Rearrangements within the facioscapulohumeral muscular dystrophy locus: mechanism, timing and consequences.

Lemmers, R.

Chapter 1.

Introduction

1.1. Facioscapulohumeral Muscular Dystrophy

1.1.1 Clinical characteristics

Facioscapulohumeral muscular dystrophy (FSHD1A, MIM 158900) is an autosomal dominant myopathy with an average age at onset in the second decade of life. Landouzy and Dejerine initially reported the disease in 1885.[56] With an incidence of 1:20,000 it is the third most common inherited neuromuscular disorder after Duchenne and myotonic dystrophy. FSHD is initially characterized by facial muscle weakness. During progression, weakness and atrophy of shoulder girdle muscles is observed in almost all cases. A gradual spread to abdominal and foot-extensor muscles, followed by clinical involvement of upper arm and pelvic girdle muscles, is seen in the majority of patients.[86] For many FSHD patients, asymmetry of muscle involvement is reported. Also, mild, often subclinical, sensorineural deafness, retinal vasculopathy and, in severe cases, mental retardation can be part of the disease.[9,10,24,25,28,77,113] FSHD displays considerable inter- and even intrafamilial variability, with severity ranging from almost asymptomatic to wheelchair-depency.[87] Furthermore, male FSHD patients are somewhat more affected than female patients.[151]

1.1.2 Genomic localization

Genomic localization of the FSHD locus was seriously hampered by the clinical variability and the phenotypic overlap of FSHD with other muscle dystrophies like the limb girdle muscular dystrophies (LGMD), Becker muscular dystrophy (BMD) and proximal myotonic myopathy (PROMM), sometimes leading to incorrect clinical counseling. The accurate definition of the clinical features of FSHD allowed the unambiguous identification of sporadic and familial FSHD cases.[86] Subsequently, DNA of several multigenerational FSHD families showing an autosomal dominant pattern of inheritance was collected and used for linkage analysis. A genome-wide scan using micro- and minisatellite markers for these families showed linkage to D4S171 on chromosome 4.[136] Further linkage analysis with additional markers linked FSHD distal to D4S139 (region 4q35–4qter)[76,97,132,139] and subsequently revealed the markers D4F104S1 (p13E-11) and D4Z4 that were directly associated with the FSHD locus.[138]

1.1.3 Causal molecular defect

Probe p13E-11 identifies a highly polymorphic *Eco*RI fragment in the subtelomere of chromosome 4q. The size-variability of this fragment is caused by an, at that time uncharacterized, macrosatellite repeat D4Z4 flanking p13E-11. The D4Z4 repeat consists of identical 3.3 kb *Kpn*I units that may vary in number between 11–100 units on normal chromosomes.[117] Sporadic FSHD patients exhibit a *de novo* contracted D4Z4 repeat of 1–10 units.[138] Similar sized D4Z4 alleles were also observed in familial cases. Transmission of this rearrangement in all affected offspring confirmed that FSHD is associated with a contracted D4Z4 repeat.[135]

An almost identical and equally recombinogenic D4Z4 repeat has been identified on chromosome 10q26, but D4Z4 contractions on this chromosome have never been associated with FSHD (Section 1.2.4, Figure 1a). To discriminate between 4qter- and 10qter-derived repeats, in addition to *Eco*RI, the restriction enzyme *Bln*I was used, which digests only chromosome 10-derived repeat units (Figure 1c).[22] New D4Z4 repeat contractions have been detected in germline as well as during embryogenesis, the latter giving rise to gonadal and somatic mosaicism (gonosomal mosaicism) for the FSHD mutation (Section 1.3, Figure 1b). A rough and inverse correlation has been observed between the severity and age at onset of the disease and the number of D4Z4 units in the repeat.[68,105] In general, familial FSHD cases have longer D4Z4 repeats (5–10 units) than non-familial cases (1–4 units) and, as a consequence, a milder phenotype. However, the complete absence of the 4q telomeric region including D4Z4 has been detected in three generations of a healthy family, showing that haploinsufficiency of the D4Z4 region does not cause FSHD.[110]

Recently, compound heterozygosity and even homozygosity has been described for FSHD alleles, which showed that both these phenomena are compatible with life. A possible phenotypic dosage effect was observed for patients that were compound heterozygous for FSHD alleles (Chapter 5).[145] This apparent dosage effect was not observed in a homozygous FSHD patient of a Brazilian family.[107] However, the many asymptomatic cases in this consanguineous family show that this FSHD allele is of unusually low penetrance. Furthermore, this family also displays another dystrophy, because a very severely affected cousin of the homozygous patient does not carry the FSHD allele.[107] Possibly, the locus associated with this other dystrophy is causing a more severe phenotype in the mother, which is masking the phenotypic dosage effect in her homozygous child.

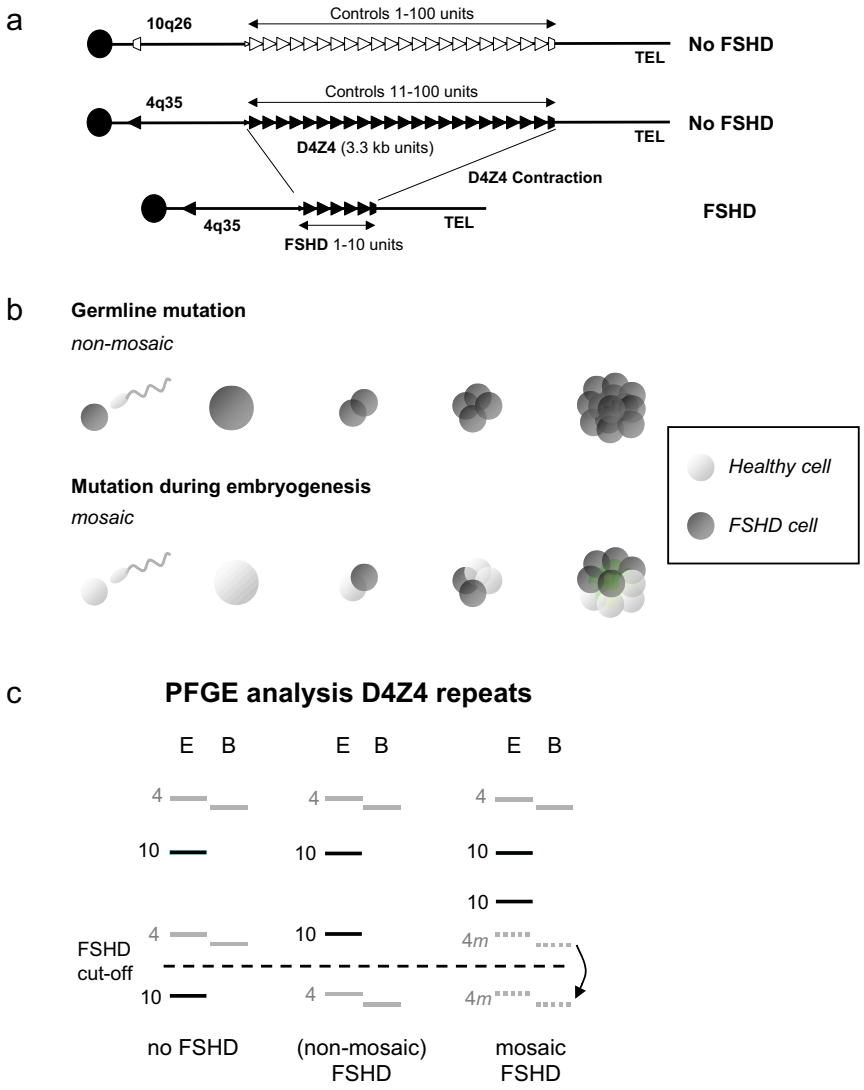


Figure 1 a) FSHD is a dominantly inherited disorder caused by a contraction of the D4Z4 repeat on chromosome 4q35 to 1–10 D4Z4 units. Each D4Z4 unit is 3.3 kb in size. Controls have a D4Z4 repeat of more than ten D4Z4 units on both chromosomes 4. An almost identical D4Z4 repeat is located on chromosome 10q, but contractions of this repeat have never been associated with FSHD

b) In 10–30% of FSHD cases, a *de novo* D4Z4 contraction has been detected. These contractions occur almost equally frequently either in the germline or during early embryogenesis. In the latter case, only a fraction of the embryonic cells carry the FSHD allele; the individual is mosaic for FSHD.

c) Representation of typical Southern blot analysis of D4Z4 repeats using pulsed field gel electrophoresis (PFGE). Genomic DNA is double digested with *Eco*RI and *Hin*dIII (E) and *Eco*RI and *Bln*I (B) and after PFGE and blotting, hybridized with probe p13E-11. The E lane typically displays four alleles, two from chromosome 4 (grey dashes) and two from chromosome 10 (black dashes). The B lane only reveals the D4Z4 repeats from chromosome 4. The left genotype has a short 10-type D4Z4 repeat and is derived from an unaffected individual. Middle and right genotypes both display a short 4-type D4Z4 repeat and are from FSHD patients. The right patient is mosaic (displaying two mosaic alleles, 4m) for FSHD, indicating that the D4Z4 contraction occurred during embryogenesis.

1.1.4 Potential molecular mechanisms

FSHD gene within D4Z4 (*DUX4*)

The association of FSHD with D4Z4 contractions made this sequence the first target for the FSHD candidate gene analysis. D4Z4 has a high GC-content (71%) with a CpG/GpC ratio of 0.8, which is characteristic for CpG islands.[137,147] Within each repeat unit a putative gene was discovered that contains a double homeobox domain, and was named *DUX4* for double homeobox 4.[30] *DUX4* has a predicted open reading frame (ORF) of 424 amino acids and is preceded by a sequence that demonstrated strong promoter activity in transient expression studies (Figure 2a).

Recently, we studied the DNA methylation of the most proximal D4Z4 unit of D4Z4 repeats on chromosome 4q35, using two methylation-sensitive restriction enzymes, of which one recognizes a restriction site in the promoter region of *DUX4* (Chapter 2).[127] For both enzymes a marked hypomethylation of the FSHD allele was shown in individuals with FSHD compared to controls. Interestingly, individuals with phenotypic FSHD but without a contracted D4Z4 repeat displayed an even more prominent hypomethylation. Conversely, our results were not in accordance with an earlier study that suggested DNA hypermethylation in control and FSHD individuals on all D4Z4 repeats.[109] This study was also performed using methylation-sensitive restriction enzymes, but in contrast to our study, a probe was used that recognizes all D4Z4 repeats on chromosomes 4 and 10, as well as many other loci. This means it compared the methylation of less than 10 D4Z4 repeats of the FSHD allele with usually about 100 D4Z4 repeats from non-pathogenic repeat arrays. As a consequence, this method is far less accurate and reliable. Interestingly, the *DUX4* promotor was recently shown to be hypomethylated in colon cancer cells that were deficient for two major DNA methyltransferases (DNMT1 and DNMT3b).[91]

Despite the marked D4Z4 hypomethylation in FSHD patients, to date *in vivo DUX4* transcription has not been demonstrated in cells of either controls or patients.[30,41,71,144] Recently, 2D Western blot experiments suggested the potential expression of the *DUX4*-protein in a primary myoblast culture derived from an FSHD patient, which was not visible in that of a control.[18] However, the same blot revealed, in addition, many other *DUX*-homeodomain-specific proteins that are differentially expressed. Therefore, this result necessitates further analyses of the detected protein spots. Moreover, the fact that an almost identical *DUX4*-ORF is identified within the D4Z4 repeats on chromosome 10, but short repeats on this chromosome have never been associated with FSHD, makes an important role for *DUX4* transcription in the etiology of FSHD difficult to envisage.

