Transcriptional analysis of the human D4Z4 and mouse Dux arrays

Amanda Hampson, BSc

Institute of Genetics

A thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy, December 2011

Declaration

I declare that this thesis is the result of my own work and has not, whether in the same or different form, been presented to this or any other university in support of an application for any degree other than that for which I am now a candidate.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Firstly, thank you to Jane for giving me the opportunity to work on this project, and to both Jane and Liz for their supervision and excellent support for which I am very grateful. Thank you also to all of the JEH lab members, both past and present, who have shared advice and feedback and provided daily entertainment.

A big thank you to all the family and friends who have given technical support, laughter and hugs when required! A special thanks to everyone who has provided babysitting over the last few years, in particular, to my mum, dad and Emma; this may be cliché, but I couldn't have done it without you and your help has been greatly appreciated.

Finally, a huge thank you to Thomas; expecting a 5/6 year old to wait for things 'until I have finished my work' is a big ask but you have understood and you have been awesome. Not only have you been really well behaved these last few months but you have even started to bring me drinks and sandwiches 'in case I need them'. You are amazing and this is for you.

Abstract

Facioscapulohumeral muscular dystrophy (FSHD) is the third most common form of muscular dystrophy in Caucasians. FSHD is caused by contraction of a 3.3kb repeat array, D4Z4, to below 11 repeat units. Each of these repeat units contains an ORF encoding the DUX4 gene and at the beginning of the work described in this thesis expression of transcripts from the most distal repeat of this D4Z4 array had been reported. However, expression data from the DUX4 gene was new and most research focussed on contraction of the array having a position affect on the expression of neighbouring genes.

There is a similar Dux array located in the mouse genome. This array also contains an ORF in each repeat and transcripts from this gene have been detected in a number of different tissues. This array is not present in the same chromosomal location as the D4Z4 array in humans so is unlikely to be a true ortholog, however it is the only Dux array in the mouse genome and may therefore be functionally equivalent.

The work described in this thesis provides evidence for transcription from multiple repeats of the D4Z4 array in a human embryonal carcinoma cell line and gives information on the sequence variation within these transcripts. In addition, this work contributes to an understanding of expression from the mouse Dux array. Similar to the human array, expression appears to come from multiple repeat units, and analysis of the sequences amplified by RT-PCR identifies variants which suggest some of these transcripts are non-coding.

The aim of this work is to determine whether human DUX4 and mouse Dux genes have equivalent functions with the long-term goal of producing a mouse model for FSHD

Chapter 1. Introduction

1.1 Identification of the causative mutation in FSHD

With an estimated incidence of 1/20 000 (Padberg, 2004a), facioscapulohumeral muscular dystrophy (FSHD) is the third most common muscular dystrophy in Caucasians. While most patients begin to show symptoms in the second decade of life, the age of onset is variable and early onset cases tend to be more severe (Brouwer *et al.*, 1995; Lunt *et al.*, 1995a; Tawil *et al.*, 1996; Klinge *et al.*, 2006). FSHD has an estimated penetrance of 95% by the age of 20 (Lunt *et al.*, 1989).

The facial muscles are usually affected first; characteristic weakening of muscles associated with the eyes and mouth cause difficulty in forming facial expressions (Padberg, 2004b). The weakness then usually spreads to the upper arms and shoulder girdle. Scapular winging may be seen and patients become unable to raise their arms above shoulder height (Padberg, 2004b). In 20% of cases the lower leg muscles are also involved and patients become wheelchair dependent (Lunt and Harper, 1991). The more severe, early onset cases are sometimes associated with non-muscular symptoms including hearing loss (Padberg *et al.*, 1995; Miura *et al.*, 1998), mental retardation (Funakoshi *et al.*, 1998; Miura *et al.*, 1998), epilepsy (Funakoshi *et al.*, 1998; Miura *et al.*, 1998) and retinal vasculopathy (Padberg *et al.*, 1995).

Locating the genetic locus for FSHD initially proved difficult. An international consortium in the late 1980s produced an exclusion map for the disease (Sarfarazi *et al.*, 1989). It was not until microsatellite markers were used in 1990 that linkage was established to the long arm of chromosome 4 (Wijmenga *et al.*, 1990). Over the next couple of years linkage to chromosome 4 was confirmed and narrowed to the 4q35 region (Wijmenga *et al.*, 1991; Fischbeck and Garbern, 1992; Gilbert *et al.*, 1992; Mathews *et al.*, 1992; Mills *et al.*, 1992; Sarfarazi *et al.*, 1992; Upadhyaya *et al.*, 1992; Weiffenbach *et al.*, 1992; Wijmenga *et al.*, 1992c).

Finally, a breakthrough in the search for the causative mutation came when the probe p13E-11 was identified in a search for homeobox genes (Wijmenga *et al.*, 1992b). This probe is a subclone of a chromosome 4 cosmid (13E) and identifies a polymorphic *Eco*RI fragment on chromosome 4q35. In familial cases of FSHD, the smallest fragments on Southern blots co-segregated with the disease (Wijmenga *et al.*, 1992b). In most sporadic cases of FSHD, p13E-11 was found to identify small, *de novo* fragments (Wijmenga *et al.*, 1992b; Griggs *et al.*, 1993; Upadhyaya *et al.*, 1993) and these new mutations were transmitted to affected children in the next generation (Wijmenga *et al.*, 1992a; Jardine *et al.*, 1993).

Restriction mapping of cosmid 13E suggested a repeated DNA structure and cloning of several rearranged fragments from patients identified a deletion of an integral number of 3.3kb repeat units (van Deutekom *et al.*, 1993b). The FSHD mutation had been identified (Fischbeck and Garbern, 1992) .

1.2 Molecular analysis of D4Z4

Cloning of the p13E-11 positive *Eco*RI fragments revealed a polymorphic array of 3.3kb *Kpn*I repeats (van Deutekom *et al.*, 1993b; Hewitt *et al.*, 1994) that has been named D4Z4 (Figure 1.1). Each repeat is GC rich (71% G+C) and contains two dispersed repetitive motifs, *LSau* and hhspm3 (Hewitt *et al.*, 1994), which are typically associated with heterochromatic regions of DNA (Zhang *et al.*, 1987; Meneveri *et al.*, 1993). Each repeat has a large open reading frame (ORF) which potentially encodes two homeodomains related to the *prd* family (Hewitt *et al.*, 1994).

In unaffected individuals, the length of the D4Z4 array varies between 11 and 100 repeats, an equivalent *Eco*RI fragment of 40kb-300kb on southern blots. FSHD patients have arrays of 10 or fewer repeats, seen as fragments of only 10-40kb (Wijmenga *et al.*, 1992b). Alleles containing only 1-3 repeats are typically associated with the most severe early onset cases (Zatz *et al.*, 1995; Tawil *et al.*, 1996). However, complete deletion of the array is not associated with FSHD (Tupler *et al.*, 1996).

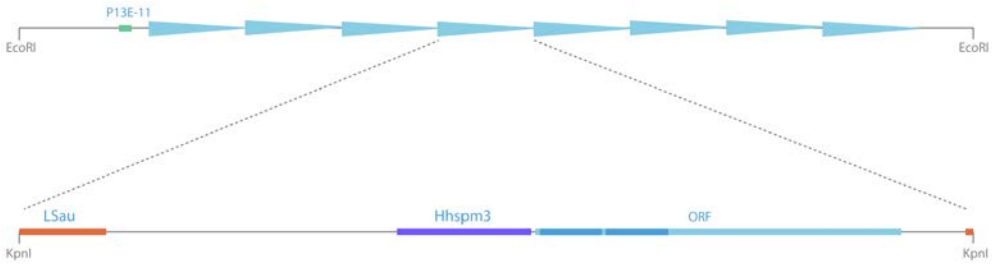


Figure 1.1 A schematic representation of the *Eco*RI fragment identified by p13E-11

Each arrowhead represents one 3.3kb *Kpn*I repeat, the number of repeats in the *Eco*RI fragment varies between 12 and 150 in unaffected individuals. The sequence motifs that are present within each individual repeat are shown and the double homeodomain within the open reading frame is highlighted in dark blue.

Approximately ten percent of FSHD patients are sporadic (Lunt and Harper, 1991) and most of these were shown to have *de novo* contraction of D4Z4 (Wijmenga *et al.*, 1992b). These mutations are thought to occur intrachromosomally, facilitated by the repetitive nature of the array (Lemmers *et al.*, 2004a). The most likely mechanism is thought to be gene conversion without crossover. However, in some cases reciprocal transfer is seen and a synthesis-dependent strand annealing (SDSA) model (Nassif *et al.*, 1994) has been proposed which allows for both cases (Lemmers *et al.*, 2004a). In this model, newly synthesised DNA does not remain base paired with the homologous template during the repair of double strand breaks (DSB). Instead, the DNA is displaced, which allows the two newly synthesised, complementary strands to anneal. In repeated sequences alignment of these strands can occur out of frame, which can result in either expansion or contraction of the region. In most cases the donor template will remain unchanged (Figure 1.2).

Examination of unaffected parents of *de novo* cases indicates that 15-20% show somatic mosaicism for the deletion (Kohler *et al.*, 1996; van der Maarel *et al.*, 2000), suggesting that the rearrangement occurs mitotically during early embryogenesis. The clinical presentation in mosaic individuals has been shown to depend on both the number of repeats left on the deleted allele and the percentage of peripheral blood lymphocytes (PBL) that carry the short D4Z4 array (van der Maarel *et al.*, 2000). A later study showed that the proportion of affected PBL and muscle cells were comparable (Tonini *et al.*, 2006). A significant percentage of mosaic individuals are asymptomatic, which suggests that there is a critical proportion of mutated cells needed for an FSHD phenotype. Consistent with this idea, the proportion of mutated cells correlates with the severity of the phenotype (van der Maarel *et al.*, 2000). It has also been shown that females need a higher proportion of cells carrying the mutation than males for FSHD symptoms to occur (van der Maarel *et al.*, 2000).

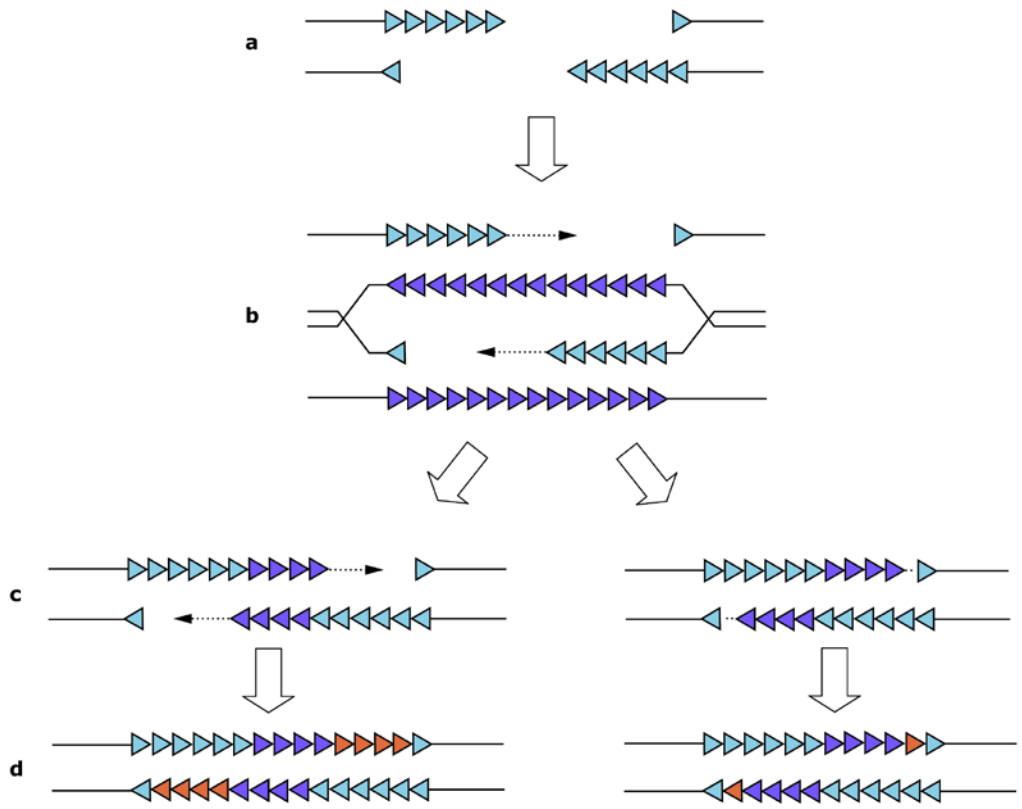


Figure 1.2 The SDSA model of double strand break repair.

a) The ends of the broken sequence (blue arrowheads) are degraded prior to invasion. b) The ends invade a homologous sequence (purple arrowheads) and DNA synthesis then proceeds from each 3' end. c) Newly synthesised DNA is displaced from the template and the two single strands align in a region of overlap. If alignment occurs out of frame contractions or expansions may be seen. d) DNA synthesis continues (orange arrowheads) with each newly made strand serving as a template for the other. See Nassif *et al*. (1994).

1.3 The 3.3kb repeat family

Initial southern blot data showed p13E-11 to hybridise to more than one locus (Wijmenga *et al.*, 1992b); a highly similar tandem array was then identified at chromosome 10q26 (Bakker *et al.*, 1995). Additionally, the short arms of the acrocentric chromosomes contain a high number of dispersed repeats similar to D4Z4 (Wijmenga *et al.*, 1992b; Hewitt *et al.*, 1994; Winokur *et al.*, 1994). This group of repeat sequences are collectively known as the 3.3kb repeat family and show two distinct types of organisation (Lyle *et al.*, 1995). On the acrocentric chromosomes and on chromosomes, 1, 3, 9 and Y the repeats are interspersed with β satellite DNA, whilst on 4q35 and 10q26 the sequences have a very homogenous tandemly repeated structure with the β satellite DNA only at the distal end of the array (Lyle *et al.*, 1995). Importantly, despite >98% nucleotide identity between the 4q35 and 10q26 arrays (Cacurri *et al.*, 1998), only contractions of the 4q35 repeat array are associated with FSHD (Bakker *et al.*, 1995).

Initially the presence of this 10q array complicated genetic diagnosis of FSHD, since small 10q fragments (which do not result in disease) could not be distinguished from those arising from 4q35 (Weiffenbach *et al.*, 1993). In 1996 a *Bln*I recognition sequence was found to be specific for the 10q repeats (Deidda *et al.*, 1996). For diagnostic analysis, a double digest (*Eco*RI/*Bln*I) is now performed so that only the 4q fragment remains intact. The restriction enzyme *Xap*I has since been shown to uniquely digest 4q derived repeat units, and in 2001, a triple analysis with *Eco*RI, *Eco*RI/*Bln*I and *Xap*I was proposed to optimise the molecular diagnosis of FSHD (Lemmers *et al.*, 2001).

In most cases this double-digestion protocol resolves ambiguity, however, variant alleles with both *Bln*I^{+} and *Bln*I^{-} repeats have been identified in 10-30% of individuals (van Deutekom *et al.*, 1996a; Lemmers *et al.*, 1998; van Overveld *et al.*, 2000; Rossi *et al.*, 2007). Variant alleles are also seen in FSHD patients (van Deutekom *et al.*, 1996a), with *Bln*I^{+} repeats present on the disease allele. Variation in the number of *Bln*I^{+} and *Bln*I^{-} repeats on these alleles is likely due to intrachromosomal sequence exchanges, which is discussed in more detail in Section 1.5.

1.4 Association of FSHD with a specific 4q telomere

There are two distinct allelic variants of the 4q telomere, 4qA and 4qB (van Geel *et al.*, 2002), which have frequencies of 42% and 58% in the population (Lemmers *et al.*, 2002). 4qA and 4qB represent the 25-40kb immediately distal to the D4Z4 array and are closely related to each other, sharing 92% nucleotide identity (Dickson, 1998; van Geel *et al.*, 2002). The 4qA telomere has 8kb of β-satellite sequence distal to D4Z4 adjacent to a divergent $(TTAGGG)_n$ array (van Geel *et al.*, 2002). Neither of these sequence elements are present on the 4qB telomere (Figure 1.3). The terminal D4Z4 repeat also differs between the variants, in 4qB this is only 570bp in length and directly abuts the subtelomeric sequence whilst the 4qA terminal repeat is adjacent to the β-satellite array (van Geel *et al.*, 2002). Shortly after the publication of this polymorphism (van Geel *et al.*, 2002), it was shown that only 4qA alleles are associated with FSHD (Lemmers *et al.*, 2002) and this was subsequently confirmed in a separate patient cohort (Thomas *et al.*, 2007). In three families, contraction of D4Z4 on a 4qB allele was shown not to segregate with the disease (Lemmers *et al.*, 2004b), arguing that the association with 4qA is not due to a difference in the frequency of deletions between the two.

To identify the subtelomere variants, DNA is digested with *Hind*III, which results in a fragment containing p13E-11, the D4Z4 array and the subtelomeric region. Probes specific to the 4qA or 4qB alleles are then used for pulsed-field gel electrophoresis (PFGE) analysis (Lemmers *et al.*, 2002). An additional rare type of 4qter telomere has been identified in four individuals (two with FSHD), which does not hybridise to 4qA or 4qB probes (Buzhov *et al.*, 2005; Thomas *et al.*, 2007). These variants have been referred to as 4qC telomeres (Lemmers *et al.*, 2010b).

