderdrukken en het verlies van Lrif1 leidt tot verminderde Trim28 binding aan *Dux*, vergezeld door verminderde H3K9me3 waarden. Dit laatste kan echter ook worden verklaard door verminderde H3 niveaus in het *Dux* locus hetgeen een meer open Dux chromatinestructuur suggereert. Deze bevindingen suggereren een geconserveerde functie voor LRIF1 in de transcriptionele onderdrukking van een pionier transcriptiefactor betrokken bij zygotische genoom activatie.

In hoofdstuk 6, tenslotte, bespreken we de bevindingen uit hoofstukken 2-5 in een bredere context en doen we suggesties voor toekomstig onderzoek.

List of Publications

1. **Šikrová D**, González-Prieto R, Vertegaal ACO, Balog J, Clemens-Daxinger L & van der Maarel SM. Lrif1 aids in Trim28-mediated repression of *Dux* repeat in mouse embryonic stem cells. *in preparation*
2. **Šikrová D**, Testa AM, Willemsen I, van den Heuvel A, Tapscott SJ, Daxinger L, Balog J and van der Maarel SM. Locus-specific differences in chromatin recruitment of FSHD2 gene products SMCHD1 and LRIF1. *Commun Biol. under review*
3. **Šikrová D**, Cadar AV, Ariyurek Y, Laros J, Balog J & van der Maarel SM. Adenine base editing of the *DUX4* polyadenylation signal for targeted genetic therapy in Facioscapulohumeral muscular dystrophy. *Mol Ther Nucleic Acids.* 2021
4. Hamanaka K*, **Šikrová D***, Mitsuhashi S, Masuda H, Sekiguchi Y, Sugiyama A, Shibuya K, Lemmers RJLF, Goossens R, Ogawa M, Nagao K, Obuse C, Noguchi S, Hayashi YK, Kuwabara S, Balog J, Nishino I & van der Maarel SM. A homozygous nonsense variant in *LRIF1* associated with Facioscapulohumeral muscular dystrophy. *Neurology*. 2020
5. Herrmannová A, Prilepskaja T, Wagner S, **Šikrová D**, Zeman J, Poncová K & Valášek LS. Adapted formaldehyde gradient cross-linking protocol implicates human eIF3d and eIF3c, k and l subunits in the 43S and 48S pre-initiation complex assembly, respectively. *Nucleic Acids Res.* 2020
6. Wagner S*, Herrmannová A*, **Šikrová D*** & Valášek LS. Human eIF3b and eIF3a serve as the nucleation core for the assembly of eIF3 into two interconnected modules: yeast-like core and the octamer. *Nucleic Acids Res.* 2016

*equal contribution

Curriculum Vitae

Darina Šikrová was born on 20th February 1992 in Michalovce, Slovakia, where she attended primary as well as high school. She started her Bachelor in Biology at Comenius University in Bratislava, Slovakia in 2010. In the last year of her Bachelor program, she joined the lab of Prof. Ľubomír Tomáška to conduct her bachelor thesis studying the yeast telomere biology. She graduated with a Bachelor degree in 2013 after which she moved to Prague to start her Master degree in Genetics, Molecular Biology and Virology at Charles University. She completed her master internship in the lab of Dr. Leoš Valášek studying the process of eukaryotic translation initiation and dissecting the role and complex assembly of the largest translation initiation factor – eIF3. Furthermore, she received an Erasmus scholarship to join for five months a lab of Dr. Luisa Cochella at the Institute of Molecular Pathology in Vienna, where she studied the role of neuronal miRNAs in avoidance behavior of *C. elegans*. After completing her master degree in 2015, she stayed in the lab of Dr. Valášek for additional year as a research technician and continued on her master internship project, which culminated in one co-first author and one co-author publication.

In November 2016, she started a PhD at Leiden University Medical Centre (LUMC) in the group of Prof. Silvère van der Maarel studying the genetic and epigenetic aspects of facioscapulohumeral muscular dystrophy, an (epi)genetic disease in which failure to silence of one repetitive element in muscle cells causes muscle wasting. Apart from conducting research, she also co-organized an MGC PhD workshop on Texel island in 2018 for her fellow PhD students from Leiden and Rotterdam and supervised two master students. The results of her doctoral research are described in this thesis.

After leaving the lab of Prof. van der Maarel, she joined for half a year the genomics core facility of LUMC (LGTC) led by Dr. Susan Kloet in order to gain more experience in next-generation sequencing. At LGTC, she worked on implementing a protocol for plate-based combined profiling of genome and transcriptome from single cells. Since July 2022, she works as an R&D Scientist at Single Cell Discoveries, where she is extending on her research experience from LGTC by working on developing and optimizing single cell sequencing methods.

Acknowledgements

It took me 50% longer of what I originally I aimed for but finally and thanks to many, this is it. First and foremost, I would like to thank my promoter, Silvère, for giving me the opportunity to undertake my PhD in his group. I appreciate that you not only supported me during my time spent in academia but also offered your support during my time deciding what to do afterwards and thanks to you I got to spend half a year at LGTC broadening my skillset, which eventually impacted my future path. Judit, you persuaded me to accept the PhD offer and encouraged me to continue with it when I had doubts. I will also never forget your hospitality and that you opened your home to me when I first arrived to the Netherlands and didn't have a place to stay at. I would also like to thank all the past and current members of the FSHD group that I crossed path with (Richard, Patrick, Anita, Bianca, Jessica, Dongxu, Marnix, Iris, Linde, Erik, Amanda, David, Muriel, Kirsten, Marlinde), who helped me during my PhD not only in a scientific but also non-scientific way. Special thanks to Mara and Remko for co-creating our wining & whining sessions and with whom I got to share both the frustrations and fun moments of PhD life. I would also like to thank two of my very first students, Alessandra and Vlad, who posed a great challenge for me as I had to switch from being a student myself to having to teach someone else, and who bravely endured my undefined supervision style.

I would like to thank also another group, which had a huge impact on my PhD - DevEpi group. Lucia, thank you for all your critical comments that pushed me to think more about my research questions and approaches how to probe them. Thanks to also all other (past and present) members of DevEpi group (Cor, Kelly, Jihed, Veronica, Haoyu) who shared their knowledge and feedback initially during our joined Monday group discussions but later also whenever I needed it. And special thanks to Maja. What initially started as a work collaboration in the end turned into a nice friendship also outside of the LUMC walls, which I am very grateful for.

Thank you Conny and Jill for helping me to get into the world of culturing mESCs, Susan and the whole LGTC team (Yavuz, Emile, Rolf, Loes, Roberta) for welcoming me for half a year to be a part of your team and the rest of the Department of Human Genetics with unfortunately too many people to name, but who created a nice working environment for me during all those years.

I would also like to thank the members of my PhD supervisory committee, Marcel Tijsterman, Annemieke Aartsma-Rus and Frank Baas for taking your time and contributing your suggestions during my yearly appraisal meetings.

Finally, I would like to thank my dad for always being there for me and reminding me to keep a perspective in life and Lukáš, for supporting me to even start my PhD journey and giving me an extra push at the end to finish it when I needed it.

A