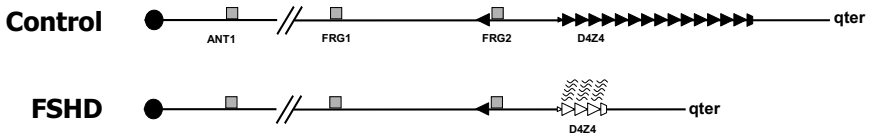
Further analysis of the D4Z4 sequence revealed the presence of two classes of repetitive DNA, *hhspm3* and *LSau*.[41] The most distal D4Z4 unit of the D4Z4 repeat has been shown to be polymorphic in length[153] ending with a 260 bp sequence, which is called pLAM.[117] Directly

distal to pLAM, a 6.2 kb ß-satellite repeat (consisting of 68 bp tandemly repeated Sau3A monomers) was identified.[117] Large β-satellite repeats and *LSau* are generally localized in heterochromatic regions of chromosomes 1, 3, 9 and 10 and the acrocentric chromosomes (13, 14, 15, 21 and 22).[41,143] The high homology of D4Z4 elements with heterochromatin regions and its subtelomeric localization suggested a heterochromatic structure for the D4Z4 repeat array.
Recently, a bi-allelic variation distal to D4Z4 was detected on chromosome 4, marked 4qA- and 4qB (Section 1.2.4). Alleles that harbor pLAM and the ß-satellite repeats are called 4qA and these elements are absent in the telomeric region of 4qB alleles.[121] Southern blot analyses have shown that both chromosome ends are almost equally common and equally recombinogenic in the population, but FSHD alleles are only associated with the 4qA distal polymorphism (Chapter 3).[57] FSHD-sized D4Z4 repeats have been detected in 4qB-type alleles but these alleles do not cause FSHD as was shown in three different families (Chapter 4).[63] For both pLAM and the ß-satellite repeat, no specific role in the pathogenesis of FSHD could be defined and, to date, it remains unclear why only D4Z4 contractions on 4qA alleles are pathogenic.

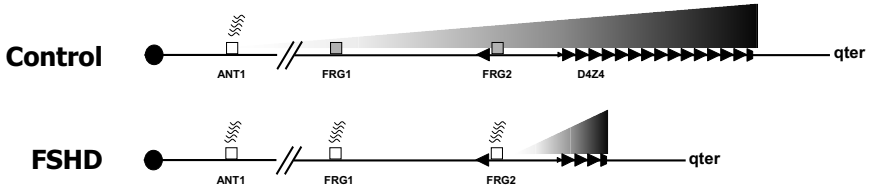
PEV-model

Until recently, the common model explaining the association of FSHD with the contraction of D4Z4 at one 4q35 allele was based on the putative heterochromatic structure of normal-sized D4Z4 arrays. Variable silencing of genes in the vicinity of heterochromatin, called position-effect variegation (PEV) has been described in *Drosophila melanogaster*.[134] In this model a euchromatic gene is downregulated because it has been placed in the vicinity of heterochromatin by a rearrangement. PEV can also be caused by expansion of a DNA repeat, with a direct correlation between repeat length and proximity of the gene to constitutive heterochromatin, and the level of gene silencing.[103] The *cis*-acting repression in PEV is probably caused by a linear gradient of heterochromatin spreading, known as heterochromatinization, and a corresponding degree of gene silencing. The distance between PEV causing rearrangements and the affected gene can be several megabases.[130] According to the *Drosophila*-PEV model, contractions of the D4Z4 repeat in FSHD patients may cause the loss of heterochromatinization and thereby an inappropriate increase of gene expression of 4q35 genes in affected cells in FSHD patients (Figure 2b).[41,142]
Recently, some evidence was presented favoring the PEV model. A transcriptional upregulation of different 4q35 genes, *ANT1, FRG1* and *FRG2* was observed in biopsies of FSHD muscles compared to control muscle using semi-quantitative PCR.[29] It was suggested that the transcriptional upregulation of 4q35 genes was inversely proportional to the distance from the D4Z4 repeat, and that the upregulation correlated inversely with the length of the D4Z4 repeat in FSHD muscle. Furthermore, a multiprotein complex was discovered that binds to a 27 bp sequence on D4Z4 (D4Z4 binding element). This so-called ‘D4Z4 repressing complex’ consists of the proteins YY1, HMGB2 and nucleolin and *in vitro* depletion of any of these components

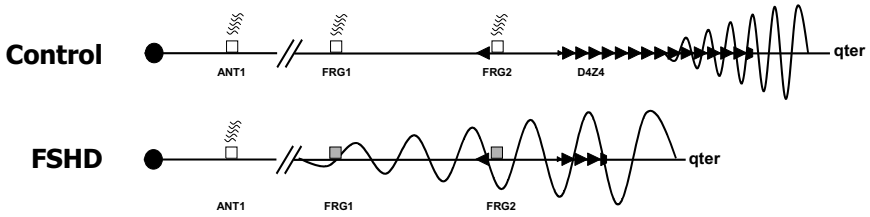
a) Expression from D4Z4



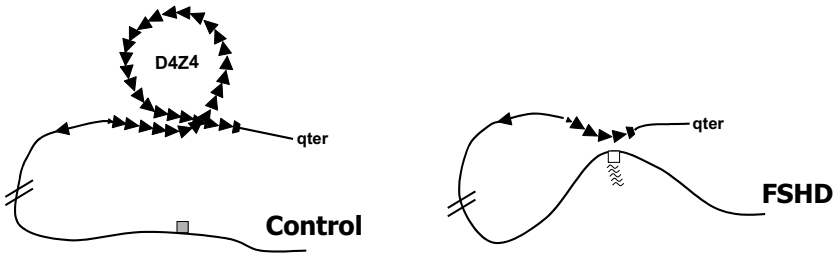
b) *Cis*-spreading model



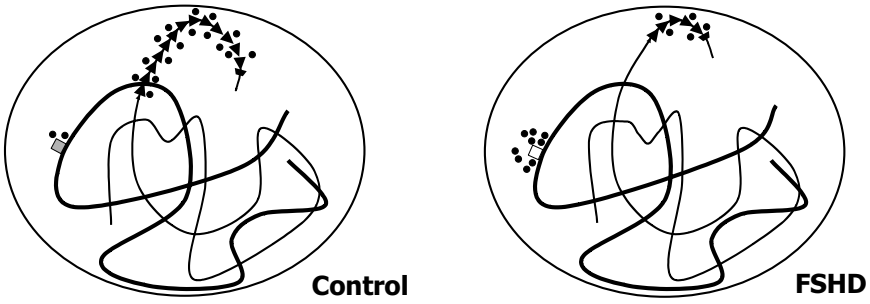
c) Insulator model



d) *Cis*-looping model



e) Nuclear organization model



resulted in an upregulation of the *FRG2* gene closest to D4Z4.[29] However, in several independent follow-up studies, the upregulation of *FRG2* and other 4q35 genes could not be confirmed by real-time PCR or array studies.[48,94,144]

The CpG-hypomethylation observed in the pathogenic D4Z4 repeat of FSHD patients may support a model for FSHD in which D4Z4 contractions result in a local chromatin decondensation that subsequently alters the gene expression on 4qter (Chapter 2).[127] Together, these observations could explain the FSHD-specific gene upregulation in *cis*, like in the 'loss of PEV' hypothesis.

Conversely, the D4Z4 repeat was suggested to display an insulator function between the distal heterochromatic and proximal euchromatic region. In this model, contraction of the D4Z4 repeat on 4q35 results in a heterochromatic spreading in proximal sequences, which would silence 4q35 genes (Figure 2c).[118]

Figure 2 Different models to explain the molecular disease mechanism of FSHD. In each section control and affected alleles are depicted and the triangles represent D4Z4, in sections a to c the location of FSHD candidate genes, *ANT1*, *FRG1* and *FRG2* are shown. The gene that is activated as a consequence of the D4Z4 contraction is indicated with an open square or triangle.

a) In the first model, FSHD is caused by expression from the contracted D4Z4 repeat. Within each D4Z4 unit a potential gene has been identified, *DUX4*. Until now, *in vivo DUX4* transcription has not been demonstrated.

b) In the *cis*-spreading model, the long D4Z4 repeat and nearby sequences share features of heterochromatin. Upon contraction, a local chromatin relaxation causes the transcriptional upregulation of 4qter genes in a distance-dependent manner, possibly through the action of the D4Z4 repression complex.

c) In the insulator model, D4Z4 acts as a spacer between heterochromatic sequences distal to the D4Z4 repeat and euchromatic sequences proximal. Upon contraction, this insulator function is incomplete allowing heterochromatinization of proximal sequences and transcriptional downregulation of genes within this region (filled squares indicate silenced genes).

d) The *cis*-looping model postulates that normally intra-array loops in arrays >10 units prevent the interaction of D4Z4 with genes in *cis* at large distances. Disruption of this interaction may cause inappropriate gene expression by long range interaction with D4Z4.

e) Finally, the nuclear organization model predicts that the interaction of 4qter with the nuclear lamina where chromatin and transcription factors (dots) are tethered is disturbed in FSHD. This alteration in turn, may lead to a misbalance of chromatin and transcription factors at 4qter and unrelated loci.

Long-distance *cis* looping model

The PEV model has been critically examined by studying the chromatin condensation of D4Z4 and proximal sequences in patients and controls.[48] This study focused on some essential features of the PEV model, the condensation of D4Z4 in controls, the spreading of this condensation (heterochromatinization) and the reduced 4q35 condensation in patients.

Heterochromatinization was tested by chromatin immune precipitation (ChIP-assay) with an antibody for acetylated histone H4 that discriminates between constitutive heterochromatin and unexpressed euchromatin.[48] Because D4Z4 repeats in controls display higher methylation levels than FSHD-sized repeats, they were considered to be highly condensed. Unfortunately, D4Z4-acetylation could not be analyzed in this ChIP-assay, because no specific primers could be designed for this sequence. Unexpectedly, the p13E-11 region proximal to D4Z4 on 4q35 and 10q26 showed acetylation levels similar to unexpressed euchromatin in normal and FSHD lymphoid cells rather than that of constitutive heterochromatin. Furthermore, no significant acetylation differences were observed between normal and FSHD-sized alleles in a limited set of samples. Although this could be due to the fact that p13E-11 recognizes 4 alleles (FSHD-sized and normal from 4 and 10), which could not be separated in this experiment, the authors claim that these findings argue against the PEV model.[48]

The same study showed that the promoter regions of FSHD candidate genes *FRG1* and *ANT1* in normal and FSHD cells display a similar H4 hyperacetylation as seen in expressed genes. Interestingly, quantitative RT-PCR analysis of seven FSHD and seven control muscle biopsies showed, in contrast to previous observations[29], no differential expression for *ANT1* between healthy and affected biopsies. Even more peculiarly, a significant *FRG1* downregulation was observed in FSHD muscles.[48]

To explain these results, another model for the molecular genetic etiology of FSHD was proposed.[48] According to this model the intrachromosomal communication of an FSHD-sized D4Z4 array and an FSHD target gene in *cis* occurs by looping rather than by a progressive spreading of heterochromatin as in the PEV model. This looping can only be formed when a D4Z4 repeat contraction impairs the formation of intra-array loops in normal sized D4Z4 repeats. It is proposed that gene expression of target genes that are more than 160 kb from the D4Z4 repeat can be altered by an abnormal looping interaction, which delivers a positive transcription factor to its promoter, or changes its local chromatin structure (Figure 2d). Otherwise, the looping interaction could influence the nuclear localization of the target gene region.[48] The observation that the size distribution of D4Z4 repeats is multimodal with equidistant peaks at 60 kb may support this model.[126]

Local chromatin alteration

The first genome-wide expression study demonstrated a global misregulation of muscle-specific gene expression in FSHD.[111] More recently, Affimetrix cDNA chips were used for these transcription studies comparing muscle biopsies from FSHD patients with controls and other muscular dystrophies.[144] Amongst the genes that were altered in an FSHD-specific and highly significant manner, many were involved in myogenic differentiation. Furthermore, genes were identified that could explain the higher sensitivity of FSHD myoblasts to oxidative stress.[141] Unexpectedly, none of the 4q35 genes were specifically upregulated in FSHD muscle biopsies, which was verified using a custom cDNA microarray containing 51 genes and expressed sequence tags (EST) from the 4q35 region. These observations are in contrast to the transcription study using radio-active end point RT-PCR[29] and made the PEV model for FSHD less likely to be correct. The lack of altered gene expression on 4q35 suggests that alterations in the D4Z4 chromatin structure do not act in *cis*. In considering other models for the disease pathogenesis, the nuclear positioning of the FSHD region was examined.[72]

The mammalian nucleus is organized in different compartments with specific domains that carry out distinct functions such as transcription, RNA processing and replication.[99] Proper nuclear positioning has been show to be essential for normal gene expression.[20] It was shown that 4qter is one of the very few telomeres that localizes to the nuclear periphery in normal myoblasts, myotubes, fibroblasts and lymphoblasts.[72] At the nuclear periphery, chromatin is anchored to the inner nuclear envelope via lamin A/C, LAP2β and BAF.[85] The nuclear envelope and associated lamina play a role in gene expression, chromatin organization and differentiation.[72] Disruption of the nuclear envelope has been shown to cause other forms of neuromuscular disease, such as Emery-Dreifuss muscular dystrophy, Limb Girdle muscular dystrophy 1B and dilated cardiomyopathy.[14,85] These diseases are caused by mutations in lamin A/C or emerin, which are both essential for the proper formation of the nuclear envelope.