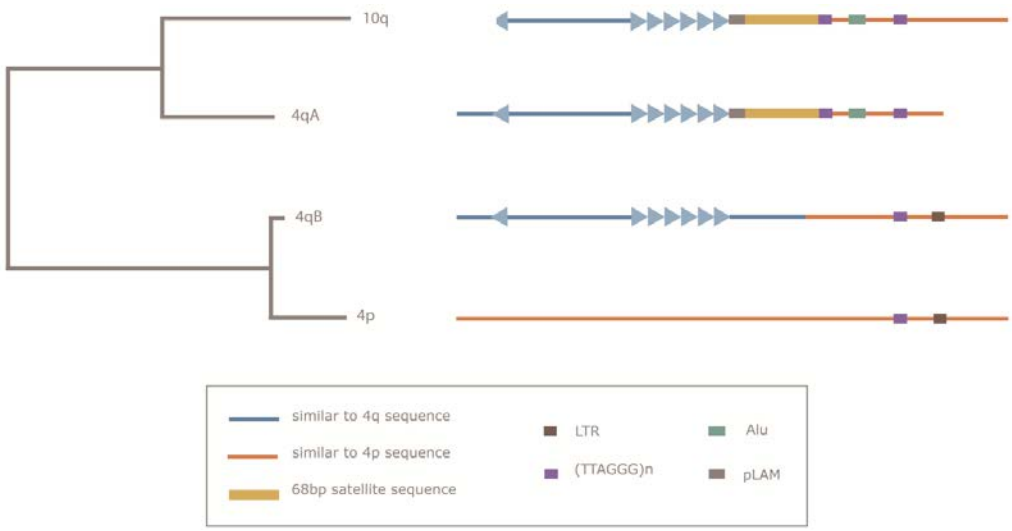


Figure 1.3 A representation of the relationship between the 4p, 4qA, 4qB and 10q telomeres

Not to Scale. The phylogenetic tree is based on sequences distal to D4Z4. Regions of sequence similarity and shared motifs are shown. Adapted from van Geel *et al*. 2002.

1.5 The evolution of the 4qA, 4qB and 10q subtelomeres

Studies of D4Z4 homologues in great apes (Clark *et al.*, 1996; Winokur *et al.*, 1996) and old and new world monkeys (Clapp *et al.*, 2007) suggest that the 4q array is ancestral. The DNA sequence and repeat content of the 4qA subtelomeric region is more similar to 10q than to 4qB (van Geel *et al.*, 2002), while the 4qB allele is more similar to the 4p telomere, lacking some of the repetitive sequence elements which 4qA and 10q share.

Both the 4q and 4p subtelomeres share sequence similarity and van Geel *et al.* (2002) proposed that a duplication event involving 4p created the 4qA telomere, which was subsequently duplicated onto 10q. The 4qB variant is likely to be the result of a more recent transfer from 4p (van Geel *et al.*, 2002). The relationship between these chromosome ends is shown in Figure 1.3. Recently, Lemmers *et al.* (2010) have produced an evolutionary network of 4qter evolution that broadly supports this model. In this network four major rearrangements occur at the D4Z4 locus, once introducing the 4qB subtelomere onto a chromosome 4qA background, and three times transferring the 4q telomere onto chromosome 10 (Lemmers *et al.*, 2010b).

The different size distribution of the 4qA, 4qB and 10q alleles argues against frequent interallelic recombination since these major transfer events. Lemmers and colleagues (2004) compared the distribution of the D4Z4 copy number of 4qA and 4qB derived repeats and showed that 4qA alleles (136 ± 7kb) were significantly longer than 4qB alleles (93 ± 4 kb). Linkage disequilibrium analysis between a *Pvu*II-RFLP within the most proximal D4Z4 repeat unit and the 4qB variant also suggests that recombination between the alleles is repressed (Lemmers *et al.*, 2004a). A difference in average D4Z4 copy number has also been observed between 4q and 10q arrays, van Overveld *et al.* found the median repeat array on chromosome 4 to be 21kb larger than that for chromosome 10 (van Overveld *et al.*, 2000), while Rossi *et al.* showed a 16kb increase in chromosome 4 array size compared with chromosome 10 (Rossi *et al.*, 2007). These observations suggest that the 4qA, 4qB and 10q subtelomere regions are evolving independently and support an intrachromosomal mechanism for D4Z4 copy number variation.

Lemmers *et al.* (2010) propose that the *Bln*I^{+} D4Z4 unit first arose on a 4qA allele, forming a hybrid array of *Bln*I^{-} and *Bln*I^{+} repeats which was then transferred to chromosome 10. Variation in the distribution of *Bln*I^{-} and *Bln*I^{+} repeats will then have occurred via intrachromosomal sequence exchanges.

1.6 Phenotypic FSHD (FSHD2)

Some FSHD families do not carry a short *Eco*RI fragment and represent a genetically distinct form of the disease, known as FSHD2. However, FSHD2 patients are clinically indistinguishable from FSHD1 patients (de Greef *et al.*, 2010). Linkage studies have failed to identify the genetic locus for FSHD2 (Iqbal *et al.*, 1992; Gilbert *et al.*, 1993; Yamaoka *et al.*, 1995; Bastress *et al.*, 2005). More detailed genetic analysis showed that a small proportion of these individuals can be accounted for by 4q35 deletions that also remove the p13E-11 probe (Lemmers *et al.*, 2003; Deak *et al.*, 2007).

1.7 The FSHD disease mechanism

Many groups initially searched for transcripts from the D4Z4 array and although related cDNA sequences were identified none originated from the 4q35 locus (Hewitt *et al.*, 1994; Altherr *et al.*, 1995; Lyle *et al.*, 1995; Gabriels *et al.*, 1999; van Geel *et al.*, 1999). In addition, as monosomy of 4q does not result in FSHD (Tupler *et al.*, 1996), haploinsufficiency of a gene within the repeat was considered to be an unlikely mechanism for the disease. Thus, D4Z4 was generally assumed to be a pseudogene and research efforts were concentrated on identifying other candidates.

Sequencing of the 4q35 region provided evidence for the involvement of chromatin organisation and epigenetic mechanisms in FSHD. The D4Z4 array is likely to reside in a heterochromatic environment. The *Kpn*I repeats contain *LSau* and hhspm3 motifs (Hewitt *et al.*, 1994), which are commonly found in heterochromatin (Meneveri *et al.*, 1985; Agresti *et al.*, 1987). Furthermore, the 4qA allele associated with the disease has a region of β-satellite DNA distal to the array (van Geel *et al.*, 2002); this repeat is also heterochromatic in nature (Meneveri *et al.*, 1985; Agresti *et al.*, 1987). D4Z4 has a high frequency of DNA methylation at CpG islands (Tsien *et al.*, 2001), which is linked to chromatin compaction (Nguyen *et al.*, 2001).

Not only does the D4Z4 array contain sequences associated with heterochromatin but, unlike other human subtelomeric regions, it has also been shown to localise to the nuclear periphery (Masny *et al.*, 2004; Tam *et al.*, 2004). This area of the nucleus is thought to provide a repressive environment and many inactive genes have been shown to localise here (Cockell and Gasser, 1999).

The predicted heterochromatic nature of this region and the failure to identify any mRNA from the repeat are the main reasons that research on the disease mechanism originally focussed on a model where the D4Z4 deletion has a position effect on neighbouring genes. A number of hypotheses have been suggested to explain the mechanism underlying such a position effect.

1.7.1 Position Effect

Position Effect Variegation (PEV) was initially described in *Drosophila* (Levis *et al.*, 1985) and is most evident in the *white* gene, which is responsible for the eye colour of adult flies. In some cases when a gene is moved into a region of heterochromatin it is silenced in some cells and expressed in others, in the case of the *white* gene, this results in red and white-patched eyes (Lewis, 1950). Telomeric position effect (TPE) has been identified in yeast (Gottschling *et al.*, 1990) and there has also been evidence for a role for position effects in mammalian genetic disease (Reviewed in Kleinjan and van Heyningen, 1998). Proteins with histone

deacetylase activity (HDACs) are recruited to heterochromatic regions of DNA, such as the centromeres and telomeres. Histone deacetylation is associated with a more condensed, repressive chromatin structure; the epigenetic marks act as binding sites for further HDAC molecules and facilitate the spreading of heterochromatin along the chromosome. It is proposed that this heterochromatinisation spreads into neighbouring genes causing them to be silenced (Figure 1.4a).

It has been suggested that the D4Z4 array acts as an insulator element (Ottaviani *et al.*, 2006); in this role it would protect the genes at 4q35 against the spread of heterochromatin from the 4q telomere. Contractions of the array are proposed to bring the candidate genes closer to the telomere thereby increasing the chance of heterochromatinisation and inactivation of normally expressed genes. Recently, there has been some evidence that D4Z4 behaves as a CTCF-dependent insulator; in a yeast model it was shown to protect a reporter gene from a telomeric position effect (Ottaviani *et al.*, 2006).

1.7.2 An alternative model: Loss of repression in FSHD

The reverse situation to heterochromatinisation has also been considered. Silencing of genes due to juxtaposition of repetitive DNA has been seen in a number of model organisms (Garrick *et al.*, 1998) and so it has been suggested that D4Z4 normally has a heterochromatic silencing effect on neighbouring genes (Gabellini *et al.*, 2002). Gabellini and colleagues postulated that in FSHD shortening of the array reduces this repressive effect, causing inappropriate upregulation of genes in the region.

In support of this model, a repression complex was identified which binds a 27bp sequence found in each repeat of the array (Gabellini *et al.*, 2002). The interaction between D4Z4 and nuclear proteins was analysed in an EMSA assay with probes which spanned the D4Z4 sequence. Three components of the repression complex were isolated; HMG2B, which belongs to a family of non-histone nuclear proteins that modulate chromatin architecture (Thomas and Travers, 2001); YY1, which interacts with co-repressors such as histone

deacetylases (Thomas and Seto, 1999) and nucleolin, a component of transcription factor complexes (Hanakahi *et al.*, 1997).

This complex has the potential to recruit proteins that can establish a repressive chromatin structure in the region. The D4Z4 array is methylated in unaffected individuals and so it is possible that a DNA methyltransferase is part of, or is recruited by, the complex. In FSHD, contraction of the array could reduce the number of bound proteins resulting in a smaller repressive effect and increased expression of the genes nearby (Figure 1.4b) (Gabellini *et al.*, 2002).

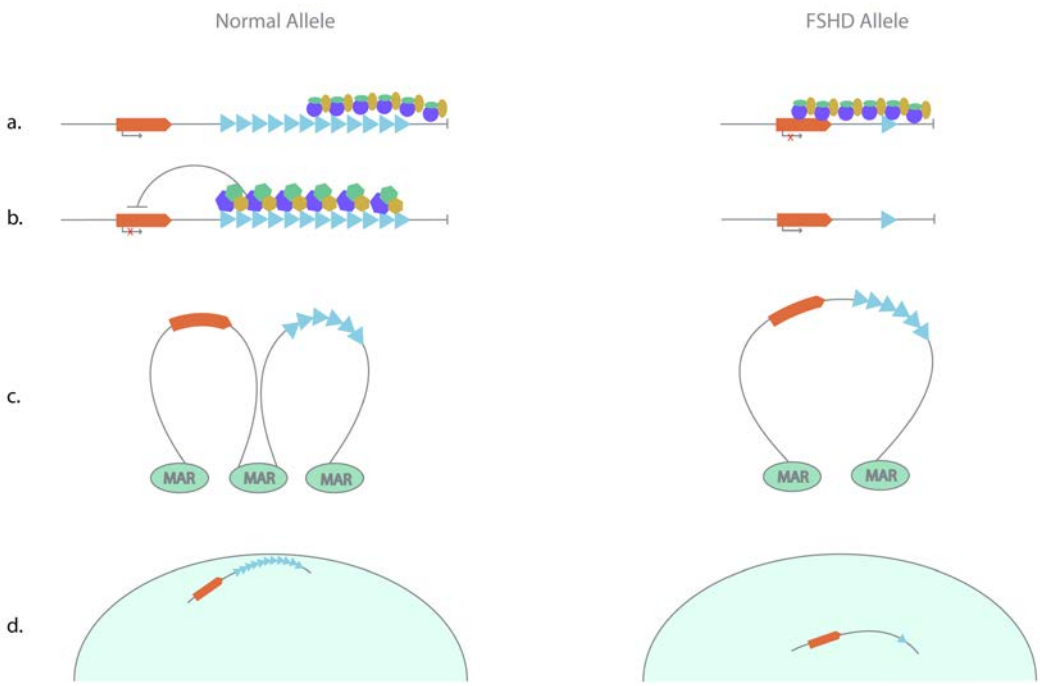


Figure 1.4 Schematic representation of position effect hypotheses for FSHD.

a) Position effect variegation (PEV). There is spreading of complexes containing histone deacetylases from the telomere. In healthy individuals D4Z4 prevents this heterochromatinisation spreading into neighbouring genes. It is proposed that in FSHD, contraction of the array allows the heterochromatin to spread into neighbouring genes and silence them. b) Loss-of-PEV. A repression complex binds to D4Z4 repeats and silences the neighbouring gene. It is hypothesized that when the repeats are lost there is reduced repression complex binding that results in inappropriate expression of the gene. c) Chromosomal looping. In healthy individuals D4Z4 and the candidate gene reside in different chromosomal loops that are separated by an FRR-MAR. In FSHD the MAR is lost and the candidate genes reside in the same loop as the array, in this situation D4Z4 may be able to influence the expression of the genes. d) Nuclear localisation. Contraction of the array brings the candidate genes into a more active environment, leading to their inappropriate expression.

In 2003, van Overveld and colleagues showed that the D4Z4 array is hypomethylated on FSHD alleles (van Overveld *et al.*, 2003; van Overveld *et al.*, 2005), which seemed to support the loss of methylating activity on contraction of the array. Recently, however, de Greef *et al.* have shown that D4Z4 is also hypomethylated in FSHD2 patients; in these cases DNA methylation is lost from both 4q alleles as well as the homologous arrays on chromosome 10 (de Greef *et al.*, 2009) and suggests that a change in chromatin structure in the region unifies the two forms of the disease (de Greef *et al.*, 2010). It should be noted that patients with ICF (immunodeficiency, centromeric instability and facial abnormalities) syndrome exhibit complete hypomethylation of D4Z4 without FSHD symptoms (Kondo *et al.*, 2000). Loss of methylation of D4Z4 in FSHD is therefore likely to be a downstream consequence of other chromatin changes, rather than a causative event.

1.7.3 Chromosomal Looping

One of the highest orders of chromatin organisation is the looping of DNA into transcriptional units. These loops are anchored to a protein scaffold via Scaffold/Matrix Attachment Regions (S/MARs) (Boulikas, 1993, 1995; Davie, 1995; van Driel *et al.*, 1995). Genes that reside in the same loop are subjected to the same environment and this organisation allows the region to be regulated independently of those in other domains.

There has been some support for chromosome looping having a role in the FSHD pathology (Petrov *et al.*, 2006). In normal myoblasts the region encompassing D4Z4 (and the candidate genes *FRG1* and *FRG2*), contains a potential S/MAR called FRR-MAR (FSHD-related region S/MAR) (Petrov *et al.*, 2006). The FRR-MAR is reported to be weaker in FSHD myoblasts (Petrov *et al.*, 2006), which could result in D4Z4 being more able to influence the expression of these candidate genes (Figure 1.4c). In 3C experiments, D4Z4 and the candidate gene *DUX4C* were reported to physically interact with the promoter of the *FRG1* gene (Pirozhkova *et al.*, 2008; Bodega *et al.*, 2009). A small reduction in loop formation between D4Z4 and *FRG1* was seen in FSHD myoblasts compared with controls, which is probably attributable to

the fewer repeats on the contracted allele (Bodega *et al.*, 2009). However, this group did not detect *FRG1* misregulation in these cells.

It has also been suggested that looping of the chromatin may allow D4Z4 to have more long-range effects on distant genes (Jiang *et al.*, 2003). There is evidence for long-range looping between locus control regions and the distant genes that they regulate (Carter *et al.*, 2002). If more support for this hypothesis is found the search for candidate genes could be extended outside of the 4q35 region.

1.7.4 Nuclear localisation

There is evidence that the position of a locus within the nucleus is correlated with the expression of its genes. Subnuclear positioning has been shown to regulate transcription of the IgH and Igkappa loci during lymphocyte development (Kosak *et al.*, 2002), while in human fibroblasts, inactive loci (albumin, cMHC and neurotensin) are located at the periphery and active loci (actin and collagen) reside in speckles rich in RNA metabolic factors (Xing *et al.*, 1995). As the nuclear periphery is associated with heterochromatic regions of the genome (Marshall *et al.*, 1997), the observation that 4q35 resides here prompted a nuclear localisation hypothesis for FSHD (Masny *et al.*, 2004). In this model, disruption of the localisation pattern due to contraction of the array might bring the candidate genes to a more active environment, leading to their inappropriate expression (Figure 1.4d).

Localisation of the 4q telomere has been compared between affected and unaffected individuals using 3D Immuno-FISH, in both lymphoblasts and myoblasts. No significant change in localisation linked to FSHD was observed for either cell type (Masny *et al.*, 2004; Tam *et al.*, 2004).

1.8 Is FSHD due to a position effect?

1.8.1 Epigenetic status

In 2003, Jiang *et al.* compared the acetylation of histone H4 at 4q35 between normal and FSHD chromosomes using chromatin immunoprecipitation (ChIP) analysis of somatic cell hybrids containing a single human chromosome 4 (Jiang *et al.*, 2003). The levels of histone acetylation observed were typical of unexpressed euchromatin rather than constitutive heterochromatin, and they found no differences in acetylation levels between FSHD alleles and controls. However, while the use of somatic cell hybrids allowed the study of D4Z4 without interference from homologous sequences, it is possible that these human chromosomes behaved differently in the rodent environment. Recently, ChIP analysis of D4Z4 has identified non-overlapping regions of euchromatin (with H3K4me2 and H3 acetylation) and heterochromatin (with H3K9me3, HPγ1 and H3K27me3) within the D4Z4 array. In FSHD cells, a specific loss of the heterochromatin markers H3K9me3 and Hpγ1 was seen on both 4q and 10q alleles, with no compensatory increase in euchromatic modifications (Zeng *et al.*, 2009).

These studies suggest that transcriptional repression is not spread uniformly along the D4Z4 array, rather that there is a more euchromatic arrangement at the proximal region (Zeng *et al.*, 2009). Loss of heterochromatin markers on FSHD alleles provides evidence against the spreading of heterochromatin in the region. This loss of heterochromatin markers is consistent with the model proposed by Gabellini *et al.* (2002). However, as the loss is not restricted to the contracted allele, it is unclear what role the other 4q35 and 10q26 arrays may play.

1.8.2 Candidate genes

The region of chromosome 4q proximal to D4Z4 is gene-poor, however some candidate genes have been identified (see Figure 1.5).

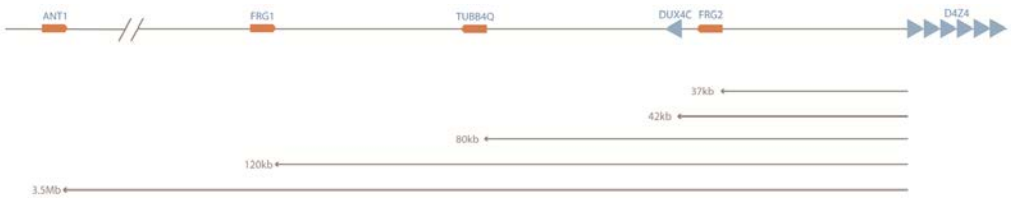


Figure 1.5 Candidate genes for FSHD at 4q35.

