Cis and trans modifiers in facioscapulohumeral muscular dystrophy

Sikrová, D.

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.

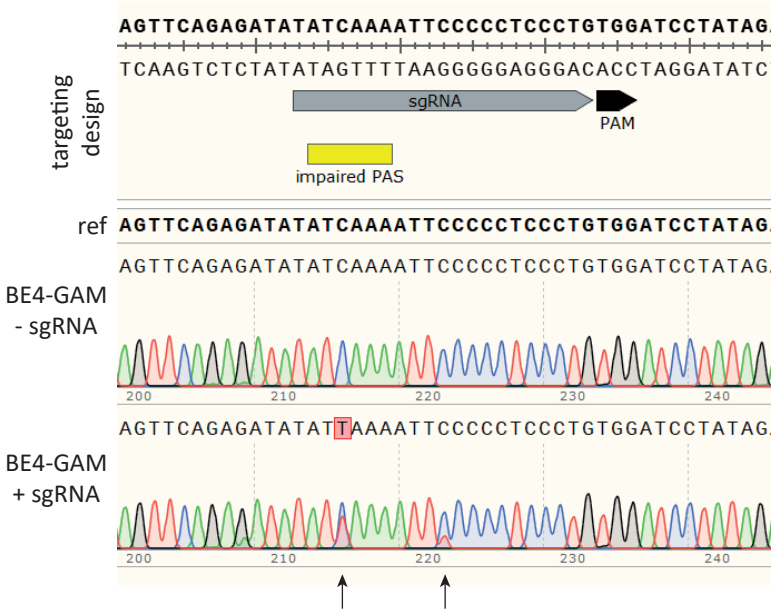

Figure 1. Cytidine base editing approach for converting the 10qA *DUX4* PAS SNP into a 4qA *DUX4* PAS SNP. Snapgene view of a section of the pLAM region in the 10qA D4Z4 allele with outlined the SpCas9 PAM site (black), the sgRNA (grey) and the impaired *DUX4* PAS sequence (yellow). Sanger sequencing 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.

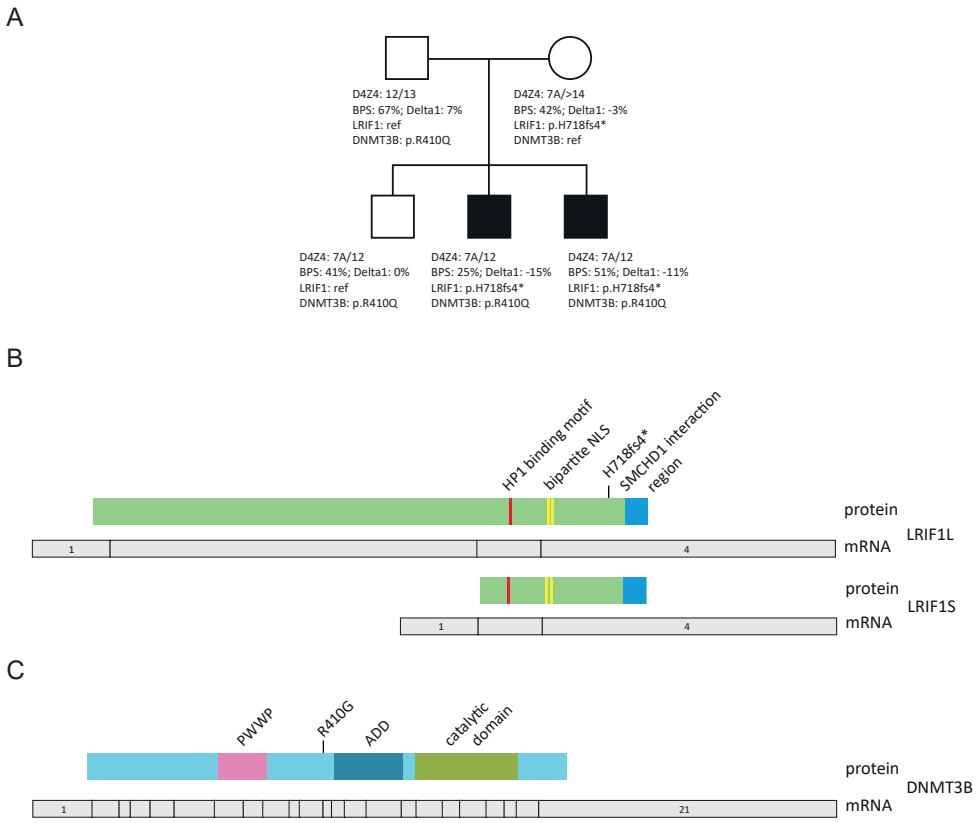

Figure 2. FSHD potentially caused by combined heterozygous *LRIF1* and *DNMT3B* variants. A) Pedigree of the family with information about D4Z4 sizing and its allelic type (A/B) if known, D4Z4 methylation (as determined by bisulfite PCR of DR1 region (BPS) and Delta1 score, which represents a difference 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.