Like these other so-called chromatin diseases, FSHD-associated D4Z4 contractions were supposed to cause an improper nuclear positioning of the FSHD region disrupting the regulation of genes elsewhere in the genome. The peripherical localization of the 4q35 region indeed suggests that it is associated with the nuclear envelope. This was strengthened by the observation that in a lamin A/C deficient fibroblasts this specific positioning was no longer observed, while the whole chromosome 4 compartmentalization did not change. However, no difference was found in the nuclear localization between normal and FSHD-alleles.[72] Therefore, FSHD is probably caused by improper interactions with transcription factors or chromatin modifiers at the nuclear envelope (Figure 2e). Interestingly, the D4S139 region (40 kb proximal to D4Z4) localizes closer to the nuclear envelope than D4Z4. This suggests that the region proximal to D4Z4 is involved in the association to the nuclear periphery and could explain the chromosome 4-specificity of FSHD.[72]

1.1.5 Identification of FSHD genes

The originally postulated PEV model initiated the isolation and characterization of genes on chromosome 4q35 proximal to D4Z4. Unfortunately, analysis of this region, using several gene identification techniques, was seriously complicated due to the subtelomeric dispersion. The subtelomeric plasticity (Section 1.2.3) duplicated some 4q35 (putative) genes to other chromosomes and consequently all gene sequences identified have many highly homologous copies on other chromosomes in the human genome.[119] Because at that time the sequence of subtelomeres and telomeres were not completed in the Human Genome Project, van Geel and colleagues sequenced 375 kb of the subtelomeric 4q35 region proximal to D4Z4, which considerably assisted the chromosome 4q assignment of these sequences (cDNA and EST clones).[123]

In total, five different genes were discovered in the 150 kb most distal region of chromosome 4q, including*; FRG1, TUBB4Q, DUX4c, FRG2* and *DUX4. DUX4* has already been discussed in section 1.1.4; the other genes will be described briefly in the following sections. Despite intensive examination, no other genes were predicted in the 225 kb region proximal to *FRG1*.

FRG1

The FSHD region gene 1 (*FRG1*) is located 125 kb proximal to D4Z4 and was the first FSHD candidate gene identified.[120] The *FRG1*-gene has nine exons and a transcript of 1042 bp. The gene is constitutively transcribed in many tissues and encodes a protein of 258 amino acids. *FRG1* is evolutionarily highly conserved in vertebrates and non-vertebrates.[35] Recently the function of *FRG1P* has been studied in greater detail.[125] It was shown that, in interphase cells, *FRG1P* was localized in the dense structures of the nucleolus, in Cajal bodies and in the speckles, by stable and transient expression. This localization suggested a role for *FRG1P* in RNA processing, which was supported by the redistribution of the protein in transcription inhibition experiments. Remarkably, two other neuromuscular disorders, oculopharyngeal muscular dystrophy (OPMD) and spinal muscular dystrophy (SMA), are caused by defects in different proteins that are involved in RNA processing and colocalize with *FRG1P*.[125]

Expression studies of *FRG1* in muscle of FSHD patients and controls by different groups did not reveal a common differential expression.[29,48,120,144] Probably, the use of different techniques, the interference of non-4q homologs and the inhomogeneity of muscle samples underly this discrepancy. However, the high conservation and the possible role in RNA processing make *FRG1* an attractive candidate gene.

TUBB4Q, FRG2 and *DUX4c*

Three other genes have been identified in the 4q35 region between *FRG1* and D4Z4. The first gene identified, *TUBB4Q*, is localized 80 kb proximal to D4Z4 and shows high homology to functional ß-tubulin genes.[124] ß-tubulin is one of the major components of microtubules, which

are involved in many cellular processes and are found in all eukaryotic cells.[51] *TUBB4Q* contains four exons encoding a putative protein of 434 amino acids. Most probably, *TUBB4Q* is a pseudogene since it has amino acid substitutions in highly conserved protein domains. Furthermore, it has a mutated start codon in about half of the cases[124] that do not segregate with FSHD-alleles (unpublished results). Finally, extensive analyses in many tissues did not reveal *TUBB4Q* transcription.[124]

Another potential gene has been identified within D4S2463, which is an inverted and truncated D4Z4 unit that is located 40 kb proximal to D4Z4.[71] This gene, centromeric *DUX4* (*cDUX4*) differs from *DUX4* in the carboxy-terminal domain and the promoter region (Marcowycz, FSHD-workshop 2003). *cDUX4* encodes a 374 amino acids long protein and has only been detected *in vitro*.

The gene closest to D4Z4, the FSHD region gene 2 (*FRG2*), is located only 37 kb proximal to D4Z4.[94] The cDNA of the gene is 2084 bp in length with four exons that encode a protein of 278 amino acids. Furthermore, a potentially strong promoter precedes *FRG2*. This promoter directs high levels of *FRG2* expression *in vitro* but is inhibited by increasing numbers of D4Z4 repeat units. Recently, *FRG2* transcription was demonstrated in FSHD muscle using a rather controversial radioactive RT-PCR with more than 40 amplification rounds.[29] However, real-time RT-PCR and cDNA micro-array experiments were unable to detect *FRG2* transcription in any tissues, including muscle.[94,144] Consistent *FRG2* transcription was shown *in vitro* for differentiating FSHD myoblasts, partly from chromosome 4 but predominantly from its homolog on chromosome 10. In contrast differentiating control myoblast only displayed transcription from distantly related FRG2 homologs (Section 1.2.3). Furthermore, control and affected human fibroblasts undergoing forced myogenesis showed predominantly transcription of the *FRG2* copy from chromosome 10. Additionally, a high *FRG2* transcription was detected in a monochromosomal cell hybrid that contained only chromosome 4 as human component.[94]

Recently three FSHD families have been described showing a contraction of D4Z4 extending to the region proximal to D4Z4 (Chapter 6).[59] The patients in these families display a normal spectrum of the disease. Detailed analysis of these deletions showed, in two independent families, the deletion encompassing also *DUX4*c and *FRG2*. In one of these families the deletion was detected in the patient (FSHD-sized D4Z4) and his healthy father and brother (normal-sized D4Z4).[60] These observations challenge the role of *DUX4*c and *FRG2* in the etiology of FSHD. However, transcriptional activity of *FRG2* from chromosome 4 and 10 in FSHD myoblasts and not in control myoblasts suggested a transvection model for *FRG2* in which chromosome pairing during interphase induces transcriptional activity of homologous genes in *trans*.[92] According to this model *FRG2* expression from chromosome 10 can be induced by a D4Z4 contraction on a chromosome 4 that lacks *FRG2* and could still play a role in FSHD pathogenesis.[94]

1.2 Subtelomeric plasticity

1.2.1 General

Human subtelomeres are unusually dynamic regions of chromosomes that form the transition between chromosome-specific sequences and the telomeric repeats ($TTAGGG_n$) that cap chromosome ends.[42,108,140] Subtelomeres contain large blocks of sequence, which are present on multiple chromosomes and are called region specific 'low-copy repeats' (LCRs), segmental duplications or duplicons.[100] Additionally, these LCRs can also be found near centromeres. In contrast, other repetitive sequences (Section 1.3.1) do not have this specific localization but can be found at distinct sites in the human genome. Usually LCRs consists of 10–400 kb genomic DNA, with >97% sequence identity. They comprise at least 5% of the human genome and can encompass genes, gene fragments, pseudogenes, endogenous retroviral sequences or repeat gene clusters.[100] Comparative analysis of subtelomeres of different human chromosomes has revealed a common organization for many chromosomes. According to this organization, two subtelomeric domains, a proximal and a distal domain, are separated by a degenerated TTAGGG repeat. Sequences at the distal domain are short (<2 kb) and repeated at many other chromosome ends. The sequences in the proximal domain are homologous to only a few ends, and the blocks of homology tend to be longer (10–300 kb) than those in the distal domain.[26,75]

Differences in copy number and chromosomal location have been observed for many subtelomeric blocks in the human genome and between the genomes of human and non-human primates.[4,42,78,95,108] This observation suggests the occurrence of segmental duplications and/or deletions during recent human evolution. To date it is unknown which exact mechanism is underlying the duplication and dispersion of the various subtelomeric blocks among the many chromosome ends. Possibly, the LCRs size and the sequence homology between the LCRs as well as the orientation with respect to each other predisposes paralogous (non-allelic) genomic fragments for homologous recombination. Depending on the orientation of paralogous LCRs, non-allelic homologous recombination can result in deletions, duplications, inversions, translocations and other complex chromosome rearrangements.[69,100]

Although the function of these recombinogenic subtelomeric regions is largely unknown, several hypotheses have been proposed. Subtelomeres may simply be a repository for genomic material, which is no longer in use. This 'junk model' is supported by the fact that subtelomeric regions encompass many pseudogenes. Furthermore, incomplete genes can be found in these regions, of which the functional copy is located elsewhere in the genome.[75] On the other hand, subtelomeres may serve in the development of new genes, the 'nursery model'. In this model, the sensitivity of subtelomeres for rearrangements and mutations plays an important role in the expansion of protein diversity (generation of new genes) through the alteration of coding sequences.[4] This is supported by the fact that some duplicated subtelomeric genes that require high diversity are transcribed, such as the olfactory receptor (OR) genes (Section 1.2.2).