The region proximal to the array is gene-poor, however some candidate genes have been identified and those most studied are shown schematically here. Adenine Nucleotide Translocator 1 (*ANT1*), FSHD-Region Gene 1 (*FRG1*), Tubulin, Beta Polypeptide 4, member Q (*TUBB4Q*) and FSHD-Region Gene 2 (*FRG2*) are represented in orange and their distance from D4Z4 shown underneath. The arrow indicates the transcriptional orientation. Each arrowhead represents an individual D4Z4 repeat. DUX4C is a single copy of a D4Z4 repeat in the reverse orientation.

1.8.2.ii *FRG1*

The first candidate gene identified at 4q35 was *FRG1* (FSHD region gene 1), which is located 100kb proximal to D4Z4 (van Deutekom *et al.*, 1996b) and is expressed in most human tissues (Bodega *et al.*, 2009). In 2002, Gabellini *et al.* showed increased expression of the gene in patient muscle biopsies compared with controls. Some later studies have also reported an increase in *FRG1* expression (Alexiadis *et al.*, 2007; Bodega *et al.*, 2009). However expression levels in these two later studies were not increased by more than 1.5 fold compared with the 30-fold increase reported by Gabellini *et al*. Many repeat studies have been unable to confirm a difference in expression levels for *FRG1* (Van Deutekom *et al.*, 1993a; Dixit *et al.*, 2007; Osborne *et al.*, 2007; Masny *et al.*, 2010) and some have even found a decrease in expression levels in patients compared to controls. (Winokur *et al.*, 1993; Jiang *et al.*, 2003)

FRG1 protein has been identified in proteomic analyses of the human spliceosome (Jurica *et al.*, 2002; Rappsilber *et al.*, 2002) and stable expression of *FRG1* in U2OS cells and in a transgenic mouse model has been shown to induce altered splicing of the *TNNT3* gene (Gabellini *et al.*, 2006; van Koningsbruggen *et al.*, 2007). Gabellini *et al.* also identified altered splicing of the *TNNT3* gene in muscle cell cultures from FSHD patients (Gabellini *et al.*, 2006). However, Osbourne *et al*. found no difference in the *TNNT3* splicing pattern between FSHD patients and controls (Osborne *et al.*, 2007). Therefore, a pathogenic role of differential splicing in FSHD remains controversial.

Several groups have overexpressed FRG1 protein, although at much higher levels than ever reported for FSHD patients. Experiments in *X. laevis* (Hanel *et al.*, 2009) and *M. musculus* (Gabellini *et al.*, 2006) provide evidence consistent with the hypothesis that *FRG1* has a role in FSHD. However, neither of these models gives an FSHD specific phenotype. Overexpression of *FRG1* in the skeletal muscle of transgenic mice results in a muscular dystrophy phenotype (Gabellini *et al.*, 2006). The severity of the phenotype in these mice is correlated with the levels of overexpression, with the most affected mice expressing high levels of human FRG1 relative to the endogenous protein (Gabellini *et al.*, 2006). As this transgene is under the

control of the human skeletal α-actin gene promoter, overexpression occurs specifically in the skeletal muscle and so it is possible that the muscle phenotype is an artefact of increased protein expression in this tissue. To generate even a mild phenotype in these mice, an increase in FRG1 10 fold higher than endogenous levels was required (Gabellini *et al.*, 2006).

When FRG1 protein levels were reduced by morpholino injection in *X. laevis* a disruption of myotome organisation and growth was observed (Hanel *et al.*, 2009). Abnormal muscle formation was also seen when levels of *FRG1* were elevated (Hanel *et al.*, 2009), and the same group have also observed an effect on angiogenesis when levels of *FRG1* expression are changed (Wuebbles *et al.*, 2009). As these studies have relied on significant over/under expression of the gene, it is possible that the phenotypes are an artefact of the experimental conditions.

1.8.2.iii *FRG2*

Of the genes studied in detail, the putative gene *FRG2* lies closest to D4Z4, located 35kb proximal to the array (Gabellini *et al.*, 2002); it has the potential to encode a nuclear protein of unknown function (Rijkers *et al.*, 2004). Closely related sequences have also been identified on a number of different chromosomes (1, 3, 4, 7, 8, 10, 16, 18, 20, 22 via BLAST searches), with FISH analysis showing the strongest hybridisation signal to chromosomes 4 and 10 (Rijkers *et al.*, 2004). One study showed that *FRG2* mRNA could be amplified by RT-PCR from differentiating myoblast cultures of FSHD cells but not cells from healthy controls. Transcripts were absent in all other tissues tested (Rijkers *et al.*, 2004). The PCR products were sequenced to determine their chromosomal origin and came predominantly from chromosome 10, although some originated from chromosome 4 (Rijkers *et al.*, 2004). Rijkers *et al.* suggest that pairing of the subtelomeric regions of 4q and 10q could explain how transcriptional activation of the 10q array occurs on contraction of D4Z4 on chromosome 4. *FRG2* transcripts have also been identified by RT-PCR in myoblasts derived from non-FSHD myopathies; however, these were derived from chromosomes 3 and 22 (Rijkers *et al.*, 2004). As the FRG2 primers used in this study amplify transcripts from multiple chromosomes, it will

be important to sequence all RT-PCR products from expression studies in order to confirm the origin of transcripts. Overexpression of *FRG2* in mouse skeletal muscle shows no phenotype (Gabellini *et al.*, 2006) and in FSHD patients with extended deletions the region containing *FRG2* is lost (Lemmers *et al.*, 2003), arguing against a primary role for this gene in the disease.

1.8.2.iv *ANT1*

ANT1 encodes an adenine nucleotide translocator, which facilitates transport of ATP over the mitochrondrial membrane (Li *et al.*, 1989; Stepien *et al.*, 1992; Doerner *et al.*, 1997; Lemmers *et al.*, 2003). Although it resides over 3.5Mb from the array (Wijmenga *et al.*, 1993), it has been studied as a functional candidate because of its high expression in heart and skeletal muscle. *ANT1* expression is increased on depletion of YY1 or MeCP2 (Forlani *et al.*, 2010). Since YY1 has been identified as part of the D4Z4 binding complex (Gabellini *et al.*, 2002), this could provide support for the hypothesis that contraction of the array reduces binding of the repression complex, thereby increasing expression of the nearby genes. However, this effect was also seen with the murine *Ant1* gene, despite it having no equivalent D4Z4 array nearby (Forlani *et al.*, 2010). There is some western blot support for changes in the levels of expression of the *ANT1* protein in FSHD patients (Laoudj-Chenivesse *et al.*, 2005; Macaione *et al.*, 2007). However, the association of *ANT1* mutations with other disorders (Hirano and DiMauro, 2001; Jordens *et al.*, 2002; Dˆrner and Schultheiss, 2007; Forlani *et al.*, 2010) suggests that the protein may have a role in secondary pathological processes, rather than the primary cause, especially since overexpression of ANT1 in the mouse has no phenotype (Gabellini *et al.*, 2006).

1.8.2.v *DUX4C*

DUX4C is inside an inverted, partial copy of a D4Z4 repeat that maps 42kb proximal to the main array (Hewitt *et al.*, 1994). The truncated repeat contains an ORF that is 150bp shorter than DUX4 and varies from the DUX4 sequence in the C-terminal domain at 33 of the final 141

positions. Ansseau *et al.* used a rabbit antiserum against a 16-residue peptide from the C-terminal domain to identify a protein of the predicted size of DUX4C in differentiating muscle cells from FSHD patients and controls using western blot. This study reported a 1.5 fold increase in DUX4C expression in FSHD samples compared with control or Duchenne muscular dystrophy cells (Ansseau *et al.*, 2009). They were also able to identify transcripts by RT-PCR in both FSHD and control cells, at proliferation and differentiation stages and sequencing showed that the RT-PCR products matched the DUX4c sequence (Ansseau *et al.*, 2009) (for primer positions see Figure 4.3). However, no other expression studies have identified transcripts from this locus and, like FRG2, *DUX4C* is lost in some patients with extended deletions (Lemmers *et al.*, 2003) so a primary involvement in the disease seems unlikely.

1.8.2.vi Expression of 4q35 candidate genes

Thus, several studies have compared expression of these candidate genes in FSHD and control cells (for a summary see Table 1.1). However, they have produced conflicting data that does not give a clear, consistant picture of the disease. In 2002 Gabellini *et al*. used end-point RT-PCR to examine expression of three genes (*FRG1*, *FRG2* and *ANT1*) in FSHD muscle samples and showed increased expression of all three loci (Gabellini *et al.*, 2002). For *FRG2*, expression was seen only in the three patient samples and not in three controls. The increase in expression correlated inversely with distance from the array, and Gabellini *et al.* suggested that D4Z4 has more influence on genes that are closer to it. While this data seemed to provide support for a position effect and loss of silencing of these genes in FSHD, a number of groups have repeated expression analyses for the genes and the results have generally not been in agreement with Gabellini *et al*. (Chen *et al.*, 2000; Jiang *et al.*, 2003; van Overveld *et al.*, 2003; Winokur *et al.*, 2003b; Macaione *et al.*, 2007; Masny *et al.*, 2010). The majority of studies that have used quantitative measures such qRT-PCR or microarrays have found no difference in expression levels of the candidate genes between FSHD samples and controls (van Deutekom *et al.*, 1996b; Jiang *et al.*, 2003; van Overveld *et al.*, 2003; Winokur *et al.*,

2003b; Dixit *et al.*, 2007). Although the lack of validation could be due to the different methods used, different sites of muscle biopsy or state of cell cultures and/or differences in the severity of disease in the patients tested, there is no clear consensus of support for the derepression model of Gabellini *et al.* (Table 1.1).

As the sample sizes in these studies are generally small, any differences seen may be in part due to normal variation in gene expression levels. As *FRG1* is a multicopy gene it is also possible that transcripts from chromosomes other than 4q were amplified by RT-PCR. In order to confirm that any variation in expression in this gene is due to changes at 4q35 it is important to determine the chromosomal origin of such transcripts.

In 2010, Masny *et al.* examined transcription of nascent RNA in single myonuclei using quantitative FISH. This method allowed them to distinguish transcription from the contracted and the noncontracted alleles, as well as the chromosome 4 specific expression of multicopy genes. None of the 16 genes tested showed a difference in expression from the FSHD alleles (Masny *et al.*, 2010). Current data therefore suggests that contraction of D4Z4 does not have a significant effect on the expression of the candidate genes outside the array.

While support for a position effect on neighbouring genes is diminishing, evidence for the direct involvement of D4Z4 in FSHD is increasing and it is now appropriate to reconsider the role that D4Z4 may play in the pathology of FSHD.

Reference	Method	Sample	Patients/Controls	*FRG2*	*FRG1*	*ANT1*	*DUX4C*
van Deutekom *et al.*(1996)	Allele specific RT-PCR	Muscle biopsy from gastrocnemius or quadriceps	4/1	n/a	no change	n/a	n/a
Gabellini *et al.* (2002)	RT-PCR	Muscle biopsy	3/3	▲	▲	▲	n/a
Jiang *et al.* (2003)	qRT-PCR	Muscle biopsy from deltoid or biceps	7/7	n/a	▼	no change	n/a
Winokur *et al.* (2003)	Microarray	Muscle biopsy from biceps and deltoid	12/9	n/a	▼	no change	n/a
Rijkers *et al.* (2004)	qRT-PCR	Differentiating myoblast cultures	3/3	▲	n/a	n/a	n/a
Laoudj-Chenivesse *et al.* (2005)	Western Blot	Muscle biopsy		n/a	n/a	▲	n/a
Osbourne *et al.* (2007)	Microarray and qRT-PCR	Muscle biopsy	19/30	no change	no change	no change	no change
Alexiadis *et al.* (2007)	RNAPol-ChIP PCR and RT-PCR	Myoblast cultures established from deltoid muscle	2/1	n/a	▲	n/a	no change
Dixit *et al.* (2007)	Microarray	Muscle biopsy	9/6	n/a	no change	no change	n/a
Bodega *et al.* (2009)	RT-PCR and Western Blot	Differentiating myoblast cultures	4/3	n/a	▲	n/a	n/a
Bodega *et al.* (2009)	qRT-PCR	Muscle biopsy	3/3	n/a	▲ *	n/a	n/a
Ansseau *et al.* (2009)	Western Blot	Muscle biopsy from quadriceps or deltoid and myoblast cultures	10/4	n/a	n/a	n/a	▲
Masny *et al.* (2010)	RNA DNA-FISH	Myotube cultures	2/2	n/a	no change	no change	n/a

Table 1.1 Summary of expression analyses for 4q35 candidate gene

A summary of published expression analyses for the main candidate genes at 4q35. Up regulation of expression in FSHD samples compared to controls is indicated by a green arrow, down regulation is indicated by an orange arrow. 'no change' - no significant difference in expression levels between FSHD samples and control samples. n/a - not assessed.* - Only 1 of the 3 samples showed up-regulation, the other two showed no change.

1.9 D4Z4 may encode the *DUX4* gene

Two cDNAs, DUX1 and DUX2, were identified during a search for target genes of the helicase-like transcription factor (HLTF). Using ChIP, Ding *et al.* showed that HLTF bound to a DNA fragment (HEFT1) with 82-86% nucleotide identity to D4Z4. HEFT1 does not map to the FSHD locus but presumably to a dispersed 3.3kb repeat (Ding *et al.*, 1998). This fragment was shown to have promoter activity when fused to a luciferase reporter gene and introduced into HeLa cells and the rhabdomyosarcoma cell line TE671. A probe from HEFT1 was then used to isolate the DUX1 and DUX2 cDNAs by library hybridisation.

Several cDNAs homologous to D4Z4 had been identified previously (Hewitt *et al.*, 1994; Lyle *et al.*, 1995). However, as these transcripts did not contain ORFs through the homeoboxes they likely represent expressed pseudogenes. The ORF of *DUX2* is truncated immediately downstream of the first homeodomain and, as no *in vivo* expression has yet been identified, it is likely that this is also a pseudogene. RT-PCR products have been obtained for *DUX1*, and two further putative DUX genes, *DUX3* and DUX5 (Ding *et al.*, 1998; Beckers *et al.*, 2001). A protein of the apparent molecular weight of DUX1 was identified in TE671 cells by western blotting with antiserum raised against a peptide from the predicted protein (Ding *et al.*, 1998).

The promoter-like region in D4Z4 maps just upstream of the homeobox ORF. This putative gene within D4Z4 was named *DUX4* (Gabriels *et al.*, 1999). Gabriels *et al*. were able to demonstrate promoter activity of this HEFT1 related sequence when inserted upstream of a luciferase reporter gene and introduced into TE671 cells.

Dixit *et al.* extracted RNA from C2C12 cells transfected with a plasmid containing a patient-derived D4Z4 array (two repeats) and used 3'-RACE to characterise the *DUX4* transcripts. They identified two different sized fragments, which they suggest is due to alternative splicing of a 136bp intron located after the stop codon (Dixit *et al.*, 2007; see Figure 1.6). This is an unusual organisation as introns in the 3'UTR are rarely found as nonsense-mediated decay (NMD) degrades transcripts containing premature stop codons (Scofield *et al.*, 2007).

In 2007, two papers reported evidence of *DUX4* expression in FSHD myoblast cell lines (Dixit *et al.*, 2007; Kowaljow *et al.*, 2007). Using RT-PCR, Kowaljow *et al.* were able to amplify a 1477bp fragment from both proliferating and differentiating FSHD myoblasts, while Dixit *et al.* were able to amplify a 1700bp fragment only from differentiating FSHD cells (Figure 1.6). However, neither of these groups published sequence data of the PCR products and since 4q and 10q arrays are very similar, this data is necessary to confirm the origin of the transcripts.

By Western blot of FSHD muscle proteins, Dixit and colleagues also detected a protein of the predicted mass of DUX4 using a mouse monoclonal antibody raised against the final 250 residues of the DUX4 ORF (Dixit *et al.*, 2007). Using polyclonal antibodies raised against a synthetic peptide corresponding to amino acids 342-356, Kowaljow *et al.* were unable to identify the DUX4 protein in FSHD or control myoblasts, however, they identified expression of a DUX4-related protein in two rhabdomyosarcoma cell lines (RD and RMS13) (Kowaljow *et al.*, 2007). None of these putative DUX4 proteins have been analysed using proteomic approaches, so verification that they represent a *bona fide* DUX4 protein will be important.

Transcription, if confirmed, would support the view that DUX4 has a coding function. Although the organisation of D4Z4 is unusual for a gene, there are examples of tandemly-arrayed protein-coding genes in humans. For example, RS447 megasatellites on chromosomes 4p and 8p encode a functional deubiquitinating enzyme, *USP17*, (Saitoh *et al.*, 2000) and the *TSPY* gene resides within the repeats of the Y-chromosome specific DYZ5 array (Manz *et al.*, 1993).

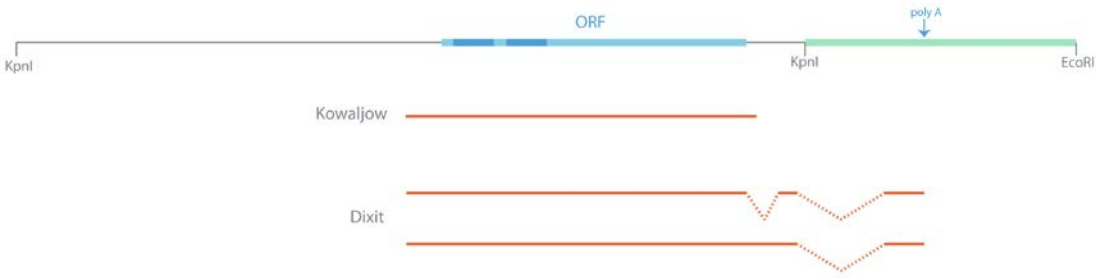


Figure 1.6 DUX4 transcripts identified in 2007

A schematic diagram of the final repeat of D4Z4. The ORF is shown in pale blue with the homeodomains highlighted in dark blue. The pLAM sequence adjacent to the array is shown in green. The potential polyA signal is indicated with an arrow. The RT-PCR products from Kowaljow *et al.* (2007) and Dixit *et al.* (2007) are shown underneath in orange. Predicited introns are identified by dotted lines.