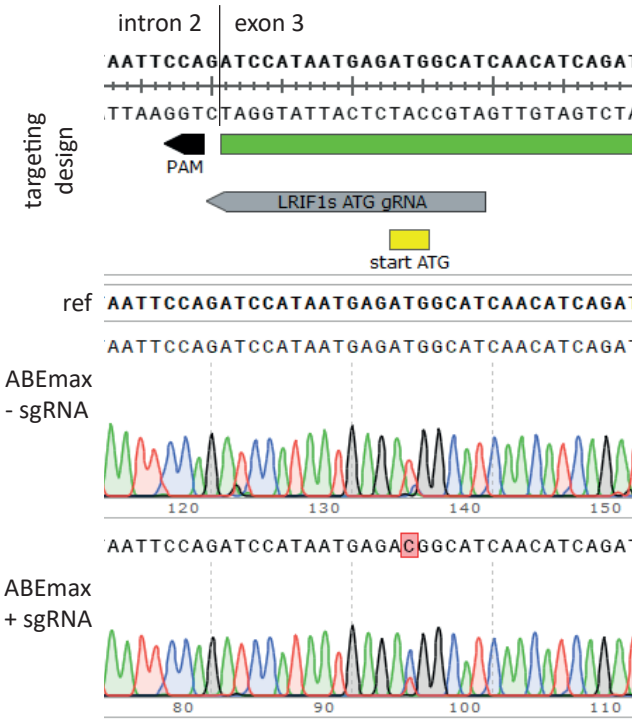

Figure 3. Adenine base editing approach for mutagenesis of the LRIF1 short isoform start codon. Snapgene view of the LRIF1 intron2/exon3 junction with the outlined SpCas9 PAM site (black), sgRNA (grey) and start codon of the LRIF1 short isoform (yellow) (top). Sanger sequencing alignment showing successful 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.

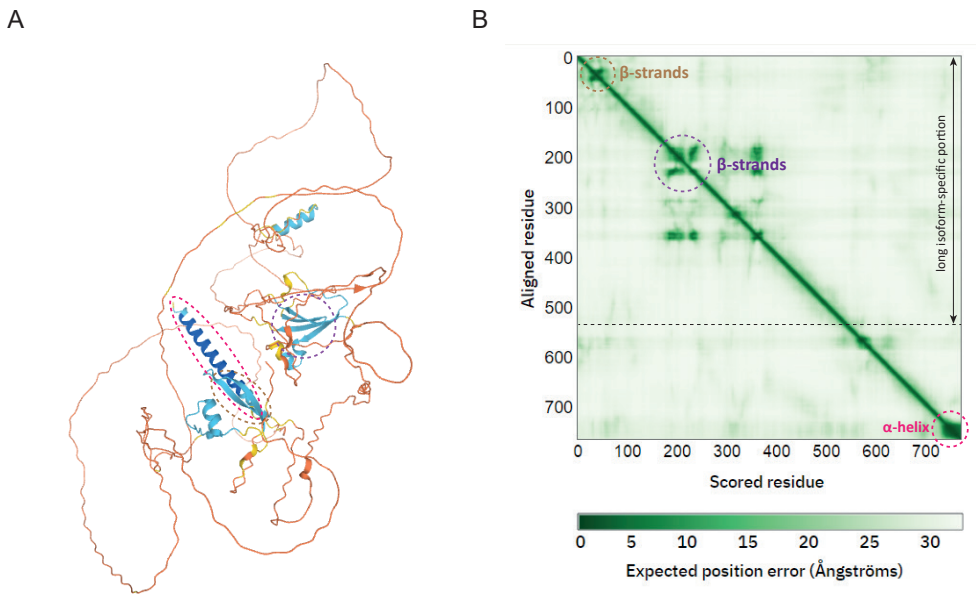

Figure 4. 3D model of the LRIF1 protein structure. A) AlphaFold prediction of the 3D structure of the long LRIF1 isoform. Colors of the domains denote the confidence of modelled structure as predicted by the algorithm (dark blue > light blue > yellow > orange from the most to the least confident prediction). **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.