Furthermore, it has been shown that LCRs in the human genome map to regions that are enriched for genes associated with immunity and defense, membrane surface interactions, drug detoxification and growth/development.[4] Therefore, subtelomeric regions are the perfect home for gene products that require rapid diversification and adaptation.
On the other hand, duplications can mediate pathogenic rearrangements. These genomic disorders are caused by DNA rearrangements resulting in a complete loss or gain of a gene(s) sensitive to a dosage effect or in disruption of a gene. These alterations in the genome can result from both inter- and intrachromosomal rearrangements by paralogous, homologous recombination involving LCRs that encompass genes or gene fragments.[100] Some of these genomic disorders are associated with duplication hotspots in subtelomeric regions like: glucocorticoidremediable aldosteronism (8q), polycystic kidney disease (16p13), alpha thalassaemia (16p), Hunter syndrome (Xq28), red-green color blindness (Xq28), Emery-Dreifuss muscular dystrophy (Xq28), incontinentia pigmenti (Xq28) and hemophilia A (Xq28).[4] Additionally, about 5 to 10% of mental retardation is caused by subtelomeric deletions.[27]

1.2.2 Gene duplications

Subtelomeric plasticity played an important role in the development of the different olfactory receptor (OR) genes that control the sense of smell. OR genes encode a large family of proteins that can identify and distinguish thousands of smells.[13] The sense of smell is essential to the survival of most species, who use their olfactory systems to identify food, smell predators and observe and interpret their environments. Furthermore, smell plays a role in mate choice, mother-infant recognition and signaling between members of a group.[150] The scientists responsible for cloning the first members of this gene family in 1991 were honored with the Nobel Prize in 2004.[12]
OR-genes are arranged in clusters that contain from 1 to 100 genes in more than 40 chromosomal locations in the mouse genome and over 100 locations in the human genome.[150] The OR family is one of the largest mammalian gene families known, with 1500 genes in mouse and 900 in human, occupying almost 1% of the mammalian genome. OR genes encode G-protein-coupled receptors containing seven transmembrane domains. In humans the OR protein is about 300 amino acids in length.[66]
The OR gene family has expanded mainly by segmental duplications, many of which have occurred since the divergence of the rodent and primate lineages. One subset of subtelomeric OR duplications is so recent that the copy number and the locations of the duplicated block are polymorphic. Sequence analyses of many copies of these OR genes, from different chromosomes and/or individuals, have shown that the main coding exon is an intact ORF with a few amino acid differences.[66] Some OR genes have become pseudogenes during human evolution, while others have been subjected to positive selection (i.e. after gene duplication, change in protein sequence is advantageous, giving rise to OR gene variation). These findings

suggest that the subtelomeric duplications resulted in species and individual specific variations in the expression of OR-genes, enabling the detection and discrimination of an increasing number of odorant molecules.[150]

1.2.3 Duplication subtelomeric 4q35 genes

The 4q35 subtelomere displays several features common to subtelomeres and because of its association with FSHD, it has been studied extensively. It is organized in a two-domain structure, of which the distal region, the 4qA/4qB region, shows homology to almost all chromosomes. Furthermore, the proximal domain has been subjected to segmental duplications, causing the dispersion of this region over the human genome.[123] Most FSHD candidate genes (Section 1.1.5) are located in this subtelomeric region and have been duplicated as discussed in the next section.

FRG1 homologs

On chromosome 4 *FRG1* is located 125 kb proximal to D4Z4. Many homologous *FRG1* sequences have been identified in the human genome, including the pericentromeric region of chromosome 9, the centromeric region of chromosome 20, the short arm of all acrocentric chromosomes (13, 14, 15, 21 and 22) and chromosomes 8 and 12 (Figure 3).[120] Most *FRG1* homologs are mapped to heterochromatin regions like the 4q35 region.

Some evidence for transcription of the *FRG1* homologs was demonstrated in rodent cell hybrids containing a single human chromosome.[120] Nevertheless, sequence analyses showed that many of these transcribed human *FRG1* homologs contain splicing artifacts such as missing exons or the presence of intronic sequences. Furthermore, in many of these sequences, sequence variations were found in the ORF.[36,120] Therefore, it was concluded that most probably the majority of *FRG1* homologs are pseudogenes.

In the great apes (chimpanzee, gorilla and orang-utan) *FRG1* is, similarly to in humans, part of a large gene family. However, in an Old World monkey, the *Macaca mulatti*, only two *FRG1* loci were detected of which one is the 4q35 ortholog.[36] These observations indicate that multiple *FRG1* duplications occurred in the great apes lineage.[36] In mouse only one *FRG1* homolog was found, which mapped to chromosome 8, the synthenic region to human chromosome 4q35.[37] Together, these observations indicate that the *FRG1* gene on chromosome 4 is the ancestral homolog of the *FRG1* gene family.

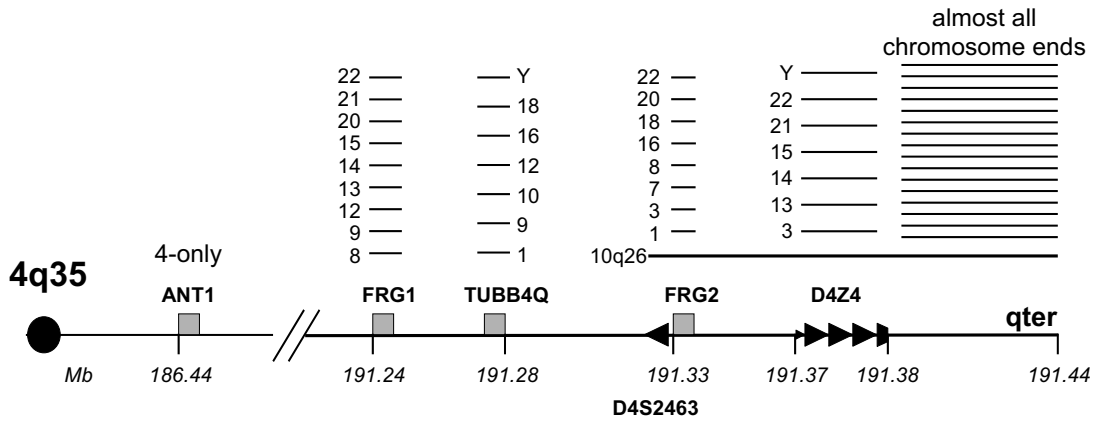


Figure 3 The distal 200 kb region of chromosome 4q35 is a typical example of subtelomeric plasticity. Homology to *FRG1, TUBB4Q, FRG2* and D4Z4 was found on 8 to 9 other chromosomes, while homology of the telomeric region distal to D4Z4 was shown with almost all other chromosomes. In contrast the more distantly located *ANT1* gene is single copy. The more recent duplication of the distal 4q35 region to 10q is indicated by a thick line. The distances of the genes relating to the chromosome 4 map (NCBI, *Homo sapiens* genome view) are denoted below the start codon of each gene in megabases (Mb).

TUBB4Q homologs

TUBB4Q was the first identified member of the *TUBB4Q* subfamily at 80 kb proximal to D4Z4. The subfamily consists of at least 10 members and represents an isotype of the large human ß-tubulin supergene family.[122,124] The β-tubulin supergene family has at least 16 members, of which about half are expressed.[122] Since β-tubulins contribute to vital processes in the cell they have highly conserved protein sequences between vertebrate species. Members of the *TUBB4Q* family displays high homology to functional β-tubulin genes, but have mutations in conserved protein domains or the start codon, and seem not to be transcribed. They are mainly located at subtelomeric and pericentromeric regions (1q42, 1q43–44, 4q35, 9q34, 10p15, 12cen–p11, 12cen–q11, 16q24, 18p11 and Yq11, Figure 3) and the 10q15 member is most likely the only functional copy.[122]

Four *TUBB4Q* orthologs have been identified in the baboon. It was shown that only the baboon ortholog of 4q35 represents a functional gene, while the others are considered to be pseudogenes. In humans, it was hypothesized that 4q35 harbors the ancestral copy of this subfamily and that this was once a functional gene.[122] Due to its subtelomeric localization *TUBB4Q* was duplicated to other chromosomes. Subsequently, the original *TUBB4Q* on 4q35 acquired mutations and recently became silenced, while the duplicated copy on 10q15 has assumed the original function. In this perspective the *TUBB4Q*-evolution might represent an example of gene gaining and silencing by duplication-mediated genome evolution.

FRG2 and *DUX4* homologs

Both *FRG2* at 37 kb distance from D4Z4 and *DUX4* in D4Z4 on 4q35 have many homologs. *FRG2* homologs are found on chromosomes 1, 8, 10, 18 and 20 and possibly on chromosomes 3, 7, 16 and 22 (Figure 3). Many DNA differences have been found within these gene copies and most of them lead to amino acid changes in the predicted protein. The 10q26 homolog only has 5 nucleotide (nt) mismatches in the ORF compared to the 4q35 gene. In fibroblasts undergoing forced myogenesis and differentiating myoblasts, predominantly the *FRG2* copy from chromosome 10 is expressed. Interestingly, a reporter assay study showed that the *FRG2* promoter region can direct high levels of expression but is inhibited by increasing numbers of D4Z4 repeat units.[94]

DUX4 is located within each unit of the D4Z4 repeat, which is a member of a dispersed 3.3 kb repeat family.[71] It has been shown that the chimpanzee, gorilla and orang-utan also contain many D4Z4-related sequences, of which one is the 4q35 homolog. However, in more distantly related primates, only two loci have been identified encompassing D4Z4-like sequences of which one is presumed to correspond to the human 4q35 repeat.[17] In human, D4Z4 homologs have been identified on 10q, 3p, Y and the pericentromeric regions of the acrocentric chromosomes (Figure 3).[71] Several homologs of *DUX4* have been identified, *DUX1*, *DUX2, DUX3* and *DUX5*, all of which map to acrocentric chromosomes.[30] The most identical *DUX4* copy is located on the subtelomere of chromosome 10q (Section 1.2.4). Until now, *in vivo* transcription and translation has only been demonstrated for *DUX1*.[30]

1.2.4 Distal 4q35 duplication

A good example of a recent segmental duplication is probably found in the subtelomeric regions of 4q35 and 10q26. In humans and all great apes the 4q35 region distal to D4S2463 is almost identical to the very tip of chromosome 10q26 (Figure 3). In Old World monkeys only one additional locus was found next to the locus syntenic to human 4q35.[17] This suggests that chromosome 4 subtelomere is the progenitor sequence that has been involved in subtelomeric plasticity over a period of at least 35 million years.[17,143] D4Z4 repeats on human chromosomes 4q35 and 10q26 show 99% homology and are equally polymorphic in length. Despite this high homology, D4Z4 contractions on chromosome 10 have never been associated with FSHD.

To explain the chromosome 4-specificity of FSHD both regions have been studied intensively for differences in gene sequence and telomeric organization. The 4q and 10q subtelomeric regions show >99% homology over a region of about 42.2 kb proximal to D4Z4, until D4S2463 (Figure 3).[121] This homologous region encompasses several genes, *DUX4*c, *FRG2* and *DUX4* (Section 1.1.5). Proximal to the region of homology both chromosomes are very different. The proximal 4q35 region is extremely gene poor, whereas the proximal 10q region encompasses many genes.[121]

The region distal to D4Z4 has been analyzed using telomeric YAC's originating from chromosomes 4 and 10. The telomeric region of one 4q YAC was almost identical to the 10q YAC over a region of 10 kb.[121] However, a second 4q YAC was almost identical to the telomeric sequence of chromosome 4p and showed only 92% homology to 10q. This suggested the presence of two distinct 4q alleles, 4qA and 4qB.[121] It was shown, by Southern blot analyses using allele specific probes, that both alleles indeed exist, and are almost equally common in the population (Chapter 3).[57] Phylogenetic analysis of the subtelomeric sequences from 4qA, 4qB, 4p and 10q confirmed a close relationship between 4qA and 10q, and 4qB and 4p. Based on the studies on *FRG1* and D4Z4 homologs, 4q harbors the evolutionarily ancestral subtelomere. Therefore, it is most likely that the 4qA-type allele was duplicated to chromosome 10q during evolution. Subsequently, the 4p (B-type) subtelomere was duplicated to 4q, resulting in the 4qA/4qB distal polymorphism at this chromosome end (Section 1.3.7, Figure 9).[121]
Subtelomeric variability has already been described for the subtelomeric region of chromosome 16p, encompassing the alpha-globin locus. Three alleles have been identified in which the alpha-globin genes lie 170 kb, 350 kb, or 430 kb from the telomere, because they encompass different telomeric segments.[140]

1.2.5 D4Z4 translocations

During meiosis, the paring of homologous chromosomes usually starts at the telomeres. As discussed before, an extensive subtelomeric homology among different chromosomes has been generated by segmental duplications, while on the other hand, these events yielded subtelomeric polymorphisms.[75] One might wonder how it is possible that the meiotic–pairing machinery always pairs the right chromosomes, despite the high subtelomeric homology between non-homologous chromosomes and subtelomeric variation between homologous chromosomes. Indeed, it has been shown that the meiotic recognition and pairing sometimes fails, leading to rearrangements between non-homologous chromosomes in yeast.[67] Mefford and Trask suggested that this failure of the meiotic–pairing machinery eventually initiates the recurrent shuffling of subtelomeres.[75] Analogously, the introduction of the 4qA/4qB distal polymorphism on chromosome 4 and the duplication between chromosome 4q35 and 10q26 created an opportunity for a further exchange between subtelomeric ends of non-homologous chromosomes. Matching and pairing of chromosome 4 homologs would become especially difficult when a cell is heterozygous for the 4qA/4qB polymorphism. This situation might promote a translocation between the highly homologous subtelomeric regions of chromosomes 4qA and 10.
Indeed, various groups have described translocations between D4Z4 repeats from chromosome 4 and 10s. A study in the Dutch population first revealed the existence of 4;10 translocations: 4-derived D4Z4 repeats on chromosome 10 and 10-derived D4Z4 repeats on chromosome 4. These translocated alleles account for about 5% of all alleles on both chromosomes.[60,116,126] In

addition, hybrid D4Z4 repeats have been identified that consist of both 4- and 10-type D4Z4 units. Other studies have shown that D4Z4 translocations have occurred worldwide and the ratio of 4;10 against 10;4 translocations varies between different ethnic groups.[73] Based on these observations, it was suggested that D4Z4 translocations occur relatively frequently.