1.10 Evolutionary analysis indicates a protein coding function for D4Z4

1.10.1 D4Z4 homologs

D4Z4 homologs in great apes and Old (rhesus) and New World (marmoset) monkeys, were originally identified by physical mapping studies in 1996 (Clark *et al.*, 1996; Winokur *et al.*, 1996). More recently, the advent of whole genome sequencing has led to the identification of homologues in a number of other species, including the murine and Afrotherian lineages (Clapp *et al.*, 2007).

In the great apes, D4Z4 orthologues align along the whole of the 3.3kb repeat. In addition, as in humans these species have numerous dispersed copies distributed throughout the genome (Clark *et al.*, 1996; Winokur *et al.*, 1996; Clapp *et al.*, 2007). In other species, including Old and New World monkeys, there is little nucleotide identity outside of the ORF, however the *DUX4* ORF itself has been maintained (Figure 1.7). This conservation since the divergence of Afrotherian and Eutherian mammals (>100mya) strongly suggests a protein coding function for the gene. Statistical comparison of the primate nucleotide sequences suggests that the ORF has been conserved by selection (Clapp *et al.*, 2007).

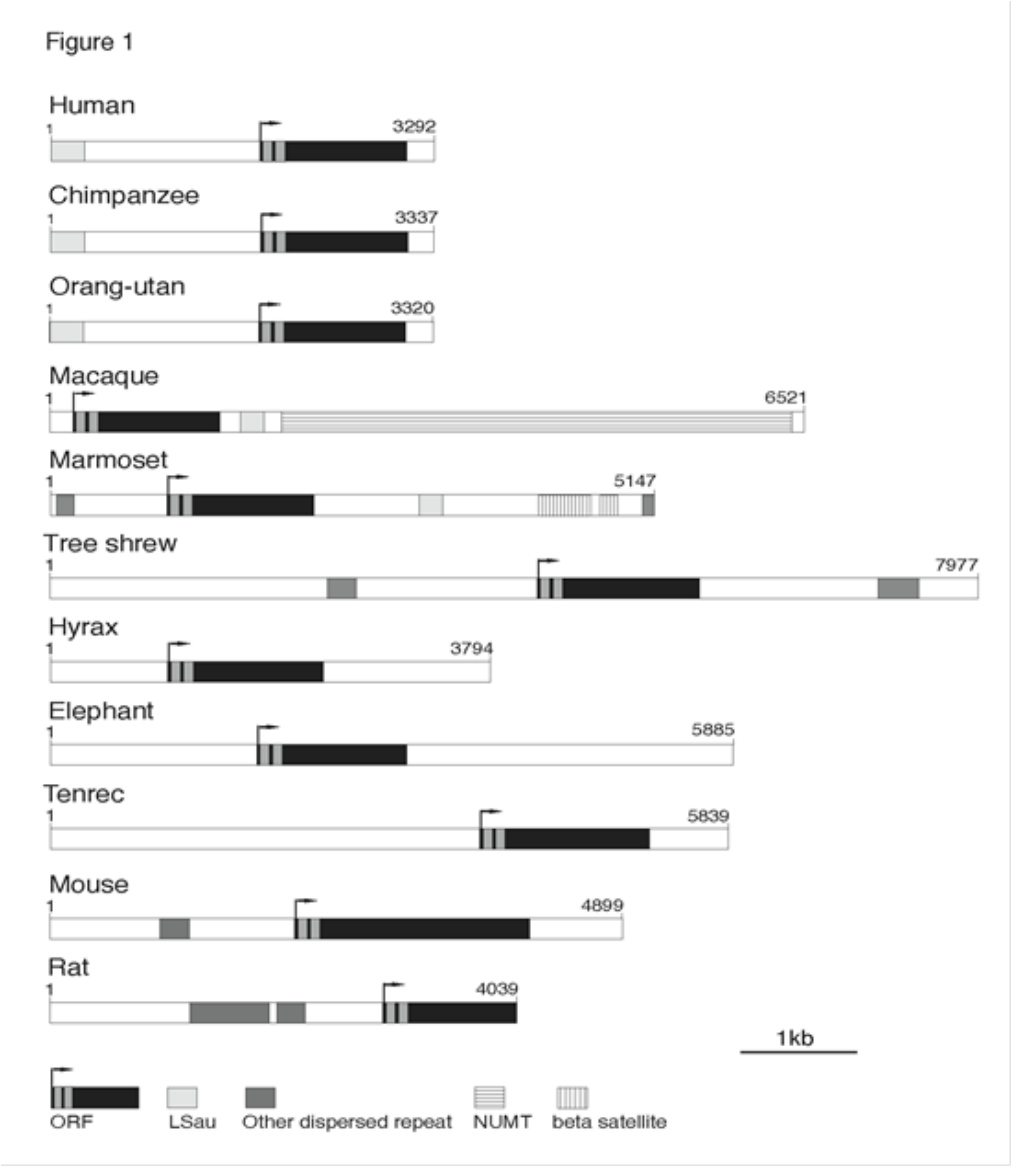


Figure 1.7 Schematic diagram of mammalian D4Z4-related repeats.

Repeats and sequence elements are drawn to scale. The ORF is conserved among all the species identified. Taken from Clapp et al. 2007

1.10.2 Intron-containing DUX genes

In 2007, Booth and Holland identified intron containing DUX genes (*DUXA* and *DUXB)* in humans (Booth and Holland, 2007). Ten retrotransposed pseudogenes of *DUXA* were also identified, indicating expression of the gene in the germline or early embryo. Orthologues of *DUXA* and *DUXB* have since been identified in a wide range of mammalian species (Clapp *et al.*, 2007; Leidenroth and Hewitt, 2010) and two new classes of intron containing DUX genes, *DUXC* and *Duxbl*, have been discovered (Clapp *et al.*, 2007).

In addition to searching publically available genome assemblies, Leidenroth and Hewitt used synteny information to catalogue both intact and inactive DUX homologues across placental mammals (Figure 1.8). This approach also allowed them to identify a homologue containing a single-homeobox gene in the non-mammalian species opossum and chicken, which they named *sDux*.

The similarity between the homeodomains of the four intron-containing DUX genes, as well as their similar gene structures and splice locations (Figure 1.9), suggest that duplication of the DUX homeobox occurred only once. As no double-homeoboxes have been identified in non-mammalian species, the phylogenetic data supports a model where duplication of the sDUX homeobox occurred in the ancestor of all placental mammals, giving rise to the ancestral DUX gene from which the others are descended. In a maximum likelihood tree of the DUX homeodomains, the first homeodomain of *DUXB*, *DUXC* and *Duxbl* cluster together, as do the second, providing further support that the two homeodomains were already duplicated and divergent in their common ancestor (Leidenroth and Hewitt, 2010).

Leidenroth and Hewitt identified a pattern of reciprocal loss and retention of DUXB and Duxbl (Figure 1.8), suggesting these genes are paralogues and more closely related to each other than to either DUXA or DUXC. The initial functional redundancy present after duplication allowed one of these paralogues to decay in each mammalian lineage. The fact that DUXA is maintained in addition to DUXB/Duxbl suggests that the genes have diversified to fill different roles.

DUXC is the only intron-containing DUX gene that is found in a tandem array organisation (Leidenroth and Hewitt, 2010) and it clusters with *DUX4* in phylogenetic analyses (Clapp *et al.*, 2007; Leidenroth and Hewitt, 2010). In addition, *DUXC* shares a conserved C-terminal domain with the *DUX4* homologs and the mouse *Dux* genes suggesting that *DUX4* probably arose from a retrotransposed copy of *DUXC*, with both genes present in the common ancestor of all mammalian orders 105 million years ago. The D4Z4 array then arose from amplification of *DUX4* in this ancestor. Current data indicates that *DUX4* has since been lost from Laurasiatheria and *DUXC* lost in the Afrotherian and Primate lineages (Clapp *et al.*, 2007; Leidenroth and Hewitt, 2010).

The *Dux* array identified in the rodent is likely to have arisen from an independent amplification of a retrotransposed copy of *DUXC*. Although these sequences are unlikely to be true orthologs of the *DUX4* arrays in humans, they share both the C-terminal domain and their tandem array organisation, and thus may be functionally equivalent.

1.11 Dux genes are expressed in mice

Gene expression studies in the mouse have been less complicated than in human due to a lack of dispersed repeats. In 2007, Clapp *et al* published evidence for expression of at least some copies of the mouse Dux gene. Overlapping primers were used to show that the whole ORF is transcribed and this was confirmed using RNA-FISH and *in situ* hybridisation.

Clapp and colleagues were also able to amplify both sense and antisense transcripts from the Dux array by RT-PCR, a transcription pattern that is also found in the R5447 tandem repeat array encoding the deubiquitinating enzyme USP17 (Saito *et al.*, 2000). Antisense transcription of Dux was confirmed by RNA-FISH (Clapp *et al.*, 2007). Antisense transcription is thought to provide a regulatory mechanism for gene expression (Lapidot and Pilpel, 2006). It has been shown that endogenous antisense sequences can down-regulate the expression of sense transcripts (Lapidot and Pilpel, 2006) and introduction of artificial complementary sequences is now used to knockdown gene expression in model systems (Sugimoto, 2004; Spaenkuch and Strebhardt, 2005; Zhou *et al.*, 2006). It is possible that these antisense

transcripts provide regulatory control for the expression of the mouse Dux array, although this needs to be confirmed experimentally.

In contrast to the apparently limited expression of *DUX4* shown in human cell lines, expression in the mouse was identified in a number of different tissues, particularly the CNS (Clapp *et al.*, 2007). Since the expression patterns seem to differ between species it will be important to establish whether human *DUX4* and the mouse *Dux* genes have equivalent functions.

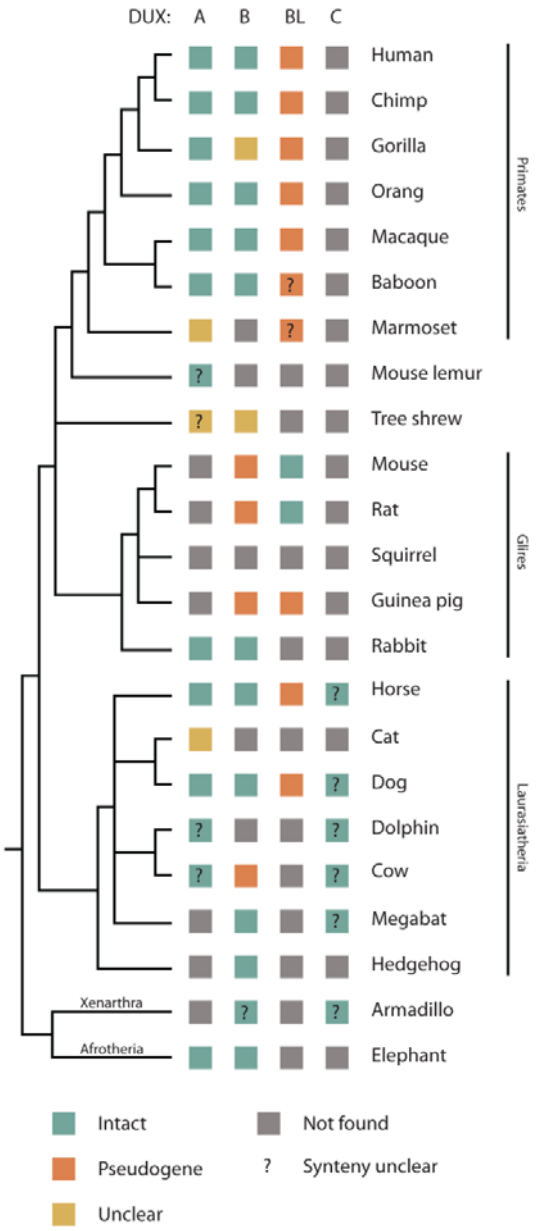


Figure 1.8 Mammalian DUX distribution

Summary of the mammalian DUX catalogue produced by Leidenroth and Hewitt (2011). Presumed functional genes (with intact ORF across all four homeobox exons, first exon may be unidentified) in green. *DUX* sequences with stop codons or deleted/missing exons in well sequenced regions in orange. Putative *DUX* homologues with unclear functional status (mainly due to gaps in assembly) in yellow. Unless marked with ?, synteny was confirmed with anchor genes.

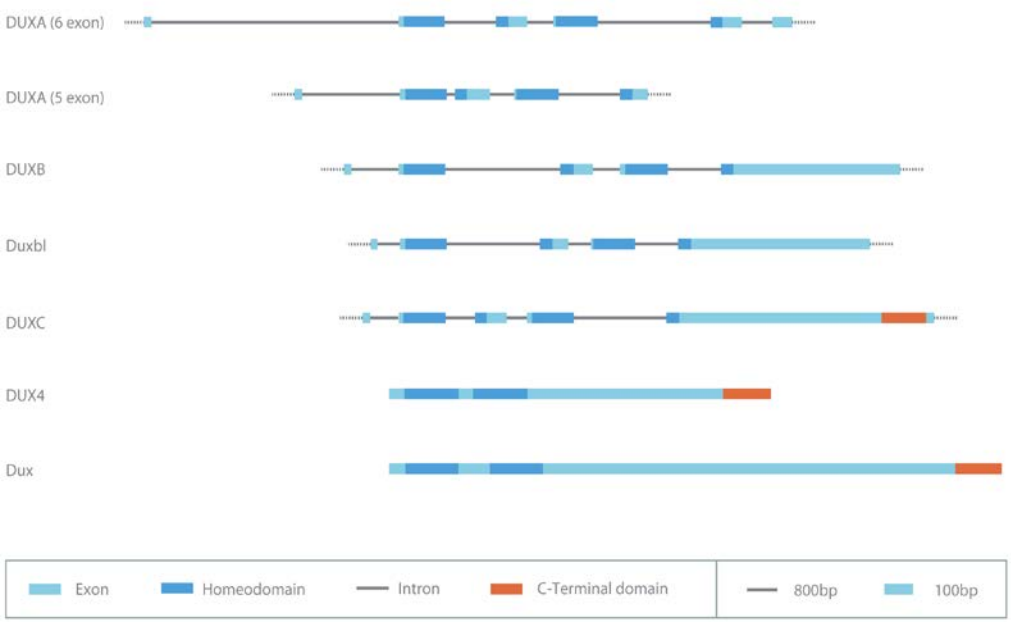


Figure 1.9 Schematic diagram of the DUX family

Structures of the DUX family members. The ORF is shown in light blue and homeboxes highlighted in dark blue. The conserved C-terminal domain is shown in orange. For the intron-containing DUX genes the transcription start and end sites are uncertain and indicated by dotted lines. Introns and exons are drawn to different scales. Adapted from Leidenroth and Hewitt 2011.

1.12 What is the likely function of DUX4 *in vivo*?

The double homeobox within *DUX4* suggests the protein is likely to be a transcription factor, it would therefore be expected to localise to the nucleus and a number of groups have shown this to be the case for exogenously expressed protein (Ostlund *et al.*, 2005; Kowaljow *et al.*, 2007). The C-terminal fragment of *DUX4* has also been shown *in vivo* to enhance the transcriptional activity of *CIC*, a HMG box transcription factor. Fusions between *CIC* and *DUX4* have been identified in two patients with Ewing-like sarcomas and their ability to activate *CIC* target genes (Kawamura-Saito *et al.*, 2006) provides evidence that *DUX4* can behave as a transcriptional activator.

1.13 What are the transcriptional targets of DUX4?

A number of global expression studies have looked for genes that are mis-regulated in FSHD muscles (Tupler *et al.*, 1999; Winokur *et al.*, 2003b; Celegato *et al.*, 2006; Dixit *et al.*, 2007). In order to identify disease-specific changes in expression, the results of such experiments (e.g. microarrays) are compared with other forms of muscular dystrophy. Genes that are consistently mis-regulated in the other disorders are likely to be the result of secondary pathological processes and can be removed from further analysis.

Some researchers have suggested that there is a global misregulation of gene expression, rather than the effect of the deletion being restricted to chromosome 4q35 genes (Winokur *et al.*, 2003b). The set of genes with apparently altered expression levels in FSHD are enriched in muscle-specific transcripts (Tupler *et al.*, 1999; Winokur *et al.*, 2003b; Celegato *et al.*, 2006), and in particular those controlled by the muscle-specific transcription factor MyoD (Winokur *et al.*, 2003b; Celegato *et al.*, 2006). The MyoD pathway is necessary for myogenic differentiation and is involved in the regeneration of muscle in response to stress and injury. This pathway has previously been shown to play a role in the nuclear envelope dystrophies, which are thought to be related to FSHD (Chen *et al.*, 2000).

It is known that cascades of transcription factors are responsible for the control of myogenesis and it is possible that *DUX4* has a role in this process. If DUX4 is a strong transcriptional activator *in vivo*, low levels of its expression could potentially affect the transcription of a large number of genes. Transfection of mouse C2C12 cells with human *DUX4* inhibited differentiation (Snider *et al.*, 2009) and overexpression of *DUX4* in zebrafish (Snider *et al.*, 2009) or *Xenopus* (Grewal *et al.*, 1999) embryos caused developmental arrest and a reduction in muscle markers. However, as observed by Leidenroth and Hewitt, the relevance of overexpressing DUX4 in species (such as zebrafish or Xenopus) which do not contain any double-homeobox genes is unclear. It is possible that the phenotypes seen in these embryos are an artefact of overexpression of a transcription factor which these species have not evolved to cope with (Leidenroth and Hewitt, 2010).

Nuclear factor-kappaB (NF-κB), shown to be activated in Duchenne muscular dystrophy, is also thought to be involved in the regulation of myogenesis and muscle repair (Monici *et al.*, 2003). NF-κB is induced in cells in response to oxidative stress; this activation is thought to be mediated by receptor for advanced glycation end products (RAGE). The RAGE-NF-κB pathway is involved in muscle regeneration in the limb girdle muscular dystrophies (Haslbeck *et al.*, 2004) and is stimulated in FSHD muscles (Macaione *et al.*, 2007). A number of studies have shown that FSHD cells are more susceptible to oxidative stress than controls (Winokur *et al.*, 2003a; Laoudj-Chenivesse *et al.*, 2005; Macaione *et al.*, 2007). However, the fact that these pathways are also activated in other forms of muscular dystrophy (Moylan and Reid, 2007), suggests that they are secondary consequences of muscle disorder rather than a causative mechanism for the disease.

Following microarray experiments it has recently been reported that the paired-like homeodomain transcription factor 1 (*PITX1*) is specifically upregulated in the muscles of patients with FSHD (Dixit *et al.*, 2007). A potential DUX4 binding site was identified in the *PITX1* promoter region and PITX1 expression was seen in some C2C12 cells following transfection with a DUX4 construct (Dixit *et al.*, 2007). These results suggest that DUX4 could be an upstream regulator of *PITX1*. There are three *PITX1* paralogues in humans (*PITX1*, *PITX2* and *PITX3*). All three of the genes are expressed in muscle; PITX3 may contribute to

myogenesis (Coulon *et al.*, 2007), while PITX1 has a role in determining hind limb identity (DeLaurier *et al.*, 2006).

1.14 Analysis of 4q haplotypes implicates DUX4 in FSHD

A more detailed analysis of polymorphisms on 4q and 10q identified 9 different haplotypes based on sequence variations within and flanking the D4Z4 array (Lemmers *et al.*, 2007; Figure 1.10a). The haplotypes are defined by the chromosomal location (4q35 or 10q26), an SSLP proximal to D4Z4, 15 SNPs within the D4F104S1 (p13E-11) sequence, 4 SNPs within the most proximal unit of the array that alter restriction enzyme recognition sites, and the distal A/B/C variation (Figure 1.10b). FSHD sized alleles on only one of the 4q haplotypes (4A161) were found to be associated with FSHD; contractions on 4qB and one 4qA allele (4qA166) were shown to be nonpathogenic (Lemmers *et al.*, 2007). The low rate of interchromosomal rearrangements has allowed these haplotypes to evolve independently and there is strong linkage disequilibrium between sequences immediately flanking the D4Z4 repeat, this means each subtelomere end is unique and can explain why FSHD is associated with only one haplotype.

Since FSHD is associated with only one of the identified haplotypes, there are likely to be some functional elements outside of and/or within the array that are necessary for the disease. For example, the 4qA161 haplotype may contain a harmful variant that affects regulation of *DUX4* or leads to altered expression levels/ratios of functional repeat variants. With respect to gene regulation the variant might increase the chance of expression, perhaps by creating a binding site for activating proteins, or it may disrupt other repression mechanisms in the region. Alternatively this haplotype may be missing an element important for repression of the gene. If antisense transcription does prove to be an important regulatory mechanism for this gene it may also be affected by variants within or outside the array.

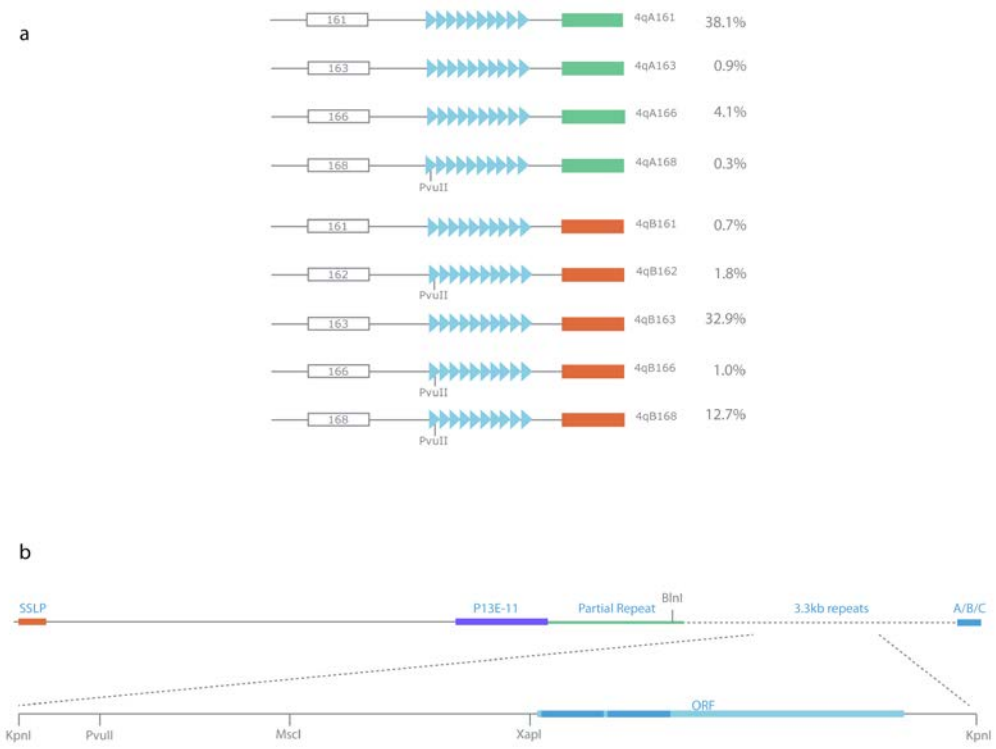


Figure 1.10 A schematic representation of the 4q haplotypes

a) The 9 main haplotypes identified by Lemmers *et al.* are shown, with their population frequencies on the right. The haplotypes are defined by three sequence variations, an SSLP proximal to D4Z4 (number of base pairs shown), a G/C SNP within D4Z4 which forms a PvuII site (C variant) and the distal qA/qB variation (the qA variant is shown in green, the qB variant is shown in orange). b) A schematic representation of the sequence variants used to define the D4Z4 haplotypes. The SSLP and P13E-11 regions are identified. The BlnI site in the first partial repeat and further restriction sites in the first proximal repeat are shown. At the distal end is the A/B/C variation.

1.15 Implications for the FSHD disease mechanism

Although conservation of the *DUX4* ORF suggests a protein coding function, expression of *DUX4* was not identified in any normal cell lines or tissues (Dixit *et al.*, 2007; Kowaljow *et al.*, 2007). However, unpublished data from Jane Hewitt's laboratory, generated by Jannine Clapp, had identified *DUX4* expression by RT-PCR from fetal skeletal muscle and testis.

Induction of high levels of *DUX4* overexpression in adult cell lines has been shown to be toxic (Gabriels *et al.*, 1999; Kowaljow *et al.*, 2007; Wallace *et al.*, 2011). These transfected cells show characteristic markers of apoptosis, which raises the possibility that inappropriate expression of the gene activates apoptotic pathways. Kowaljow *et al.* saw an increase in the release of cytosolic LDH and increased expression of caspases 3 and 7 at 48 hours after overexpression of *DUX4* in TE671 cells. Recently, Wallace *et al.* produced a *DUX4* construct with a mutation in the first homeodomain; expression of this mutant did not cause apoptosis, suggesting that DNA binding is required for the toxic effects (Wallace *et al.*, 2011).

The tandemly repeated structure of D4Z4 has also been conserved among species (Clapp *et al.*, 2007; Leidenroth and Hewitt, 2010), which suggests this organisation plays a functional role. One possibility is that the tandem repeats help to maintain a heterochromatic environment, silencing expression from the *DUX4* gene. Following the deletion, heterochromatin may open up resulting in inappropriate expression. This hypothesis could explain the observation that fewer repeats tend to correlate with a more severe phenotype (Brouwer *et al.*, 1995; Lunt *et al.*, 1995b; Tawil *et al.*, 1996; Klinge *et al.*, 2006); if a decrease in repeat number results in a less repressive structure, the chromatin could open to a greater extent, increasing the possibility of expression from the remaining repeats. However, as small arrays on 10q or 4qB do not result in disease, FSHD cannot result simply from the property of array organisation and so there are likely to be functional elements on 4qA alleles that are necessary for the disease. This hypothesis would also explain why monosomy for the 4q35 region does not result in FSHD, a factor that is difficult to account for when considering a position effect mechanism.

1.16 Future perspectives

The identification of a D4Z4 homolog in the mouse raises the possibility of manipulating the mouse *Dux* array to mimic FSHD. A model organism for the disorder would enable studies of the mechanisms involved in the FSHD pathology, and potentially increase our understanding of the pathways involved. Before a mouse model can be created it will be important to establish whether transcripts from the mouse *Dux* array play the same role in the mouse that DUX4 does in humans and to determine the relationships between mouse *Dux* genomic variants and transcripts.

Therefore, the aim of the work described in this thesis was to investigate whether the mouse and human arrays are functionally equivalent and initial work was focussed on whether PITX1 was a target for Dux in the mouse (Chapter 3), as had been described for human DUX4 (Dixit *et al*., 2007). The second aim of this work was to test for expression of human DUX4 in normal cell lines by RT-PCR and confirm expression in FSHD myoblasts. In order to map sequence variants within the D4Z4 arrays any expressed transcripts would be compared with genomic sequences (Chapter 4). Finally, sequence data would be gathered from the mouse repeat by RT-PCR and compared with the genomic sequence data (Chapter 5), in order to investigate whether sequ3ence variation and expression pattern is similar between mouse and human repeats.

Chapter 2. Methods and Materials

2.1 Cell Culture

2.1.1 Maintenance and passaging of cells

C2C12 (mouse myoblast), TE671, RD, RMS13 (human rhabdomyosarcoma) and hEC (human embryonic carcinoma) cells were cultured in 75cm^2 flasks containing 12ml media (see Table 2.9) and incubated at 37°C with 5% CO_2. Growth was checked daily and cells were passaged at around 80% confluency. When passaging the cells, all solutions were warmed to room temperature and cells were washed with 5ml phosphate buffered saline (PBS). Cells were incubated at 37°C for 5 to 10 minutes in 3ml of 0.25% trypsin and then dislodged from the bottom of the flask by gentle tapping against the hand. Appropriate media (7ml) was added to neutralise the trypsin and the cells were pipetted up and down to break up any clumps. Cells were typically split ~1:10 and so 1ml of suspension was added to a new flask with 11ml fresh media.

Primary human myoblast cell lines (GM17731, GM17940 and GM17869, Coriell Institute) were cultured in 75cm^2 flasks coated in 0.7% gelatin containing 12ml media (see Table 2.9) and incubated at 37°C with 5% CO_2. Growth was checked daily and cells were passaged before they reached 80% confluency to prevent them ceasing division in preparation for differentiation. Passaging of the cells was performed as above with a weaker concentration of trypsin (0.1%). Cells were typically split in a 1:5 ratio.

For immunocytochemistry, myoblast cell lines were cultured in 6 well plates on coverslips coated in poly-L-Lysine at a concentration of 500µg/ml.

2.1.2 Coating flasks in gelatin

Using a cell scraper, 150µl of 0.7% gelatin was spread evenly over the bottom of each 75cm^2 flask. Flasks were left for 30-60 minutes to allow the gelatin to solidify before cells

were added. Flasks were stored at 4°C for up to a month and stored flasks were warmed to 37°C before use.

2.1.3 Coating flasks in Matrigel

Matrigel was thawed on ice, overnight in the fridge and pipettes, racks and flasks were pre-chilled in the freezer overnight. Matrigel was diluted 10 times in differentiation medium without horse serum and 3ml added to each flask 75cm^2. Matrigel was spread evenly over the surface of the flask using a cell scraper and excess Matrigel was then removed. Flasks were left at room temperature to set then used immediately.

2.1.4 Differentiation of myoblasts

Myoblasts were grown in 75cm^2 flasks coated in Matrigel containing 12ml media. When cells reach 100% confluency, the growth media was removed and cells were washed in PBS. Differentiation media one (12ml) was added for 24 hours and then replaced with differentiation media two. Differentiation typically took 5-7 days.

2.1.5 Storage in liquid nitrogen

Freeze media was prepared in advance and kept on ice. Cells were trypsinized (as described in section 2.1.1), resuspended in media and the cell suspension transferred to 15ml Falcon tubes. The tubes were centrifuged at 600rcf for 5 minutes, the supernatant removed and the cells resuspended in the remaining droplet of media. Freeze media (1ml per 75cm^2 flask) was added to each tube, mixed by pipetting and then 0.5ml of cell suspension was transferred to a cryotube. The tubes were stored at -80°C for 48 hours before being transferred to liquid nitrogen.

2.1.6 Recovery from liquid nitrogen

Appropriate media was warmed to room temperature and added to a 75cm^2 flask. Cells were removed from liquid nitrogen and thawed quickly by incubating at 37°C. As soon as the cells were defrosted they were transferred to the flask and incubated at 37°C.

2.1.7 Transfections

Cells were transfected with PITX1 constructs that were cloned during this project (section 3.2.2.i) or with constructs that had been made previously in this lab (Figures 2.1 to 2.4).

All transfections were carried out in 6 well plates with a total volume of 3ml per well. C2C12 cells were seeded at a density of 2x10^5 and TE671 at a density of 4x10^5 and then incubated overnight prior to transfection. For immunocytochemistry, coverslips were placed at the bottom of each well. Non-transfected controls were treated exactly the same except for the addition of DNA. After transfection, cells were incubated at 37°C for 24-48 hours before being used for immunocytochemistry (section 2.2).

If required, proteosome inhibitor (Z-Leu-Leu-Leu-al), at a final concentration of 25 μM, was added to the wells and the cells left to incubate for 5 hours before immunocytochemistry was performed.

2.1.7.i Preparation of coverslips

A Duran bottle containing coverslips immersed in 20ml nitric acid (69%) and 10ml HCL (100%) was covered in foil and incubated at room temperature for 2 hours. The coverslips were washed four times with 75% ethanol, followed by a further two washes with 100% ethanol. Treated coverslips were stored in 100% ethanol.

When used for myoblasts, treated coverslips were placed in a single layer in a petri dish and covered with Poly-L-Lysine (500μg/ml) for 1 hour at 37°C. Coverslips were then allowed to air dry before use. Coated coverslips were covered with parafilm and stored at 4°C.

2.1.7.ii Transfection with Effectene

Cells were washed with 3ml PBS and 1.6ml media was added to each well. 400ng of purified plasmid DNA was diluted to 100µl with buffer EC. Enhancer (3.2µl) was added, vortexed for 1 second and then incubated at room temperature for 5 minutes. 10µl Effectene reagent was added and mixed by pipetting before being incubated at room temperature for 10 minutes. Media (600ml) was mixed with the transfection complexes and then transferred the appropriate well. Plates were swirled gently to ensure even dispersion of the transfection mixture. Non-transfected cells were treated the same except that no DNA was added.

2.1.7.iii Transfection with FuGene

Cells were washed in 3ml PBS and 3ml of fresh media was then added to each well. Purified plasmid DNA (2µg/well) was mixed with 100µl serum free media and 7µl FuGene reagent, vortexed and incubated for 15 minutes at room temperature. The transfection mixture was then added to the well under the surface of the media. The plate was swirled gently to ensure an even distribution of the complexes.

2.1.7.iv Transfection with GeneJuice

Cells were washed in 3ml PBS and 3ml of fresh media was then added to each well. GeneJuice reagent (3µl) was added to 100µl serum free media, vortexed and incubated at room temperature for 5 minutes. Purified plasmid DNA (1µg /well) was then added and mixed by pipetting before incubation at room temperature for a further 15 minutes. The transfection mixture was then added to the correct well and the plate was swirled gently to ensure an even distribution of the transfection complexes.

2.1.7.v Transfection with Transpass

Cells were washed in 3ml PBS and 3ml fresh media added to each well. 3µg purified plasmid DNA was added to 250µl serum free media and 6-12µl Transpass reagent. The mixture was incubated at room temperature for 30 minutes and then added to the cells. The plates were swirled gently to ensure an even distribution of the transfection complexes.

2.2 Immunocytochemistry

Cells to be used for immunocytochemistry were seeded onto coverslips in 6 well plates; coverslips for myoblasts were first coated with Poly-L-Lysine. If cells were transfected, immunocytochemistry was carried out 24 or 48 hours after transfection.

A volume of 3ml per well was used for each solution. Cells were washed twice in PBS and then fixed at room temperature for 10 minutes with 4% PFA. After three washes in PBS, cells were permeablised with 0.5% Triton in PBS for 10 minutes on ice and then washed a further three times. Next, the cells were incubated for 30 minutes at room temperature in 10µg/ml bovine serum albumin (BSA)/5% goat serum in PBS to block non-specific binding. Coverslips were then placed cells down onto 100µl primary antibody diluted in 10mg/ml BSA/5% goat serum in PBS and incubated in the dark for 1h. After washing with PBS 3 times for 5 minutes, the process was repeated with 100µl secondary antibody diluted in 10mg/ml BSA in PBS. After the 1h incubation the cells were washed 3 times for 5 minutes in PBS then mounted onto slides using 15µl vectashield with DAPI. The edges of the coverslips were sealed with nail varnish and slides were stored at 4°C in the dark. Antibody dilutions are provided in Table 2.1.