An earlier study showed that translocated D4Z4 repeats on chromosome 4 were equally frequently composed of homogeneous 10-type and hybrid D4Z4 repeats, while those on chromosome 10 are often composed of homogeneous 4-type repeats.[126] Recently, the chromosomal origin and allele type was examined in more detail for many translocated D4Z4 repeats. Unexpectedly, we have found that most translocated D4Z4 alleles on chromosome 4 (4;10 translocated alleles) carried a hybrid D4Z4 repeat. In contrast, about 90% of all 10;4 translocated alleles carried an homogeneous 4-type D4Z4 repeat, while only 10% were hybrid. Approximately half of the homogeneous 4-type D4Z4 repeats on chromosome 10 displayed a 4qB distal variation, while the others were of the 4qA-type (manuscript in preparation).

Further analyses of the 10;4 translocated alleles with the 4qA variation demonstrated that all these alleles carry an identical microsatellite expansion close to D4Z4. Furthermore, a restriction fragment length polymorphism (RFLP) was identified in half of these 4;10 translocated alleles (manuscript in preparation). Therefore, we suggest that all these translocated alleles are derived from a single founder translocated allele, despite the fact that they all carry a different-sized D4Z4 repeat. Therefore, the variability of the D4Z4 repeat sizes of the translocated alleles most probably results from regular intrachromosomal D4Z4 rearrangements and not from a high frequency of D4Z4 translocations.

1.3 Plasticity of repetitive DNA sequences

1.3.1 General

More than half of the human genome consists of repetitive DNA sequences. Some of these repetitive sequences encode multigene families, but most of them are noncoding. Roughly, repeats are categorized into five classes: (1) interspersed repeats; (2) pseudogenes; (3) simple sequence repeats (microsatellite repeats); (4) segmental duplications (LCRs, Section 1.2.1); and (5) blocks of tandemly arrayed sequences.[55]

The most abundant class, occupying about 45% of the human genome, are the interspersed repeats, like short interspersed sequences (SINEs, for example Alu-repeats), long interspersed sequences (LINEs), LTR retrotransposons and DNA transposons.[55] Tandemly repeated sequences are categorized mainly according to their repeat unit size. We distinguish microsatellite (<10 nt), minisatellite (10–100 nt), macrosatellite (0.1–4 kb) and megasatellite (>4 kb) repeats. Tandem repeat arrays are generally very polymorphic in length and display a high mutation frequency.[81,93]

Alteration of both micro- and minisatellite repeat number have been observed in germline and somatic cells.[44,131] Furthermore, an increased instability of these repeats is observed in many human cancer cells[2,146] and after ionizing radiation.[45] Repeat rearrangements can cause disease by influencing gene expression, modifying the ORF within genes or generating fragile sites. Several inherited neuromuscular disorders are caused by trinucleotide instability, including myotonic dystrophy, Huntington disease, Oculopharyngeal muscular dystrophy (OPMD), several spinocerebellar ataxias, and Friedreich ataxia.[98]

Tandemly arrayed macrosatellite repeats may include transcribed genes, such as U2, histone and ribosomal RNA genes, as well as noncoding sequences. Rearrangements in most of these repeats, like RNU2, DXZ4 and RS447 (Section 1.3.4), have not been associated with disease. However, many diseases have been described that are associated with rearrangements of only two direct or indirect repetitive sequences. Disease in these cases can be caused by the disruption of one or more gene(s) within the repetitive sequence, that is 4 kb up to 500 kb in size (Section 1.1.2).

1.3.2 Models for repeat instability

It is generally accepted that repeat instability can be initiated by the repair of a DNA double strand break (DSB). DSB can occur spontaneously during DNA replication of single strand DNA nicks and can disturb the integrity of chromosomes and the viability of cells. Therefore efficient repair of these DSBs is essential for cell survival.[88]

Eukaryotes have developed several mechanisms to repair DSBs, including nonhomologous DNA end joining (NHEJ) and homologous recombination. Homologous recombination requires a template with sufficient sequence identity to the damaged DNA sequence to allow direct repair. Homologous DNA can be found on the homologous chromosome, the sister chromatid or on the same allele when the sequence is repetitive.[88]

Mainly based on the studies of micro- and minisatellite rearrangements, two types of mechanisms have been proposed to be involved in repeat instability: replication slippage and homologous rearrangement. Small changes in microsatellite copy number are most probably the result of replication slippage.[154] Homologous rearrangements are most often involved in the expansions and contractions of minisatellite and larger repeat arrays. Initially, two different rearrangement mechanisms were defined: the crossover mechanism (reciprocal transfer of genetic information) and the gene conversion mechanism (non-reciprocal transfer of genetic information) (Figure 4a).[93]

Gene conversions are most often explained by the DSB repair model as proposed for recombination events in yeast.[102] Initially within this model the resolution of the gene conversion was postulated to occur through cutting and resolving of two Holliday junctions, either with or without crossover.[43] When subsequent experiments showed only very low crossover rates, other models were proposed that did not require Holliday junctions, termed SDSA, for synthesis-dependent strand annealing.[79] In these models a 3' end of a resected DSB invades the donor template and primes DNA synthesis. Next, the newly synthesized DNA strands are unwound from the template and returned to the broken molecule, allowing them to anneal to each other. Out-of-frame annealing of the newly synthesized strand can lead to repeat contractions or expansion, while the donor template remains unchanged.

The mechanism underlying rearrangements of mini- and macrosatellite repeats has been studied extensively in *Saccharomyces cerevisiae*.[88] Mitotic rearrangements were studied by the introduction of a DSB in a repeat on a yeast chromosome, which could be repaired using the homologous repeat on a plasmid (donor sequence).[89] After DSB repair, donor and recipient were recovered and analyzed for repeat rearrangements. In these studies, most often the rearrangement was observed in the recipient molecule, while the donor allele was extremely stable. Consequently, most rearrangements could be explained by a gene conversion without crossover as in the SDSA model. The few rearrangements that were accompanied by a crossover were explained by a modified SDSA model that, like the DSB repair model, includes the possibility of crossover through cutting and resolving of two Holliday junctions.[89]

1.3.3 Repeat Instability in humans

Most studies on human minisatellite repeats (MS31, MS205, MS32 and CEB1) were performed in the germline, where some display a high instability.[44] Remarkably, for most loci, the male germline displays a higher repeat instability than that of the female.[128] In general, no exchange of flanking markers was observed, suggesting that these rearrangements occurred intrachromosomally. It was concluded that meiotic minisatellite rearrangements occur via an SDSA model with only few associated crossovers. Remarkably, a striking polarity was observed in the mutational process, i.e. repeats were preferentially expanded or deleted at one end of the repeat array. Minisatellite repeats MS32 and CEB1 also display mitotic instability, but at a much lower frequency than during meiosis. Small expansions and contractions were found in blood, which were allocated to replication slippage and showed no polarity. It was concluded that germline and somatic minisatellite mutations occur via distinct pathways.[7,11,47]

Only a few examples of macro- and megasatellite repeat instability have been described in humans. DXZ4 is a human macrosatellite repeat, which is localized on chromosome Xq24. This polymorphic repeat is comprised of 3 kb *Hin*dIII units and varies between 50 and 100 units. The repeat has a high GC-content (63%) and is hypermethylated on the active X chromosome and hypomethylated on the inactive X.[33] RS447 is a human megasatellite repeat, which is comprised of 4.7 kb units and varies in length from 20 to 103 units.[34] Similar to D4Z4, the individual RS447 units each contain a putative ORF. Furthermore, the repeat is localized on chromosome 4p16.1, and a highly homologous repeat array can be found on the distal part of chromosome 8p.[34] The RS447 repeat displays high meiotic instability, but also somatic instability has been described. Until now, the rearrangement mechanism has not been elucidated, but RS447 rearrangements are accompanied by repeat contractions as well as expansions.[81] Another megasatellite repeat is the U2 tandem repeat array in the RNU2 locus that maps to 17q21–22. The U2 repeat varies from 5 to 40 repeat units (30 to 250 kb), each 6.1 kb in size and encodes the U2 spliceosomal small nuclear RNA snRNA.[90] The RNU2 locus has mainly been studied in the repeat homogenization process called concerted evolution (Section 1.3.6). The rearrangement mechanism for this intrachromosomal homogenization is believed to occur via gene conversions and unequal sister chromatid exchanges.[65]

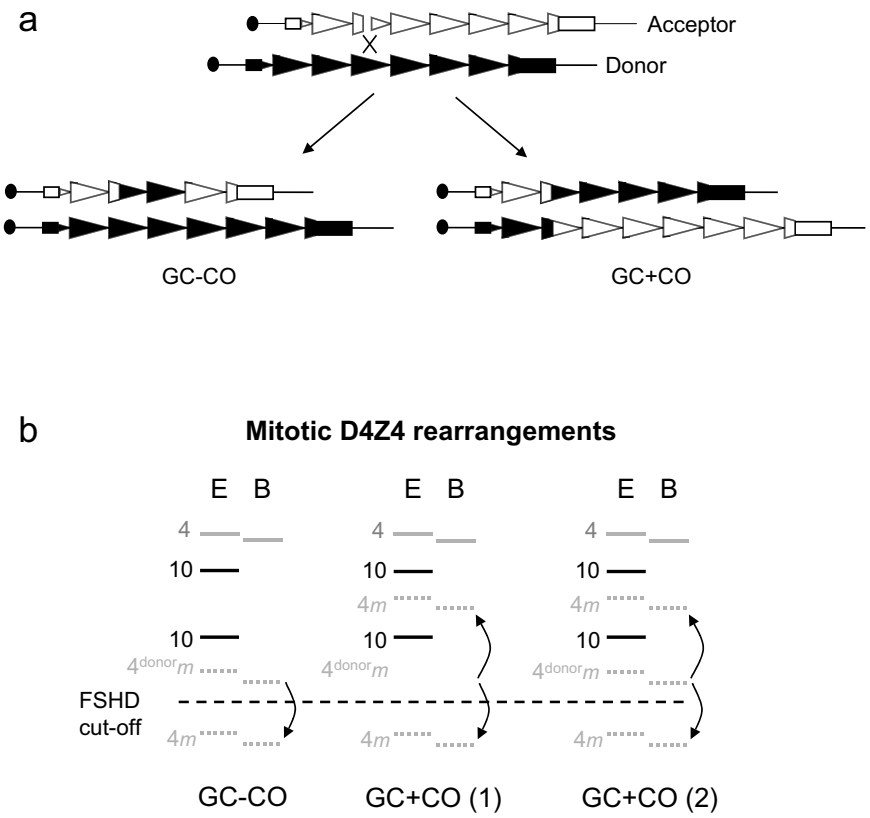


Figure 4 a) Generally repeat rearrangements are induced by double-strand DNA breaks in meiotic and mitotic cells. Break repair through gene conversion with crossover (GC+CO) or without crossover (GC-CO) requires pairing of the damaged DNA (acceptor) with an homologous DNA donor (Figure 5a). Repair of repetitive DNA by either GC+CO or GC-CO is often associated with a contraction or expansion of the repeat.
b) D4Z4 repeat contraction during embryogenesis results in mosaicism for FSHD. Depending on the rearrangement mechanism, different mosaic genotypes can be expected. In GC-CO rearrangements the donor repeat length remains unchanged (left). Two cell populations arise: one with the contracted acceptor and one with the unchanged donor repeat as was shown by the lower allele intensities (dashed lines). In GC+CO rearrangements, the length of the acceptor as well as the donor repeat is changed (often one contracted and one expanded repeat). If the contraction occurs before the first zygotic division (middle, GC+CO(1)), then cells with the original donor repeat length are absent (only mosaic expanded and contracted repeat visible, Figure 6b). If the GC+CO rearrangement occurs during the following zygotic divisions (right, GC+CO(2)), then mosaic ancestral sized repeats are visible next to mosaic expanded and contracted repeats (mosaic individual carries three D4Z4 cell populations).