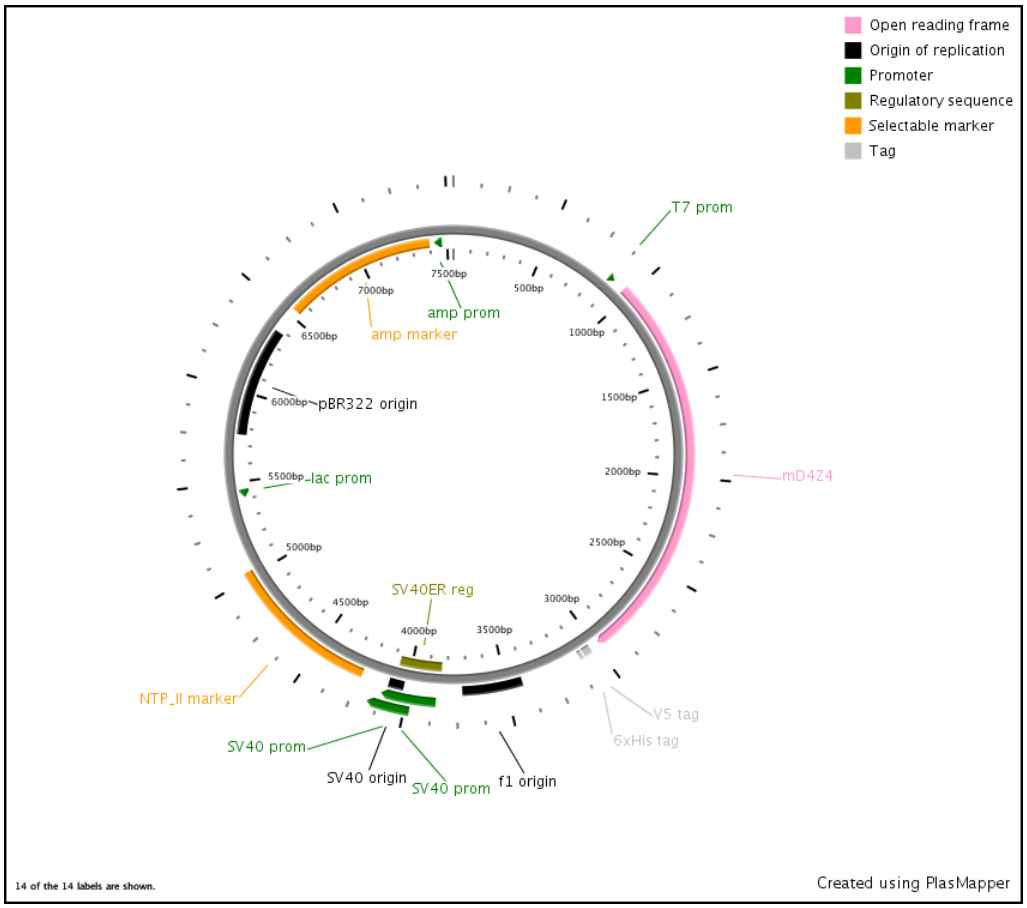


Figure 2.1 Map of the FlmD4Z4 V5 construct

This construct will produce a full length mouse D4Z4 protein tagged at the carboxy terminus with the V5 epitope. This construct was generated by Jonathan Rowlinson.

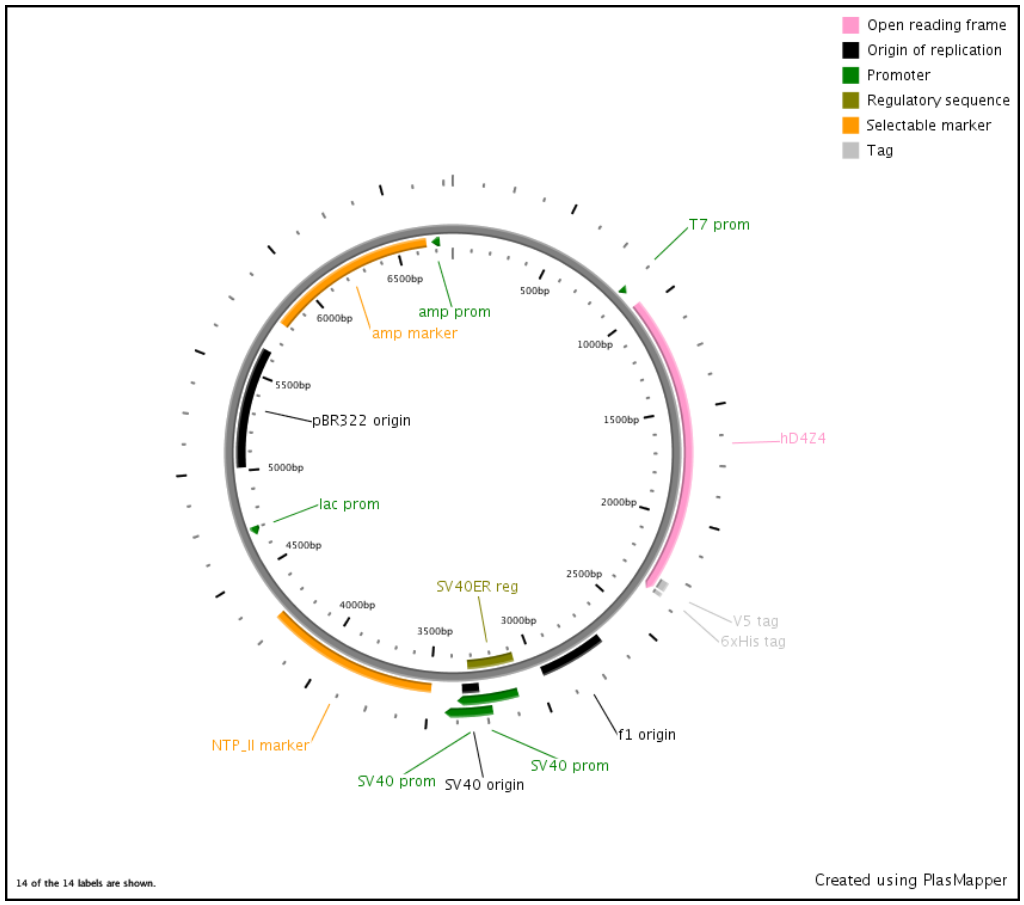


Figure 2.2 Map of the FlhD4Z4 V5 construct

This construct will produce full length human D4Z4 tagged at the carboxy terminal end with the V5 epitope. This construct was generated by Jonathan Rowlinson.

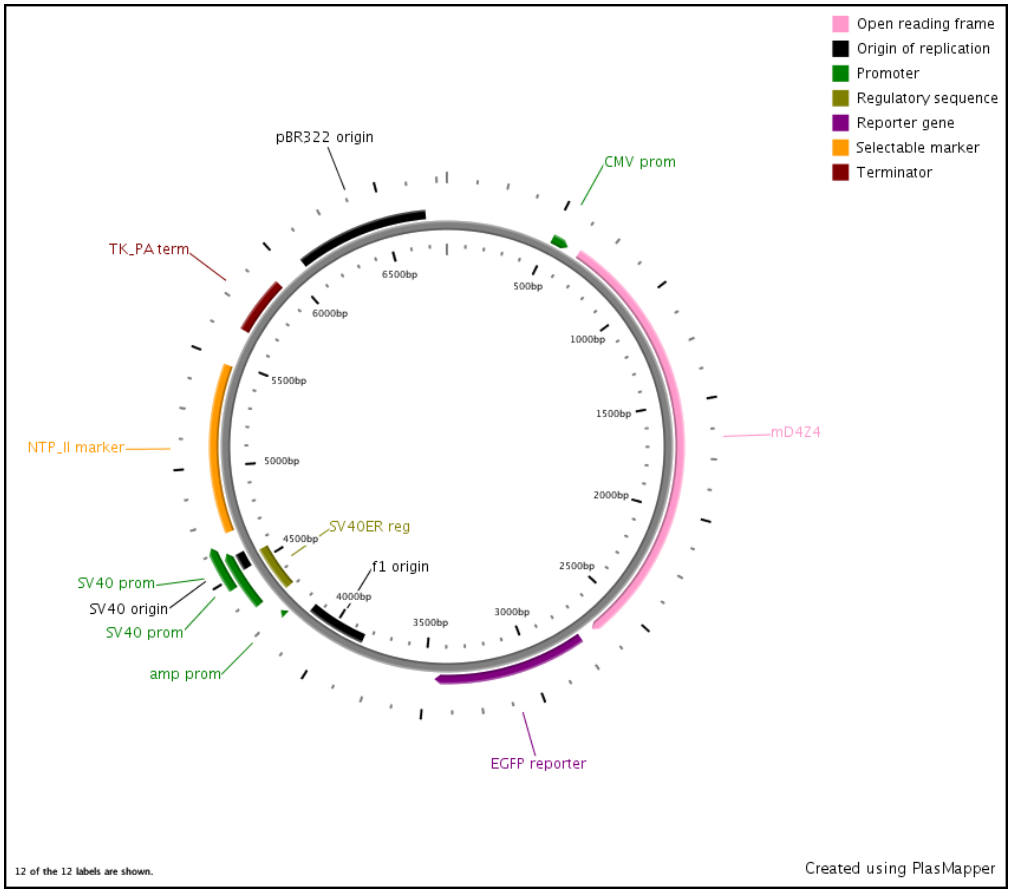


Figure 2.3 Map of the FlmD4Z4 GFP construct.

This construct will produce full length mouse D4Z4 tagged at the carboxy terminus with EGFP. This construct was generated by Jannine Clapp.

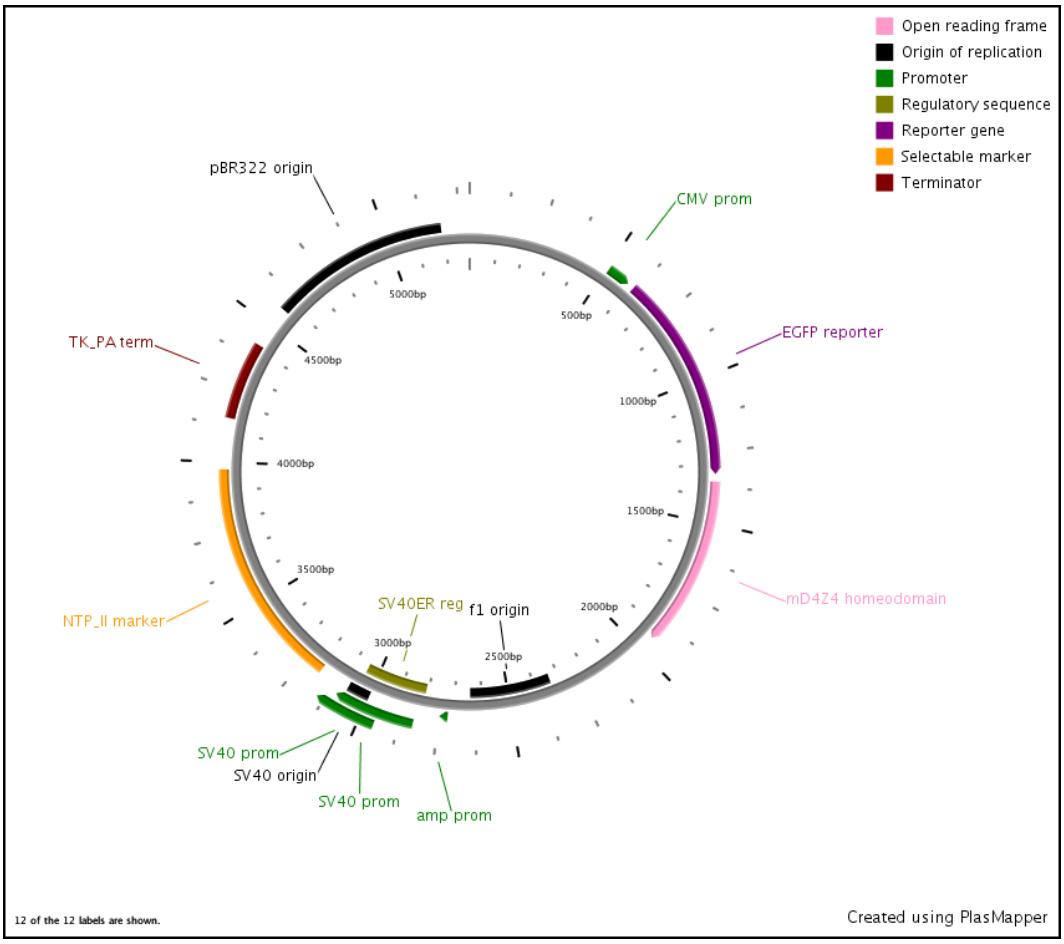


Figure 2.4 Map of GFP NtermD4Z4.

This construct will produce the N-terminus and homeodomain region tagged at the amino terminus with EGFP. This construct was generated by Jannine Clapp.

a.

Antibody	Raised in	Company/Source	Dilution
Anti-V5	Mouse	Invitrogen	1/200
Anti-PITX	Rabbit	Dr Yi-Wen Chen, Children's Hospital, Boston	1/1000
Anti-myc (9E10)	Mouse	Iowa Hybridoma Bank	1/800
Anti-Desmin	Rabbit	Sigma	1/500
MF20	Mouse	Donald A. Fischman, Cornell University, New York	1/50
9A12	Mouse	Dr Yi-Wen Chen, Children's Hospital, Boston	1/1000

b.

Antibody	Company	Dilution
Goat anti-rabbit rhodamine redex	Molecular Probes	1/500
Goat anti-mouse rhodamine redex	Molecular Probes	1/500
Goat anti-mouse IgG FITC	AbCam	1/128

Table 2.1 Antibodies used for immunocytochemistry

Source and antibody dilution used in immunocytochemistry for primary antibodies (a) and secondary antibodies (b).

2.3 Imaging and Photography

Cells were viewed using an Axioscope 2 compound microscope (Zeiss) and photographs were taken using a MTI 3CCD camera and the OpenLab software package (Improvision).

2.4 RNA Purification

2.4.1 RNA extraction from mammalian cells

RNA was extracted from cultured cells using TRIzol reagent. Cells were trypsinized as described previously (section 2.1.1) and pelleted by centrifugation at 600rcf for 5 minutes in 15ml Falcon tubes. The supernatant was removed and the cells were washed with PBS. Cells were re-pelleted by centrifugation for a further five minutes, the supernatant was removed and the cells were kept on ice. The pellets were resuspended in 2ml TRIzol per 75cm^2 flask and mixed by pipetting before incubating at room temperature for 5 minutes. Chloroform (0.2ml per 1ml TRIzol) was added and the tubes shaken vigorously before incubating at room temperature for 5 minutes. The cells were then centrifuged at 4°C for 15 minutes at 16100rcf. The upper aqueous phase was transferred to a new tube and 0.5ml propan-2-ol added per 1ml TRIzol. Samples were incubated at room temperature for 10 minutes and then again centrifuged at 16100rcf for 10 minutes at 4°C. The supernatant was removed and the pellet was washed in 75% ethanol, vortexed and then centrifuged at 5900rcf for 5 minutes at 4°C. The supernatant was removed and the pellet air-dried before resuspending in 10µl sterile distilled water. The RNA was stored at -80°C. RNA was examined by gel electrophoresis on 1% agarose gel (Figure 2.5a) and the quality of the RNA was checked with a nanodrop (ND-1000 Spectrophotometer, using software version 3.2.1); RNA with an A260/A280 ration of 1.8-2.0 was used in further reactions. The integrity of all RNA samples was confirmed by a control RT-PCR using primers for β-actin (Figure 2.5b).

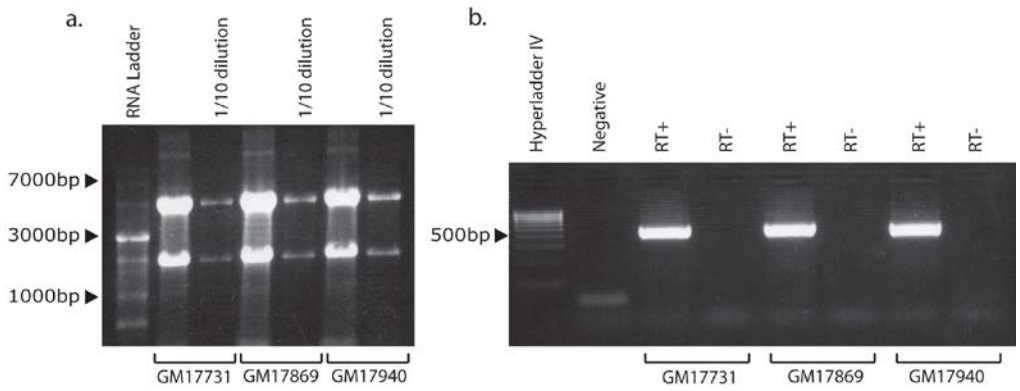


Figure 2.5 Quality analysis of RNA extracted from myoblast cells

a) Gel electrophoresis of 2µl undiluted and 1:10 diluted RNA isolated from myoblast cell lines. RNA ladder from New England Biolabs. b) RT-PCR of myoblast RNA with primers for β-actin (see Table 2.5 for sequences). Reverse transcription was carried out with Superscript III and the PCR step with Biotaq polymerase.

2.4.2 Poly (A)$^+$ RNA purification

Poly (A)+ mRNA was isolated from total RNA using the PolyATract mRNA Isolation system (Promega). Total RNA (500µg) was made up to 500µl volume with RNase-free water and incubated at 65°C for 10 minutes. After incubation, 150pmol Biotinylated-Oligo(dT) Probe and 13µl 20x SSC was added and the reaction allowed to cool to room temperature for 10 minutes. Streptavidin MagneSphere Paramagnetic Particles (SA-PMPs) were rinsed three times with 300µl 0.5X SSC while in the magnetic stand and then resuspended in 100µl 0.5X SSC.

The contents of the annealing reaction were then added to the washed particles and incubated for 10 minutes at room temperature with gentle inversion every 2 minutes. The SA-PMPs were captured using the magnetic stand, the supernatant was removed and the particles washed four times in 300µl 0.1X SSC. The final SA-PMP pellet was resuspended in 100µl for RNase-free water and captured with the magnetic stand. The eluted mRNA (supernatant) was transferred to a new tube and the elution step repeated with 150µl RNase-free water.

2.5 DNase treatment of RNA

DNase treatment was carried out using TURBO DNA-*free* (Ambion). 10µg RNA was added to 1X reaction buffer, sterile distilled water and 2U DNase in a total volume of 18µl and then incubated at 37°C for 30 minutes. A further 2U DNase was added and the sample incubated for another 30 minutes at 37°C. To inactivate the enzyme 2µl inactivation reagent was added to the sample and incubated at room temperature for 2 minutes with occasional gentle mixing. The sample was centrifuged at 9300rcf for 90 seconds in an Eppendorf microfuge and the supernatant transferred to a new tube; the concentration of RNA was typically 0.5µg/µl.

2.6 DNA Purification

2.6.1 Genomic DNA extraction from mammalian tissues

Tissues were washed in PBS, cut into small pieces and transferred to sterile tubes. The tissue was washed twice in PBS then resuspended in 4.12ml DNA buffer. After the addition of 100µl of 20mg/ml proteinase K and 480µl of 10% SDS the tissue was incubated at 40°C overnight. If there were still visible tissue pieces, the samples were homogenised gently and another 100µl proteinase K was added before incubation at 40°C for 4 hours. A phenol/chloroform extraction was then performed on the samples. The samples were mixed with 4.8ml phenol (pH 7.8) on a rotator for 10 minutes and then centrifuged at 720rcf for 5 minutes. The supernatant was removed to a new tube and mixed with 2.4ml phenol and 2.4ml chloroform:isoamyl alcohol (24:1) and the rotator and spin were repeated. Finally, the supernatant was removed to a new tube and 4.8ml chloroform:isoamyl alcohol (24:1) was mixed to the sample and the rotator and spin repeated. After centrifugation, 100µl 3M NaAc and 20ml of 100% ethanol were added and mixed gently to precipitate the DNA. The DNA was transferred to a new tube, washed in 70% ethanol and allowed to air dry. DNA was dissolved in 1ml sterile distilled water on at rotator at 4°C overnight.