1.3.4 Instability D4Z4 repeat

Chapter 7 describes the mechanism by which D4Z4 rearranges.[62] Almost half of the new FSHD mutations occur post-fertilization, resulting in gonosomal mosaicism for D4Z4 (Figure 4b). Most information on the mechanism of D4Z4 rearrangements was obtained by studying mitotically rearranged D4Z4 repeats in *de novo* FSHD kindreds. A unique feature of mitotic D4Z4 rearrangements is the high frequency (25% of all cases) of gene conversion associated with crossover.

Homologous D4Z4 repeats required for the rearrangement can be found on the sister chromatid, the chromosome 4 homolog, and also on chromosome 10q26. Moreover, the repetitive nature of D4Z4 could enable an intrachromatid DSB repair within the D4Z4 repeat array (Figure 5a). In all studied D4Z4 rearrangements, DNA markers proximal (*Pvu*II-RFLP) and distal to D4Z4 (4qA/4qB) showed no allelic exchanges. Furthermore, a significant difference was shown between the D4Z4 repeat length distribution of 4qA and 4qB alleles. Together, these observations exclude a frequent recombination between D4Z4 repeats from chromosomes 4qA and 4qB. Most of the studied mosaic individuals were heterozygous for the 4qA/4qB polymorphism on chromosome 4. Because we never observed a change of the alleletype as a result of the D4Z4 rearrangement, even by gene conversions with crossover, it was suggested that mitotic D4Z4 rearrangements generally occur intrachromosomally (Chapter 7, Figure 5b).[62]

Most mitotic D4Z4 rearrangements can be explained by a gene conversion without crossover as in the SDSA model. However, about 25% of the studied mitotic rearrangements are explained by a gene conversion mechanism associated with crossover (Figures 4b and 6). This is in contrast to many other gene conversions studied that appear to be rarely associated with crossovers.[89] The D4Z4 rearrangements that were accompanied by crossover can be explained by the DSB repair model or by SDSA models that allow crossover.[6,84,89] In all models, repair of a DSB leads to the formation of two Holliday junctions, of which cutting and resolving may result in either crossover, or no crossover.[43]

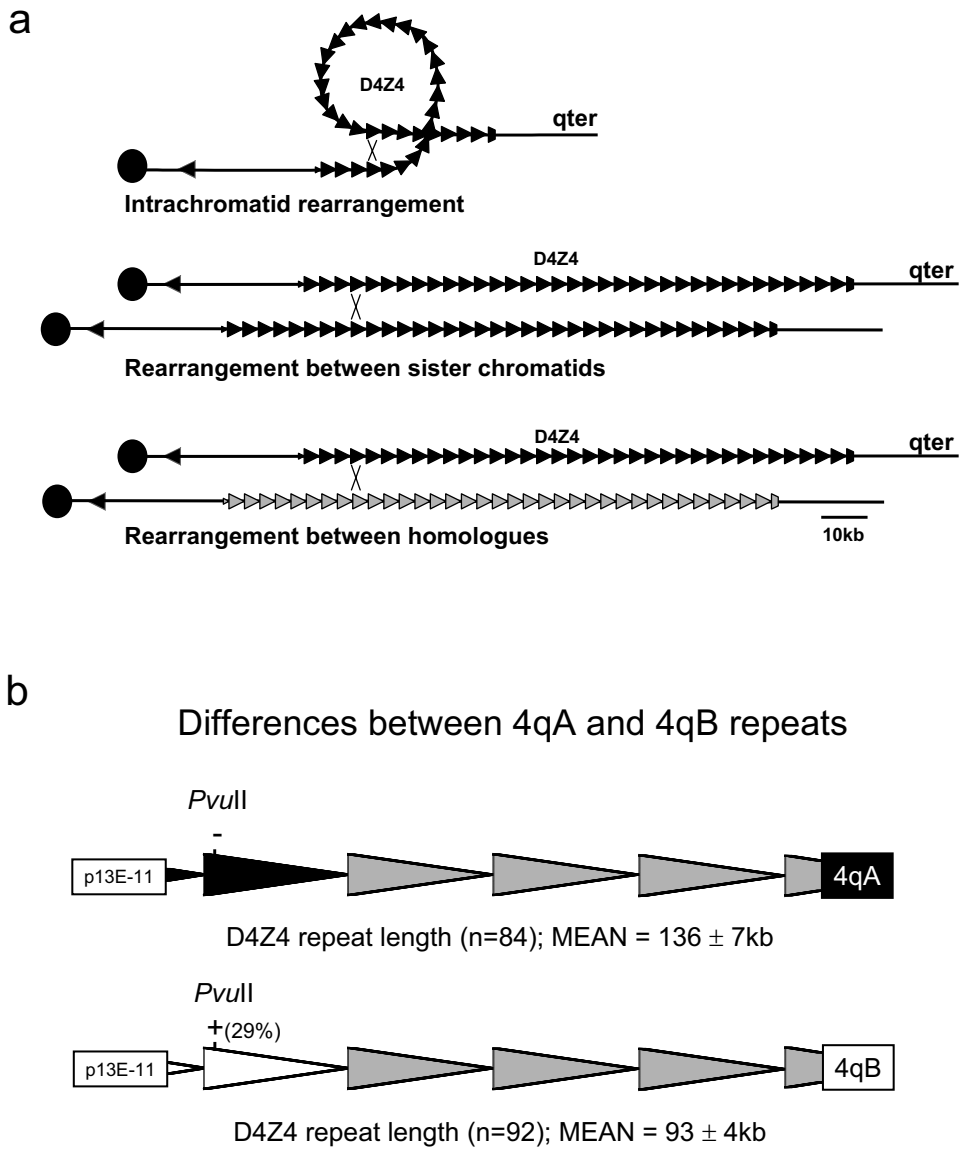


Figure 5 a) DNA break repair of a repetitive sequence by homologous recombination requires pairing of the damaged DNA with an homologous DNA sequence. Homologous D4Z4 repeats can be found on the sister chromatid, the homologous chromosome 4 or on chromosome 10q26 (Figure 1). In addition, the formation of a D4Z4 loop enables an intra-allelic break repair.
b) To reveal the common partner for D4Z4 rearrangements the D4Z4 repeat length distribution of 4qA and 4qB alleles were compared. The widely different length distribution ($p<0.001$) indicates that 4qA and 4qB alleles do not often interact. Furthermore, a *Pvu*II-RFLP within individual D4Z4 repeat units was studied. *Pvu*II was present in 29% of the most proximal D4Z4 units of 4qB repeats and in only 1% of 4qA repeats. These results suggest that D4Z4 rearrangement generally occur intrachromosomally.

1.3.5 Timing of mitotic D4Z4 rearrangement

The co-existence of mosaicism for D4Z4 in peripheral blood lymphocytes (PBL), muscle and fibroblast cells, as well as in the germline, indicates that mitotic D4Z4 rearrangements occur early in embryogenesis.[31,53,61] As described in chapter 7, detailed analysis of eleven mosaic individuals revealed eight D4Z4 rearrangements by a gene conversion without crossover and three with crossover (Figure 6a).[62] From the latter group, one patient (Family 36) displayed a mosaic mixture of a contracted and an expanded D4Z4 repeat in almost equal proportions of PBL. The absence of the mosaic parental sized D4Z4 repeat in this patient indicates that the rearrangement has occurred before the first embryogenic cell division (Figure 4b (GC+CO (1)) and 6b). Analogously, five patients (families 5, 6, 24, 26 and 35) with mitotic D4Z4 gene conversion without crossover that displayed equal proportions of cell populations, most probably also occurred at this period (Figure 4b (GC-CO) and 6b). The gene conversion without crossover for families 7 and 55611 may have occurred before the second cell division (expected percentages: 75% for parental and 25% for *de novo* alleles), which is also the time point that gene conversion with interchromatid crossover may have occurred for family 1 (expected percentages: 50% for parental, and 25% for both *de novo* alleles) (Figure 4b (GC+CO (2)) and 6c). Mitotic D4Z4 rearrangements that occurred at a later stage of development, will generally result in *de novo* mosaic alleles in <25% of the cells as detected in asymptomatic carriers of the FSHD allele.[115] In family Rf120 and 12 the two *de novo* cell populations are not equally present (Figure 6a). This is because the proportion of affected cells depends both on the timing of the rearrangement and on stochastic events related to which of the early embryonic cells contribute to the embryo and which cells in further stages of embryogenesis contribute to the different tissues (Section 1.5.2).

Different mechanisms could underlie the apparent early occurrence of D4Z4 rearrangements. The one-cell embryo is formed from two highly different sets of chromatin; sperm chromatin is highly compacted by protamines, whereas oocyte chromatin is much less condensed.[52,96] From the one-cell up to the four-cell embryonic stage, the maternal and paternal chromosomes do not mix and still display characteristics of their gonadal cells regarding DNA methylation. Immediately after fertilization, methylation of paternal DNA is rapidly reduced, whereas maternal DNA displays a replication-dependent DNA demethylation (Figure 7a).[39,96] It is tempting to speculate that changes in DNA conformation result in DNA breaks that initiate mitotic D4Z4 rearrangements.[8]

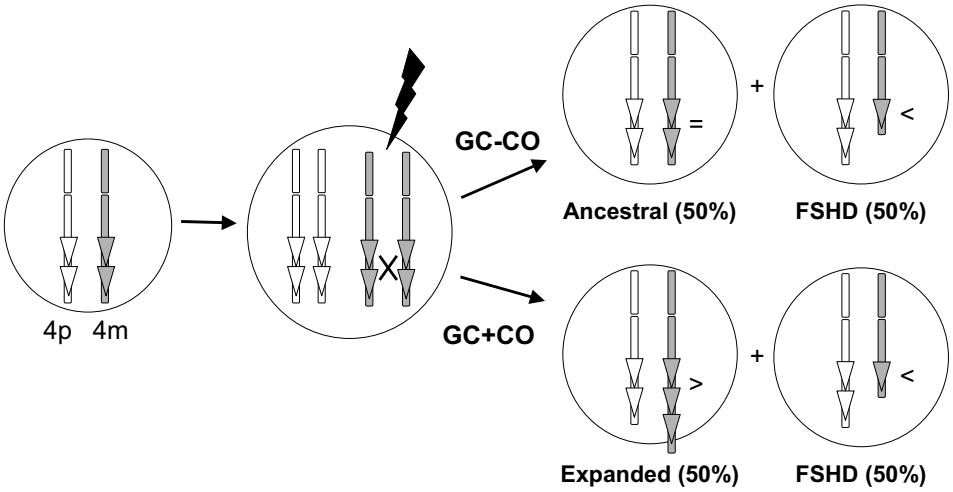
Surprisingly, amongst the patients in whom we suggest that the D4Z4 rearrangement occurred before the first cell division, we observed a rearrangement of the ancestral maternal allele in six out of seven cases (Figure 7b). While being close to significant ($p=0.058$, Fisher's exact test), this suggests a preference for the maternal allele being rearranged in the earliest stages of embryogenesis. This observation seems in contrast to the meiotic minisatellite instability showing a preference for the paternal allele (Section 1.3.3).

a Analysis 11 mosaic patients with *de novo* FSHD

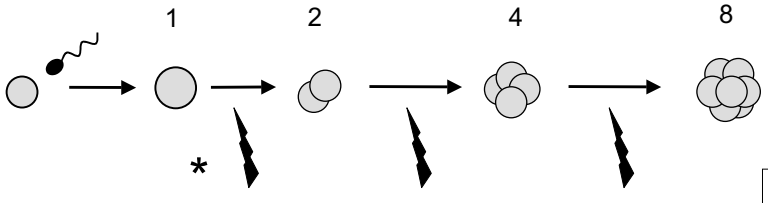
GC - CO				
family	gender	ancestral	FSHD	*de novo* 2
Fam 5	male	40%	60%	0%
Fam 6	male	60%	40%	0%
Fam 12	female	10%	90%	0%
Fam 24	male	50%	50%	0%
Fam 26	male	50%	50%	0%
Fam 35	male	50%	50%	0%
Fam 7	male	70%	30%	0%
Fam 55611	male	80%	20%	0%

GC + CO				
family	gender	ancestral	FSHD	*de novo* 2
Fam 36	male	0%	60%	40%
Fam 1	male	37%	26%	37%
Rf 120	female	20%	70%	10%

b D4Z4 rearrangement before 1st zygotic division (*)



c Proportion FSHD cells after D4Z4 contraction in first few zygotic divisions



	P	F	E	P	F	E	P	F	E
GC - CO	50%	50%	0%	75%	25%	0%	88%	13%	0%
GC + CO	0%	50%	50%	50%	25%	25%	75%	13%	13%

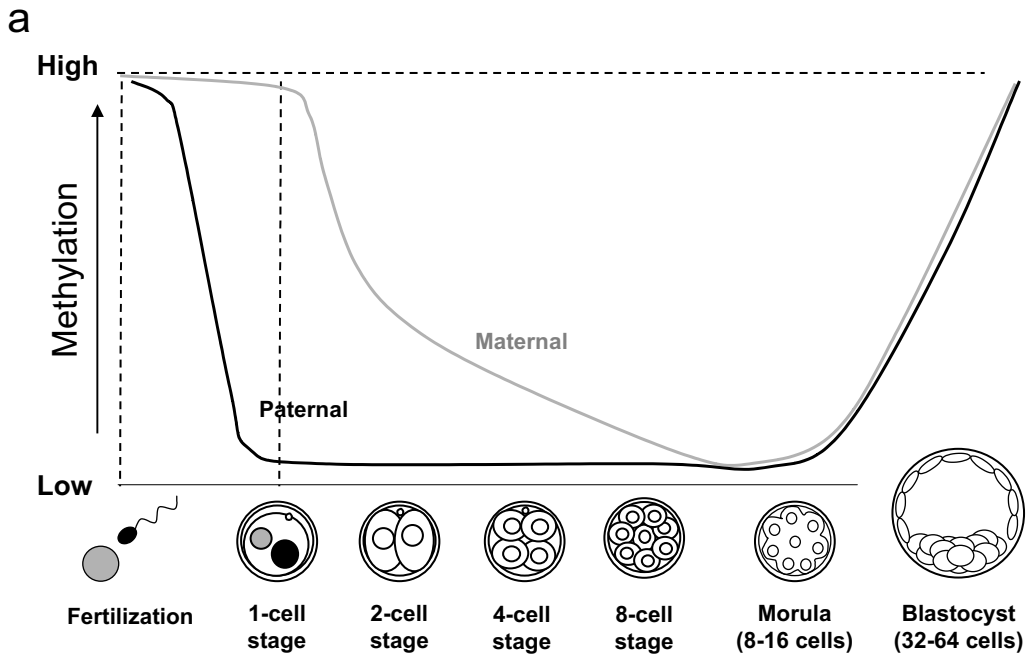
P = parental cell population (unchanged D4Z4)
F = FSHD cell population (contracted D4Z4)
E = extra cell population (expanded D4Z4)

Possibly, this maternal preference is only visible for D4Z4 rearrangements before the first zygotic cell division because then maternal and paternal genomes display the largest differences (Figure 7a).[39,74]

Previously, an increased incidence of 10;4 translocations (chromosome 10 with 4-type D4Z4 repeat, Section 1.2.5) has been reported in Dutch FSHD patients with a mitotic D4Z4 contraction.[115] Seven (27%) 10;4 translocated chromosomes were detected in thirteen D4Z4 mosaic individuals[115], while on average, the frequency of these translocated chromosomes is 6.5% in the control population.[126] Recently we analyzed a larger group of individuals with a mitotic D4Z4 contraction (FSHD patients and carriers) for 4;10 translocations using the chromosomal assignment method described in section 1.4. Seven (17%) translocated 4-type repeats on chromosome 10 were identified in 21 individuals with a mitotic D4Z4 contraction (unpublished results). Although the proportion of translocated chromosomes is lower than initially found, it is still significantly higher than in controls ($p=0.019$, Fischer's exact test). This observation was confirmed in a Chinese patient group in which 18.8% 10;4 translocated chromosomes were observed against 4.5% in the Chinese control population.[148] These data suggest that translocated 4-type repeats on chromosome 10 enhance mitotic D4Z4 contractions. Recently, the lack of FSHD patients amongst Black South Africans has been explained by an enrichment of 4;10 translocated repeats in this population.[82] Surprisingly, these observations seem in contrast with the D4Z4 rearrangement mechanism described in chapter 7, in which all mitotic rearrangements occurred within the chromosome.[62]

Figure 6 a) Detailed D4Z4 analysis of eleven mosaic FSHD patients revealed a mosaic mixture of a contracted FSHD-sized repeat and the unchanged donor repeat in eight cases, which is suggestive of a mitotic gene conversion without crossover (GC-CO). However, in three cases the D4Z4 rearrangement resulted in two different sized D4Z4 repeats, indicative of a gene conversion with crossover (GC+CO).
b) Most of the mosaic GC-CO cases and one of the GC+CO cases suggest that the mitotic D4Z4 contraction occurred before the first zygotic division. As shown in the figure, GC-CO zygotic D4Z4 contractions give rise to equally present cell populations with ancestral-sized and FSHD-sized repeats (Families 5, 6, 24, 26, 35). When the D4Z4 rearrangement is accompanied by a crossover, then one of the two cell populations has an expanded repeat and the other a contracted repeat (Family 36).
c) If the GC+CO rearrangements of D4Z4 occur after the first divisions then not only are the cell populations with expanded and contracted D4Z4 repeats present, but also cells with unchanged parental-sized repeats. The proportion of mitotic *de novo* FSHD cells becomes smaller for both GC-CO as GC+CO rearrangements when they occur later during embryogenesis.

A previous FISH analysis of interphase PBL of a limited set of FSHD patients and controls suggested an increased somatic pairing of the subtelomeric regions of chromosomes 4q and 10q in patients.[101] However, subtelomeric domains of chromosomes 4 and 10 were recently shown to occupy distinct regions in the nucleus. Whilst the subtelomere of 4qter was localized near the nuclear periphery, those of 10q occupy the interior of the nucleus.[72] The nuclear peripheric localization of 4q35 was corroborated in another study.[104] It is well established that chromosomes occupy distinct territories in the mammalian nucleus and there is evidence that the spatial organization is largely retained after mitosis.[32] The apparent role of 10;4 translocated repeats in intrachromosomal D4Z4 rearrangements on chromosome 4 could partially be explained if the presence of translocated repeats on chromosome 10 influences its nuclear positioning. Possibly, dislocalized, translocated 4-type repeats on chromosome 10 have a subtle impact on the integrity of D4Z4 repeats. A study on the positioning of chromosomes 18 and 19, which occupy different positions in the nucleus, did not provide evidence for a disturbed localization of translocated 18;19 and 19;18 chromosomes. However, the relative orientation of the translocated arms of both derivative chromosomes seems to reflect their chromosomal origin.[21] In addition, a recent study demonstrated that the preference for the nuclear periphery is an intrinsic feature of 4qter as the presence of 4 Mb of 4qter sequences on the derivative chromosome X in a cell line with a 4;X translocation causes a more peripheral localization of this chromosome.[104]



GC - CO	ancestral	FSHD	*de novo* 2	
expected	50%	50%	0%	origin
Fam 5	40%	60%	0%	maternal
Fam 6	60%	40%	0%	paternal
Fam 12	10%	90%	0%	maternal
Fam 24	50%	50%	0%	maternal
Fam 26	50%	50%	0%	maternal
Fam 35	50%	50%	0%	maternal

GC + CO	ancestral	FSHD	*de novo* 2	
expected	0%	50%	50%	origin
Fam 36	0%	60%	40%	maternal

GC - CO	ancestral	FSHD	*de novo* 2	
expected	75%	25%	0%	origin
Fam 7	70%	30%	0%	paternal
Fam 55611	80%	20%	0%	maternal

GC + CO	ancestral	FSHD	*de novo* 2	
expected	50%	25%	25%	origin
Fam 1	37%	26%	37%	paternal
Rf 120	20%	70%	10%	maternal

Figure 7 a) Chromatin alterations during early embryogenesis. Detailed analysis of mitotic D4Z4 rearrangements suggests that they often occur during the first few zygotic division of embryogenesis. It is exactly during this period of development that considerable changes occur in DNA conformation and methylation. Possibly, these changes cause DNA breaks that initiate the D4Z4 rearrangements. Figure partially adapted from Santos and Dean.[96]
b) Based on the proportion of FSHD cells, D4Z4 genotypes of the 11 mosaic FSHD patients (from mitotic D4Z4 rearrangements by GC-CO and GC+CO) are divided in two groups, representing D4Z4 rearrangement before first and second zygotic division. Unexpectedly, six out of seven rearranged alleles from the first group appear to be of maternal origin. This might suggest a rearrangement preference for the maternal allele. Possibly, the differences in both DNA-methylation and packaging for maternal- and paternal-derived chromosomes underlies this apparent difference in rearrangement susceptibility

1.3.6 Repeat Homogenization (Concerted Evolution)

Another feature of repetitive sequences is the homogenization of individual units within a repeat array, which is called concerted evolution. Initially, concerted evolution was detected when a greater sequence similarity (if not identity) was detected between individual repeat units of a tandem repeat within each species, than between orthologous repeat units of closely related species. This suggested that repeat units within a repetitive family do not evolve independently from each other.[23,65] Concerted evolution has been described in many organisms for coding- and noncoding RNA multigene families. Since concerted evolution mostly involves abundant RNA or protein molecules (rRNAs, snRNAs and histones) it has been hypothesized that it functions to retain the high production of homogeneous transcripts.[65] On the other hand, some apparent cases of concerted evolution might in fact reflect recent repeat amplifications and transpositions. Furthermore, not all multigene families evolve in a concerted fashion. For example, some members of the major histocompatibility complex and immunoglobulin gene families are not more closely related to one another than to the orthologous genes from different species.[65]

In primates, concerted evolution has best been studied for the RNU2 locus harboring the U2 tandem repeat array (Section 1.3.3). The differences in the U2 repeat were established early in the primate lineage and concerted evolution has occurred over the last 35 million years.[90] More specifically, Old World monkeys have a U2 repeat array that consist of 11 kb repeat unit, and primates have a U2 repeat array that consists of 5 kb repeat unit. This difference in repeat unit length was caused by a 5 kb deletion in the primate copy. Subsequently, concerted evolution spread the 5 kb deletion from one unit to the whole repeat array in the ancestral primate genome.[65]

The human RNU2 locus is localized on chromosome 17q21–22 and contains 5–40 U2 repeat units. The number of U2 repeat units varies extensively between different repeats, suggesting a high level of repeat instability. Analysis of individual markers has indicated that individual units within a single U2 repeat array are entirely homogeneous for specific polymorphisms. However, homologous chromosomes carrying different homogenous U2 repeats can be found in any combination, indicating that repeat homogenization primarily occurs intrachromosomally. The markers flanking the U2 repeat have been shown to display only two haplotypes (left+ and left-, right+ and right-). Surprisingly, left+ was always associated with right+, and left- with right-. Therefore, it was concluded that individual U2 repeats do not exchange flanking markers and interchromosomal rearrangements most probably occur infrequently and via gene conversions without crossovers. The homogenization of the U2 repeat most probably occurs via rapid intrachromosomal gene conversions with or without crossover (Figure 8).[65]