2.6.2 Genomic DNA extraction from mammalian cells

Cells were trypsinised as described in section 2.1.1 and pelleted by centrifugation at 600rcf for 5 minutes in 15ml Falcon tubes. The supernatant was removed and the cells were washed with PBS. Cells were re-pelleted by centrifugation for a further five minutes, the supernatant was removed and the cells were kept on ice. The pellet was resuspended in 4.12ml DNA buffer then 100µl of 20mg/ml pronase and 480µl 10% SDS were added and the sample incubated at 40°C overnight. The following day phenol chloroform extraction was performed as described previously (Section 2.6.1) and DNA was dissolved in 1ml SDW at 4°C overnight.

2.6.3 Plasmid DNA minipreps

To analyse inserts by restriction enzyme digestion, plasmid minipreps were used to isolate DNA. 500µl of an overnight culture was centrifuged at 16100rcf for 1 minute to pellet the cells. The supernatant was removed and the pellet resuspended in 200µl of solution 1. Solution 2 (200µl) was added to lyse the cells and mixed gently by pipetting. Next, solution 3 (300 µl) was added and the solution mixed in by inversion. The samples were centrifuged at 16100rcf for 10 minutes and the supernatant transferred to a fresh tube. To precipitate the DNA, 1ml of 100% ethanol was added to each sample and then stored at -20°C for 20 minutes. The samples were centrifuged at 16100rcf for 10 minutes, washed in 70% ethanol and then centrifuged for a further 5 minutes before leaving the pellets to air-dry. The DNA was resuspended in 20µl digest mix.

2.6.4 Plasmid DNA minipreps using columns

When a higher purity of DNA was required, the GenElute plasmid miniprep kit (Sigma) was used according to the manufacturer's instructions. The DNA was eluted from the columns using 50µl sterile distilled water.

2.6.5 Purification of PCR and digestion products

The GenElute PCR Clean-up kit (Sigma) was used to remove contaminants from PCR and digestion products. Manufacturer's instructions were followed and the DNA was eluted into 30µl of sterile distilled water. To purify fragments <200bp the MinElute PCR purification kit (Qiagen) was used and the DNA eluted into smaller volumes, typically 10µl.

2.6.6 Extraction of DNA from agarose gels

The correct sized bands were excised from the gel using a clean scalpel blade and the agarose removed by centrifugation through polymer wool for 10 minutes at 9300rcf. The supernatant was transferred to a spin column, the sample centrifuged for a further 5 minutes at 16100rcf and any further supernatant recovered was also transferred. The MinElute PCR purification kit (Qiagen) was used to purify the DNA as described in Section 2.6.5.

2.7 Digestion of DNA

2.7.1 Plasmid DNA

Miniprep digest reactions contained 1µl restriction enzyme, 1X restriction buffer, 0.5µl RNase A, 1µg DNA and sterile distilled water to a total volume of 20µl. DNA that was purified using the GenElute plasmid miniprep kit was digested in the same way but without the addition of RNase A. The reactions were incubated at 37°C for 2-6 hours. Restriction enzymes are shown in Table 2.8.

2.7.2 Genomic DNA

Typically, 500ng of DNA was digested with 100units of restriction enzyme, 1x restriction buffer and 10µg BSA in a total volume of 100µl. Samples were incubated at 37°C overnight. Restriction enzymes are listed in Table 2.8.

2.8 Modification of DNA ends

2.8.1 Pfu treatment

Treatment with Pfu was used to polish the ends of Taq-generated PCR fragments prior to ligation into the pSMART vector. The PCR fragments were added to 3U Pfu polymerase,

100µM dNTP, 1X Pfu buffer with MgSO4 in a final volume of 50µl. Reactions were incubated at 72°C for 30 minutes and then cleaned using the GenElute PCR Cleanup kit (see Section 2.6.5).

2.8.2 T4 Polynucleotide Kinase treatment

T4 Polynucleotide Kinase was used for the addition of 5'-phosphates to fragments prior to ligation into the pSMART vector. PCR products were added to reactions containing 1X T4 DNA ligase buffer and 10U T4 Polynucleotide Kinase to a final volume of 50µl. Reactions were incubated at 37°C for 30 minutes followed by 65°C for 20 minutes and then cleaned using the GenElute PCR Cleanup kit (see section 2.6.5).

2.8.3 Calf intestinal phosphatase treatment

Digested vectors were treated with calf intestinal phosphatase to remove 5' phosphate groups and reduce self-ligation. 2U CIP contained in 1X NEB buffer 3 was incubated at 37°C for 30 minutes. The reactions were then incubated at -20°C to inactivate the CIP and then cleaned up with the GenElute PCR Cleanup kit (Section 2.6.5).

2.9 Nucleic Acid Electrophoresis

2.9.1 Gel electrophoresis

After electrophoresis all gels were imaged using a Gel Doc™ XR+ System with Quantity One software (version 4.6.5, Build 094)

2.9.1.i DNA

Agarose gels were prepared in 1X TAE buffer with 0.5µg/ml ethidium bromide. Gels were run in 1X TAE buffer for 1 hour at 50V. Commercial ladders from Bioline (Hyperladder I or IV) were used as size standards.

2.9.1.ii RNA

Agarose gels (1%) were prepared in 1X MOPS buffer. 2µl RNA sample or ladder were mixed with 2µl ethidium bromide and 6µl formaldehyde loading buffer, incubated at 80°C for 10 minutes and then put immediately onto ice. Samples were briefly centrifuged before being loaded onto the gel. Gels were run in 1X MOPS buffer for 3-4 hours at 50V.

2.9.2 Capillary electrophoresis

Products of the SSLP genotyping reaction modified from Lemmers *et al.* (2010) were sized by capillary electrophoresis on Applied Biosystems 3130*xl* Genetic Analyzer. 10µl of a 16-reaction master mix (2µl ROX Genescan and 170µl Hi-Di Formamide) was added to each sample. The plate was covered with a grey rubber lid, heat denatured at 96°C for 3 minutes and then transferred to ice for a further minute before being loaded onto the machine. Samples of known alleles were run as size standards.

2.10 Subcloning

2.10.1 Ligations

2.10.1.i pcDNA

To make constructs for ligation, reactions were set up as follows: 1µl CIP treated vector, 1µl purified insert, 1X ligation buffer and 1µl T4 DNA ligase were mixed and incubated at room temperature for 15 minutes.

2.10.1.ii pSMART vectors

In each reaction, 50-100ng of blunt-ended, 5'-phosphorylated insert DNA, 2.5µl 4X CloneSmart Vector premix and 1µl CloneSmart DNA ligase were added to sterile distilled water to a final volume of 10µl. Reactions were incubated at 16°C for 16 hours and then heat denatured at 70°C for 15 minutes.

2.10.1.iii pGEM-T Easy vector system

Typically, reactions were carried out in a 10µl volume containing 50-100ng of PCR product, 50ng of pGEM-T Easy vector, 3U/µl T4 DNA ligase and 1x rapid ligation buffer. Reactions were incubated at 4°C overnight.

2.10.2 Transformations

Competent cells (see Table 2.2) were incubated on ice for 30 minutes with 1µl ligation mixture. The cells were then heat shocked at 42°C for 45 seconds before being returned to ice for a further 2 minutes. 200 µl of SOC media (DH5α) or 960 µl Recovery Medium (*E.cloni*) were added to the cells and incubated with shaking at 80rcf for 1hr at 37°C. 200µl of cells were then plated onto agar plates containing the appropriate antibiotic (see Table 2.2) and incubated overnight at 37°C.

Vector	Competent Cells	Agar	Antibiotic	Concentration
pcDNA 3.1	DH5α	LB	Ampicillin	50µg/ml
pGEM-T Easy	DH5α	MacConkey	Ampicillin	50µg/ml
pSMART	*E.coni*	YT	Kanamycin	30µg/ml

Table 2.2 Antibiotic requirements of cloning vectors

2.11 Glycerol Stocks

Once clones that carried the correct sequence were identified, glycerol stocks were made for long-term storage. 1ml of an overnight culture was added to a cryotube with 0.5ml 90% glycerol and mixed by inversion. Samples were stored at -80°C.

2.12 PCR

Typical reactions and cycling conditions are described in Table 2.3. Primers sequences are listed in Tables 2.4 and 2.5.

2.13 RT-PCR

2.13.1 Standard RT-PCR

RNA was first treated with DNase to remove contaminating DNA (see Section 2.5) First strand cDNA was synthesised using either the ImProm-II reverse transcription system (Promega) or SuperScript III reverse transcriptase (Invitrogen). The quality of cDNA was checked using a control PCR with primers for β-actin (Tables 2.4 and 2.5).

Typically, 5µl of first strand cDNA was used in a 25µl PCR reaction. For standard PCR conditions see Table 2.3.

Polymerase	Reaction Mix		Cycling Conditions					
	Reagents	Total Volume	Initial Denaturation	Denaturation	Annealing	Extension	Number of Cycles	Final Extension
Pfu (Promega)	Template as required 2-3U/μl Pfu 1X reaction buffer 10mM dNTP 1μM each primer	50μl	95°C 1 min	95°C 30 sec	30 sec	72°C 30 sec	30	72°C 5 min
Proofstart (Qiagen)	Template as required 2.5U/μl Proofstart 1X reation buffer 300μM dNTP 1μM primer 1X Q-Solution 10% DMSO	50μl	95°C 5 min	94°C 1 min	1 min	72°C 1 min	35	72°C 1 min
LA Taq (Takara)	Template as required 5U/μl LA taq 1X reaction buffer 2.5mM dNTP 1μM each primer	50μl	94°C 1 min	94°C 30 sec	30 sec	72°C 2 min	30	72°C 5 min
Primestar (Takara)	Template as required 2.5U/μl PrimeSTAR 1X reaction buffer 2.5mM dNTP 0.3μM each primer	50μl		98°C 10 sec	5 sec	72°C 1 min	30	
Taq (Abgene)	Template as required 5U/μl Taq 1X reaction buffer 0.2mM dNTP 0.5μM each primer 1.5mM $MgCl_2$	25μl	94°C 1 min	94°C 30 sec	30 sec	72°C 1 min	30	

KOD (Novagen)	Template as required 1.5U/µl KOD 1X reation buffer 0.2mM dNTP 0.4µM each primer	50µl	98°C 5 min	98°C 20 sec	30 sec	72°C 1 min	35	72°C 5 min
KAPA HiFi (GRI)	Template as required 1µl KAPA HiFi 1X reaction buffer 0.2mM dNTP 0.3µM each primer	50µl	95°C 1 min	98°C 20 sec	15 sec	68°C 30 sec	35	68°C 5 min
BioMix Red (Bioline)	Template as required 1X BioMix 10pmol each primer	25µl	95°C 2 min	95°C 30 sec	30 sec	72°C 30 sec	30	
Platinum Taq (Invitrogen)	Template as required 1X PCRx buffer 0.2mM dNTP 1.5mM $MgSO_4$ 0.2µM each primer 2.5U Platinum *Taq*	20µl		94°C 30 sec	30 sec	68°C 2.5 min	35	68°C 7 min
Phusion (NEB)	2ng template 1X Phusion HF buffer 0.2mM dNTP 0.5U Phusion enzyme 0.5µM each primer	25µl	98°C 3 min	98°C 15 sec	30 sec	73°C 15 sec	32	72°C 30 sec

Table 2.3 PCR conditions

Reagents and cycling conditions for PCRs described in this work. Annealing temperature varied with primer pair, see Table 2.4 for details.

Primer Name	Sequence	Annealing Temp	Product Size	Reference
MPITX1F	CCG GCG GAA TTC ATG GCG CAA GCG GGA GCG GA	55°C	972bp	
MPITXRSful	TCC CTA TTC GAA GTT CAC GGG TTG AGG CGC TC			
mD4Z4_ATG_F2	AGT CGA TTC TCC CAA GGT GA	59°C	2633bp	Clapp *et al.* (2007)
mD4Z4_TGA_R2	CAC AGC TCT GCA TGA AGC AT			
mD4Z4_nested_F	CTG TTG GTG GCA GTG GTG T	64°C	2071bp	
mD4Z4_nested_R	CGG GTC TCT TCA GGA CTT TG			
mD4Z4_TGA_F	AAC TGC TGA CCG AAG TCC AA	57°C	586bp	Clapp *et al.* (2007)
mD4Z4_TGA_R2	CAC AGC TCT GCA TGA AGC AT			
m_D4Z4_F2	GCA CTC AAG CAG ACA GCA CA	57°C	360bp	Clapp *et al.* (2007)
m_D4Z4_R2	GTG TCC ATT TGC TCC CAT GT			
mD4Z4_F8	GCC CAC AGC TCA AGA TCA AG	59°C	271bp	Clapp *et al.* (2007)
mD4Z4_10	ATC AAG GAG GGG TTC CAG AG			
mD4Z4_ATG_F	TTT AAG GGG CAG TGG TCA CA	59°C	310bp	Clapp *et al.* (2007)
mD4Z4_ATG_R	CCA GCT CCT TCC TCT CCT TG			
mB-actin-f	GAG AGG TAT CCT GAC GAA G	59°C	1400bp	
mbeta-actin-R	ACG CGA CCA TCC TCC TCT TA			
MspI1-F	TCA AGG AAG GGG GTC TGA AA	61.9°C	556bp	
MspI1-R	TGT GAC CAC TGC CCC TTA AA			
MspI2-F	GTA GTG GGC GGG ACT ACC TG	65.6°C	451bp	
MspI2-R	AGC TCT TCC TGG AGC TGT GG			
MspI3-F	GGA TCC TAG GGC AAG CCT TT	65.6°C	413bp	
MspI3-R	AAG GAG GGG TTC CAG AGA GC			
MspI4-F	TAG AAA CCC GAG AGG CCA AA	67.1°C	408bp	
MspI4-R	GAG TCC ACT CTC CCC ATT GC			
MspI5-F2	TGG ACT CCT TCC CTG GCT TA	63.1°C	447bp	
MspI5-R2	CTG GAC ATA GCC CAG CCT TC			
PgkI_F	CAC GCT TCA AAA GCG CAC GTC T	56.1°C	170bp	
PgkI_R	CTT GAG GGC AGC AGT ACG GAA T			
M114	ATT CTG GTT GTG CCG AGT TGC GAG	60°C	489bp	Cristina Tufarelli (personal communication)
M115	CGT GGC ACT TTT GAG TTC ATC TCT C			
MspI_ctrl_F2	GAA GGT TCG TGG GAG AGC	64°C	503bp	
MspI_ctrl_R2	CCC TGC TCC TCA AGT TGG AC			

Table 2.4 Primers for amplification of mouse sequences

Primer Name	Sequence	Annealing Temp	Product Size	Reference
HDUX4_F1	GGA GGA GCT GGC CAG AGA GAC	59°C	806bp	Jannine Clapp (Personal communication)
HDUX4_R1	CCT CCG TTT CTA GGA GAG GTT G			
1797	ATG GCC CTC CCG ACA CCC T	65°C	109bp	Snider *et al.* (2009)
1906	CGA TGC CCG GGT ACG GGT TCC GCT CAA AGC			
1797	ATG GCC CTC CCG ACA CCC T	67.5°C	438bp	Snider *et al.* (2009)
2235	GCC CTG CCA CCC TGT CCC GGG TGC CTG GC			
2672	GCC GCA GGG CCA AGG GGT GCT TGC GCC ACC	66°C	280bp	Snider *et al.* (2009)
2952	CTA GGA GAG GTT GCG CCT G			
4q7F	GCA GAG CTC TCC TGC CTC T	60°C	352bp	
4q7R	GAG AGA GGA ACG GGA GAC CT			
4q8F	ACG GAG GTT CAG TTC CAC AC	62°C	342bp	
4q8R	CTC ACC GCC ATT CAT GAA G			
D4F104S1 F	CCC AGT TAC TGT TCT GGG TGA	58°C	523bp	Lemmers *et al.* (2007)
D4F104S1 R	GAA AGC CCC CTG TGG GAG			
Dux4C#49	CGC GTC CGT CCG TGA AAT TCC	64°C	1233bp	Ansseau *et al.* (2009)
Dux4C#167	GGG AGC TGA AGG ACC TGG			
EX2BF	CCA GAG TCC AGC TCA GGC	55°C	684bp (G) 360bp (mRNA)	Gabelline *et al.* (2002)
SSC8	CTC ACA GGT AAG TGG AGA ATG G			
#222	AGT GCA CAG TCC GGC TGA	59°C	1477bp	Kowaljow *et al.* (2007)
#219	GTA GCC AGC CAG GTG TTC C			
#222	AGT GCA CAG TCC GGC TGA	56°C	2000bp or 1700bp	Dixit *et al.* (2007)
#407	TCC AGG TTT GCC TAG ACA GC			
mycSfuF	GGG CCC TTC GAA CAA AAA CTC ATC TCA GAA GAG G	60°C	57bp	
MycAgeR	CTC AGA AGA GGA TCT GAA TAT GCA TAC CGG TCA T			
HPITX1F	GTC CCC GAA TTC ATG GAC GCC TTC AAG GGG GG	55°C	966bp	
HPITX1RSful	GCG GGG TTC GAA GCT GTT GTA CTG GCA CGC GT			
myh2F	GGA CCA ACT GAG TGA ACT GAA A	59°C	156bp	
myh2R	GTG TCT CAG TTA TCA AGA GGC AA			
hdesminF	CCT ACT CTG CCC TCA ACT TC	61°C	2564bp (G) 519bp (mRNA)	
hdesminR	AGT ATC CCA ACA CCC TGC TC			
D4F104SF1	CCC AGT TAC TGT TCT GGG TGA	60°C	2494bp	
pLAM1	GGC CGG TTT GGA ACC TGG			
14A	CCC CGA GCC AAA GCG AGG CCC TGC GAG CCT	62°C	1572bp or 1436bp (DUX4-fl) 630bp (DUX4-s)	Snider *et al.* 2010
174	GTA ACT CTA ATC CAG GTT TGC CTA GA			
15A	CGG CCC TGG CCC GGG AGA CGC GGC CCG C	62°C	1399bp or 1263bp (DUX4-fl) 457bp (DUX4-s)	Snider *et al.* 2010
175	TCT AAT CCA GGT TTG CCT AGA CAG C			
SSLPF1	GGT GGA GTT CTG GTT TCA GC	59°C	165bp	
SSLPR1	CCT GTG CTT CAG AGG CAT TTG			
2131-FAM	[6FAM] GGT GGA GTT CTG GTT TCA GC	61°C	163bp	
SSLPR	CCT GTG CTT CAG AGG CAT TTG			
hbeta-actinF	GAG GCA TCC TCA CCC TGA AG	59°C	500bp	
hbeta-actinR	GGC CAT CTC TTG CTC GAA GT			

Table 2.5 Primers for amplification of human sequences

2.13.1.ii ImProm-II reverse transcription system

Reaction tubes were cooled on ice, 1µg DNase-treated RNA was combined with 0.5µg Oligo(dT)$_{15}$ primer and sterile distilled water in a final volume of 5µl. The samples were incubated at 70°C for 5 minutes and then immediately put on ice for a further 5 minutes. Reaction mix containing reaction buffer, 25mM MgCl2, 10mM dNTP mix, 20u RNasin and 1µl Reverse Transcriptase in a total volume of 20µl. Reactions were then incubated at 25°C for 5 minutes and then at 42°C for one hour. The reverse transcriptase was inactivated by incubation at 70°C for 15 minutes.