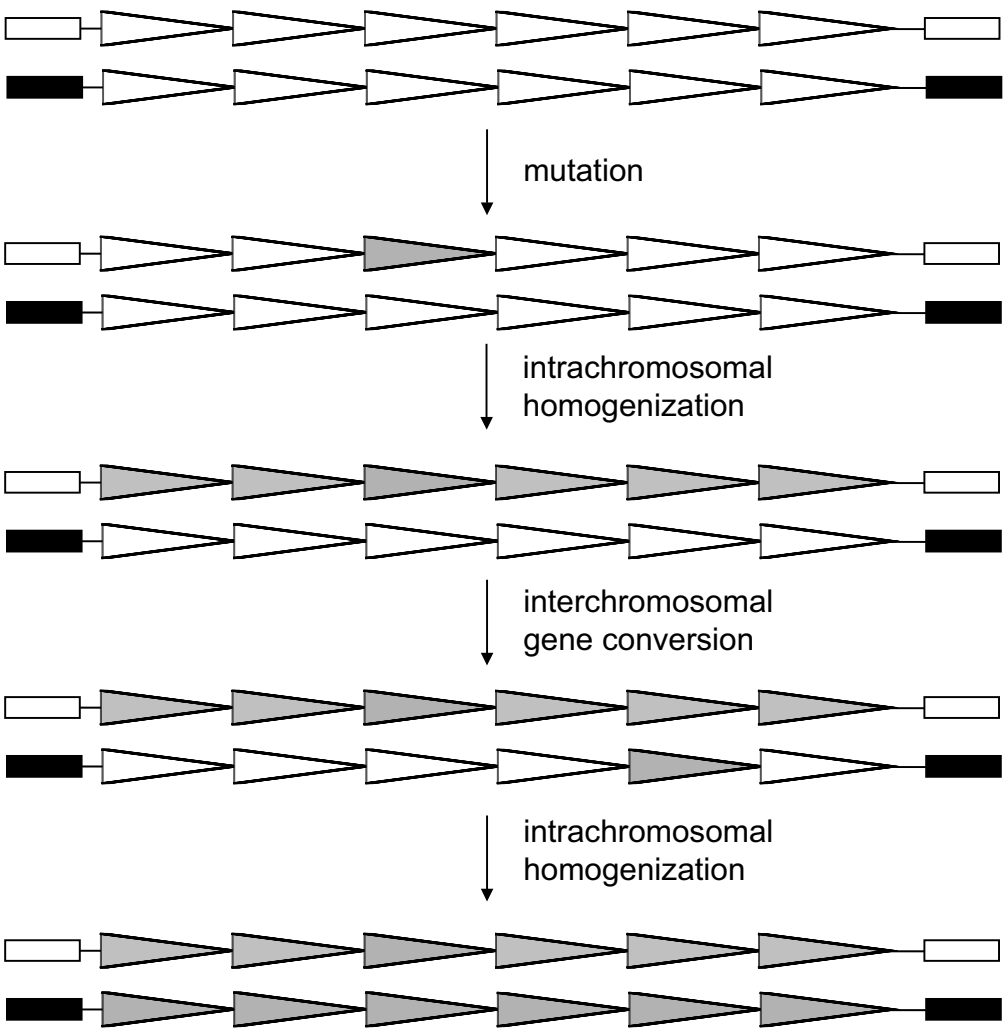


Figure 8 A model for concerted evolution (repeat homogenization) of the RNU2 locus in humans and primates (adapted from Liao).[65] The U2 repeat array (open triangles) on two homologous chromosomes is depicted together with distinct flanking chromosomal DNA sequences. One repeat unit acquires a mutation (gray triangle), which is rapidly homogenized through the repeat array by intrachromosomal rearrangements. The mutation is then spread to the repeat on the homologous chromosome by an interchromosomal gene conversion that occurs at a much lower frequency than the intrachromosomal homogenization. Subsequently, the homologous repeat is rapidly homogenized. Most probably, repeat homogenization occurs by intrachromosomal gene conversions with or without crossover.

1.3.7 Evolution of D4Z4 in human

As discussed in section 1.2.4, D4Z4 has been subjected to multiple duplications during primate evolution. While Old World monkeys only have one D4Z4 locus, humans and all great apes have D4Z4 copies on chromosomes 4q35, 10q26 and the pericentromeric regions of the acrocentric chromosomes.[17]

Comparative sequencing of D4Z4 units from different chromosomal origins in humans (4qA, 4qB and 10q alleles), in combination with analysis of the distal 4qA/4qB polymorphism on chromosome 4 possibly permits elucidation of D4Z4 evolution. Therefore, individual *Kpn*I units of D4Z4 repeats from several chromosomes 4 and 10 have been sequenced and subsequently aligned. This alignment shows that the 3.3 kb D4Z4 sequences from chromosome 4 and 10 are 98% identical. Some allele-specific single nucleotide polymorphisms have been identified of which some generate allele-specific restriction sites. The restriction enzyme *Bln*I is specific for chromosome 10-derived D4Z4 repeat units[22] and the restriction enzyme *Xap*I is specific for chromosome 4-derived D4Z4 units (Chapter 9).[58] Both the restriction enzymes *Bln*I and *Xap*I are completely homogenized through chromosome 10- and 4-derived D4Z4 repeats, respectively. More recently a *Pvu*II-RFLP was detected in D4Z4 (Chapter 7).[62] The *Pvu*II restriction site has been detected in 29% of the most proximal D4Z4 units (the first unit of a repeat array, directly adjacent to p13E-11) in D4Z4 repeats on 4qB-type alleles, and in only 1% in this position on 4qA-type alleles. Conversely, this RFLP has not been homogenized through the whole D4Z4 repeats on 4qB-type alleles and as a consequence mixtures of *Pvu*II sensitive and resistant units can be found within one array. Moreover, also internal D4Z4 repeat arrays on 4qA-type alleles are hybrid for this *Pvu*II-RFLP (Figure 5b).

Van Geel proposed a model for the evolution of the 4q telomere (van Geel 2002, thesis), which can be refined using the new data for *Xap*I, *Pvu*II and 4qA/4qB (Figure 9). In this model the subtelomeric 4qA region, with D4Z4, was duplicated to chromosome 10q (Section 1.2.4). Most likely this duplication was followed by several single nucleotide changes within a single D4Z4 unit generating specific restriction sites *Xap*I and *Bln*I in D4Z4 repeats on chromosomes 4 and 10, respectively. Most probably, these repeat arrays were then homogenized by concerted evolution. Next, the 4p-telomeric region (B-type) was duplicated to the chromosome 4 region distal to D4Z4, generating the 4qA/4qB polymorphism. On the one hand, the presence of *Xap*I in D4Z4 repeats on both 4qA- and 4qB-type chromosomes (manuscript in preparation) indicate that the introduction of the distal 4qA/4qB polymorphism occurred after *Xap*I homogenization on 4q (Section 1.2.4). Otherwise, the *Xap*I site was present on the ancestral D4Z4 repeat, altered in a single D4Z4 unit on chromosome 10 after duplication, and subsequently spread over the repeat array.

Most probably, the *Pvu*II restriction site was created later in the D4Z4 evolution in a single D4Z4 unit on a 4qB allele and has not yet been completely homogenized throughout the entire repeat, possibly due to the short time span (Figure 1 in Chapter 7).[62] Most likely, an

interchromosomal gene conversion copied *Pvu*II from 4qB to 4qA and subsequently spread over this repeat by intrachromosomal rearrangements (Figure 9). The proximal D4Z4 unit, in contrast to the distal D4Z4 unit, on 4qA repeats has until now escaped this homogenization, suggesting that D4Z4 rearrangements are polarized to the distal end of the repeat array, as described for meiotic minisatellite rearrangements (Section 1.3.3).

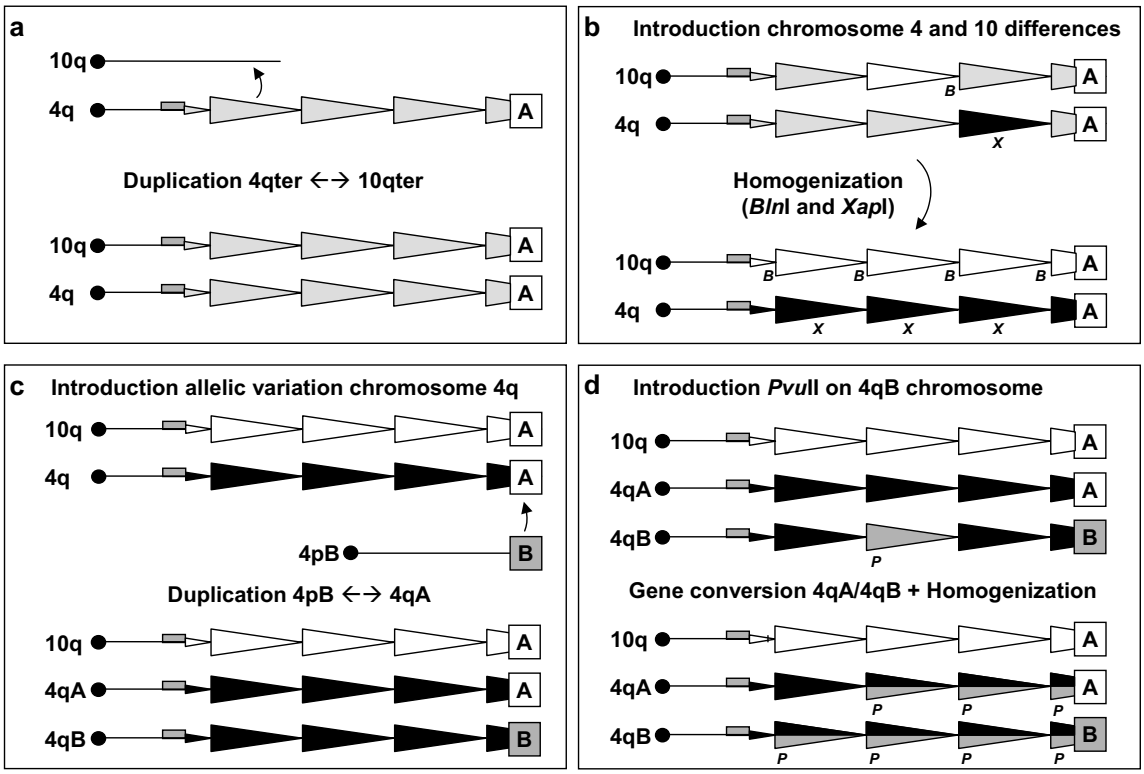


Figure 9 Proposed mechanism for the evolution of the 4q telomere harboring the D4Z4 repeat.
a) First the telomeric 4qA region, with D4Z4, was duplicated to chromosome 10q.
b) Introduction of 4- and 10-specific restriction sites *Xap*I and *Bln*I within single D4Z4 units, followed by intra- and interchromosomal homogenization created 4- and 10-repeat specificity.
c) Introduction of 4qA/4qB allelic variation by the duplication of the 4p-telomeric region (B-type) to the chromosome 4 region distal to D4Z4.
d) Finally the *Pvu*II restriction site was introduced. Possibly due to a lack of time, *Pvu*II has not been completely homogenized throughout D4Z4 repeats on 4qB chromosomes. Therefore most 4qB-type D4Z4 repeats are polymorphic for the *Pvu*II-polymorphism between individual D4Z4 units. Also internal D4Z4 units on 4qA chromosomes are polymorphic for *Pvu*II, but surprisingly the most proximal D4Z4 unit lacks the *Pvu*II restriction site (Chapter 7). This suggests that *Pvu*II was most probably introduced in a D4Z4 repeat unit on a 4qB chromosome, and after incomplete homogenization on 4qB, has more recently been converted to the D4Z4 repeat on chromosome 4qA.