2.13.1.iii Superscript III reverse transcription

DNase treated RNA (2µg) was mixed with 0.5µg oligo(dT)$_{15}$ primer and 10mM dNTP in a volume of 24µl. The sample was incubated at 65°C for 5 minutes and then transferred to 55°C. During the incubation 2X First-Strand buffer, 25mM $MgCl_2$, 0.1M DTT, 200U SuperScript III RT were mixed and heated to 55°C. Once both samples were at 55°C, the reactions were mixed and incubated at 55°C for 1 hour followed by 85°C for 5 minutes.

2.13.2 OneStep RT -PCR

OneStep RT-PCR (Qiagen) was carried out according to manufacturer's instructions. 1.2µg DNase-treated RNA was used in a reaction volume of 50µl containing 1X reaction buffer, 400uM of each dNTP, 0.6uM primer, 1X Q solution and 2µl enzyme mix. Samples were then incubated at 50°C for 30 minutes followed by 95°C for 15min to inactivate the reverse transcriptase and activate the DNA polymerase. 40 cycles of: 94°C for 1 minute (denaturation), 59°C for 1 minute (annealing) and 72°C for 1 minute were followed by a final extension at 72°C for 10 minutes. For no reverse transcriptase (–RT) controls, RNA was added at the end of the 15-minute inactivation step. Genomic DNA was added instead of RNA for a positive control.

2.14 DNA sequencing

2.14.1 Sequencing reaction

Sequencing reactions were carried out using Big Dye Terminator V3.1 at a 1/12 dilution, with 200ng DNA, 1X sequencing buffer and 2pmol/µl primer in a total volume of 5µl. Sequencing primers are shown in Table 2.8. The reaction was then incubated at 95°C for 5 minutes, followed by 45 cycles of 95°C for 30 seconds, 50°C for 20 seconds and 60°C for 4 minutes.

2.14.2 Precipitation in individual tubes

The volume was adjusted to 100µl with sterile distilled water and 1 µl of 10mg/ml glycogen, $1/10^{th}$ volume 3M NaAc and 2 volumes 100% ethanol were added. The reactions were then stored at -20°C for 20 minutes. The samples were centrifuged at 16100rcf for 10 minutes, washed with 70% ethanol and then centrifuged at 16100rcf for 5 minutes. Pellets were left to air dry and then analysed by the Geneservice DNA sequencing facility in Nottingham.

2.14.3 Purification of DNA sequencing products in microtitre plates

When sequencing reactions were performed in 96-well plates, the CleanSEQ system was used for purification. The CleanSEQ was shaken to resuspend the magnetic particles and 10µl was added to each sample and mixed with 42µl of 85% ethanol by pipetting. The reaction plate was left on the magnet for 3 minutes to separate the beads from the solution, the supernatant was discarded and the beads washed twice with 100µl of 85% ethanol. The plate was then allowed to air dry for 10 minutes. The reactions were analysed by the Durham Sequencing Service (DBS). Primers used for sequencing are shown in Table 2.8.

Primer name	Primer Sequence
T7	AAT TAA CCC TCA CTA AAG GG
BGH	TAG AAG GCA CAG TCG AGG
LT7	GTA ATA CGA CTC ACT ATA GGG C
SP6	GAT TTA GGT GAC ACT ATA G
SL1	CAG TCC AGT TAC GCT GGA GTC
SR2	GGT CAG GTA TGA TTT AAA TGG TCA GT

Table 2.6 Primers used for sequencing

2.15 Computer analysis of DNA sequences

2.15.1 BLAST

BLAST analysis was carried out using the web interface at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Sequences were copied and pasted into the input window and generally, all options were left at the default settings.

2.15.2 Sequencher

Detailed sequence analysis was carried out using the Sequencher program version 4.10.1 (GeneCodes, USA).

2.16 Protein analysis

2.16.1 Protein extraction from mammalian cells

Protein extraction was carried out on ice. Media was removed from the culture flasks and the cells were washed with ice cold PBS. Cells were scraped off the flask into fresh PBS, transferred to eppendorf tubes and centrifuged at 720rcf for 3 minutes. The cell pellet was resuspended in 100µl RIPA buffer per 75cm3 flask and incubated for 30 minutes on ice. After

incubation, cells were centrifuged at 16100rcf for 10 minutes at 4°C. The supernatant was transferred to a new tube and stored at -80°C.

2.16.2 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Polyacrylamide 10% gels (1.5mm thick) were cast using Mini-PROTEAN 3 casting frame and stand (Biorad). Resolving gels were made up with 1X lower buffer and polymerised with 0.05% ammonium persulphate (APS) and 0.13% TEMED. Stacking gels were made up with 1x upper buffer and were polymerised with 0.03% APS and 0.05% TEMED.

Protein samples were mixed with 1X DTT and 4µl of loading buffer in a final volume of 20µl and heated to 80°C for 5 minutes prior to loading into the wells. Electrophoresis was performed in 1X running buffer at 100V until the protein had migrated. Commercially available protein markers were used as size standards (section x).

2.16.3 Western blotting

Hybond-P membrane was activated by soaking in methanol and then soaked in 1X transfer buffer along with gels from SDS-PAGE for equilibration. Transfer was carried out in 1X transfer buffer in a mini transfer cell at 100V for 1 hour at room temperature. The membrane was then used in immunodetection (Section 2.17.4).

2.16.4 Immuno-detection

The membrane was blocked overnight at 4°C with 5% Marvel in 0.05% PBST. The primary antibody (9A12) was diluted 1/1000 in 2% marvel in 0.05% PBST, poured over the membrane and incubated at room temperature for 3 hours. The membrane was then washed with PBST three times for 15 minutes each wash. The secondary antibody (HRP-conjugated goat anti-mouse IgG) was diluted 1/5000 and incubated on the membrane for 1 hour before repeating

the three washes. Enhanced chemiluminescence (ECL) detection was carried out using the ECL-Plus kit (Amersham Biosciences) according to manufacturer's instructions. The membrane was exposed to X-ray film for 5 minutes and the film developed using an SRX-201 developer (Konica Minolta).

2.17 Materials

2.17.1 General reagents

All standard laboratory reagents that are not otherwise stated below were from Sigma. Glycogen was from Roche. TRIzol was from Gibco.

2.17.2 Kits

Spin miniprep kits used in purifying DNA and PCR products were from Sigma. The ImProm Reverse Transcription kit and the Poly(A)+tract mRNA isolation kit were from Promega. The OneStep PCR kit was from Qiagen. The TURBO DNA-*free* kit was from Ambion. The BigDye Terminator sequencing kit was from Perkin Elmer Applied Biosystems. The ECL-Plus kit was from Amersham biosciences. The CleanSEQ clen up system was supplied by Agencourt Bioscience Corporation.

2.17.3 Bacterial Reagents

MacConkey agar, tryptone and yeast extract were from Difco Laboratories. All other bacteriological agents were purchased from BDH. Antibiotics were from sigma.

2.17.4 Enzymes and sequencing

Restriction enzymes and buffers are shown in Table 2.8. DNA polymerases and buffers are shown in Table 2.3. Rnase A was from sigma. PNK and CIP were from NEB.

Restriction Enzyme	Buffer	Company
*Eco*RI	Buffer H	Roche
*Age*I	Buffer K	Promega
*Sfu*I (*Asu*II)	Buffer H	Roche
*Pst*I	Buffer H	Roche
*Msp*I	NEBuffer 4	NEB
*Hpa*II	NEBuffer 1	NEB

Table 2.7 Restriction Enzymes

2.17.5 Nucleic Acids and Protein Standards

DNA markers were from Bioline (Hyperladders I and IV). Oligonucleotides were designed using Primer3 software (http://frodo.wi.mit.edu/primer3/) and ordered from Sigma or Invitrogen. The Precision Plus protein marker was from Biorad. dNTPs were purchased from Bioline.

2.17.6 Competent bacterial cells and vectors

E. coli DH5α cells were purchased from Invitrogen. The pGEM-T Easy Vector System was from Promega. pcDNA vectors from Invitrogen. The Clone Smart Blunt Cloning Kit was produced by Lucigen.

2.17.7 Cell culture

Media, trypsin EDTS, penicillin/streptomycin solution, L-Glutamine, PBS and Fetal Bovine Serum (FBS) were purchased from Lonza. bFGF was from Invitrogen. Dimethyl sulphoxide (DMSO), Dexamethasone, proteasome inhibitor and Poly-L-Lysine were from Sigma.

2.17.8 Electrophoresis, transfer membrane and film

Molecular grade agarose was purchased from Geneflow. Type VII low gelling agarose and ethidium bromide were from Sigma. Hybond-P membrane was from Amersham Biosciences. Formaldehyde loading buffer was from Ambion and X-ray film was from Kodak.

2.17.9 Solutions

2.17.9.i General Solutions

DNA Buffer	0.2M Tris-HCL (pH8.0), 0.1M EDTA
0.5X TBE	45mM Tris-borate, 1mM EDTA
1X MOPS	10mM EDTA, 0.2M MOPS, 50mM NaAc, 100mM NaOH
1X RIPA Buffer	150mM NaCl, 40mM Tris, 5mM EDTA, 1% NP40, 1% deoxycholate, 0.1% SDS
1X TAE	40mM Tris-acetate, 1mM EDTA
10X PBS	135mM NaCl, 2.5mM KCL, 10mM Na_2HPO_4, 1.8mM KH_2PO_4 pH7.4
20X SSC	3M NaCl, 0.3M trisodium citrate
TE	10mM Tris, 2mM EDTA

2.17.9.ii Miniprep solutions

Solution 1	50mM glucose, 25mM Tris-HCL pH8, 10mM EDTA pH8
Solution 2	0.2M NaOH, 1% SDS
Solution 3	5M KOAc

2.17.9.iii SDS-PAGE and Western Blotting

Running Buffer	25mM Tris, 2M glycine, 1% SDS
Resolving Buffer	0.1% SDS, 375mM Tris
Stacking Buffer	0.1% SDS, 250mM Tris
Transfer Buffer	25mM Tris, 2M glycine, 20% methanol
PBST	1X PBS, 0.1% Tween-20
Block Solution	5% Marvel in PBST

2.17.10 Media

2.17.10.i Cell Culture

Media and supplements used for cell culture are listed in Table 2.9.

Freeze media	5% DMSO in complete media (see Table 2.9)
Myoblast freeze media	10% DMSO in FBS
Differentiation Media 1	DMEM without L-Glutamine and sodium pyruvate, 2% Horse Serum, 2mM L-Glutamine, 100U/ml of penicillin/streptomycin solution
Differentiation Media 2	DMEM without L-Glutamine and sodium pyruvate, 15% Horse Serum, 2mM L-Glutamine, 100U/ml of penicillin/streptomycin solution

2.17.10.ii Bacteriological

LB (per litre)	10g tryptone, 5g yeast extract, 10gNaCl
MacConkey (per litre)	50g MacConkey agar
SOC (per litre)	20g tryptone, 5g yeast extract, 0.59g NaCl, 1.86g KCl, 10mM $MgSo_4$, 10mM $MgCl_2$, 20mM glucose
YT (per litre)	16g tryptone, 10g yeast extract, 5g NaCl

Cell Line	Growth Media	Supplements
C2C12	DMEM with 4.5g/L Glucose	10% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 4mM L-Glutamine
TE671	DMEM with 4.5g/L Glucose	10% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 4mM L-Glutamine
RD	DMEM with 4.5g/L Glucose	10% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 4mM L-Glutamine
RMS13	RPMI 1640	10% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 4mM L-Glutamine 10mM Hepes
Myoblasts	Hams F-10 with L-Glutamine	20% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 10ng/ml bFGF 1µM Dexamethasone
hEC	DMEM with 4.5g/L Glucose	10% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 4mM L-Glutamine
GCT27	DMEM with 4.5g/L Glucose	10% fetal bovine serum 120U/ml penicillin 120µg/ml streptomycin 4mM L-Glutamine

Table 2.8 Growth media requirements for cell lines

Chapter 3. Does DUX4 regulate PITX1?

3.1 Introduction and aims

The use of immortalised cell lines in the study of human diseases is widespread. These uniform populations of proliferative cells are easy to manipulate and tend to have accelerated growth rates and a reduction in growth requirements in comparison with primary tissues. The cell lines allow investigation of the function of the cells and the factors affecting growth rate and differentiation. They can be used to observe the cellular effects of gene manipulation or of expressing a protein of interest. While primary cell lines have a limited lifespan, in disease research they crucially allow a comparison of disease and control cells.

In addition to myoblast lines (see chapter four), a number of other commonly used cell lines have been utilised in FSHD research. Transfection of D4Z4 constructs into the mouse myoblast cell line C2C12 was used to show the localisation of the DUX4 protein to the nucleus (Ostlund *et al.*, 2005), investigate proposed promoter or enhancer regions (Kawamura-Saito *et al.*, 2006; Petrov *et al.*, 2008), and to observe the effect of D4Z4 copy number on expression of reporter genes (Yip and Picketts, 2003; Ottaviani *et al.*, 2009). Cells ectopically expressing high levels of DUX4 display characteristic markers of apoptosis (Gabriels *et al.*, 1999; Kowaljow *et al.*, 2007; Wallace *et al.*, 2011). As most overexpression studies used C2C12 or the human rhabdomyosarcoma TE671 cell lines, these were adopted for most of the work described in this chapter.

In 2007, based on microarray experiments, Dixit *et al.* reported that the paired-like homeodomain transcription factor 1 (*PITX1*) is specifically upregulated in the muscles of patients with FSHD and proposed that DUX4 was an upstream regulator of this gene (Dixit *et al.*, 2007). *Pitx1* is expressed predominantly in the developing hindlimb in mice (GibsonBrown *et al.*, 1996; Lanctot *et al.*, 1997; DeLaurier *et al.*, 2006) and has been shown to have a role in limb patterning and growth in both the chick (Logan and Tabin, 1999) and mouse (DeLaurier

et al., 2006). In one family, a *PITX1* mutation has been shown to segregate with clubfoot over three generations (Gurnett *et al.*, 2008) and 9% of *Pitx1*$^{+/-}$ mice show a clubfoot-like phenotype (Alvarado *et al.*, 2011). Misexpression of *Pitx1* in the developing chick wing bud changes the morphology such that it resembles a leg (Logan and Tabin, 1999), whilst misexpression of *Pitx1* in the mouse causes the forelimb to assume some structures characteristic of the hindlimb (DeLaurier *et al.*, 2006). Loss of *Pitx1* expression in the mouse causes the hindlimb to assume the morphology of the forelimb (Lanctot *et al.*, 1999). Alterations in the pattern of *Pitx1* expression have been proposed to underlie the evolutionary adaptation of vertebrate hindlimb structures (Shapiro *et al.*, 2004; Shapiro *et al.*, 2006).

By sequence analysis, Dixit *et al.* identified a potential DUX4 binding site in the promoter region of *PITX1* and cloned a 369bp fragment containing this site into a luciferase reporter construct. Co-transfection of C2C12 cells with the reporter construct and a DUX4 expression vector resulted in a 7.4 fold increase in luciferase activity after 24 hours. A mutation of the TAAT sequence in the putative DUX4 binding site reduced the induction of luciferase activity 4-fold. Electrophoretic mobility shift assays (EMSA) of nuclear extracts from the transfected C2C12 cells showed a supershift on incubation with an anti-DUX4 antibody (9A12). However, no chromatin immunoprecipitation (ChIP) experiments were performed, so the *in* vivo binding of DUX4 to the *PITX1* promoter has not been validated.

Expression of PITX1 was observed by co-immunofluorescence staining in C2C12 cells transiently transfected with constructs encoding DUX4. DUX4 was detected using the 9A12 antibody, and PITX1 was detected with a rabbit serum raised against a Pitx1 specific peptide. Both DUX4 and PITX1 proteins were detected in the same nuclei after 24 hours, however, expression was not confirmed by RT-PCR or western blotting.

This publication therefore identified *PITX1* as a possible target for DUX4 and the discovery provided an avenue to study its targets and in particular, to establish whether the mouse Dux protein has the same function as the human DUX4. The aims of this part of the study were to confirm the finding that overexpression of DUX4 results in activation of the endogenous *PITX1* gene and to investigate whether the mouse *Dux* gene can also induce *PITX1*.

3.2 Results

3.2.1 Establishing conditions for the efficient transfection of DUX4

In initial transfection experiments, C2C12 and TE671 cells grown on coverslips were transfected with DUX4 constructs made previously in this laboratory (Figures 2.1 and 2.2) using Effectene. Immunocytochemistry was performed using anti-V5 antibodies and the transfection efficiency was found to be only 1-2% per coverslip. To try and improve transfection of the cell lines a number of different reagents were trialled.

Firstly, different seeding densities were tested so that the culture would be 50-80% confluent at the start of the transfection and near to confluency by the end. In case the miniprep reagents were affecting the transfection efficiency, constructs were purified with either the Sigma or the Qiagen kits (with and without lyse blue) and transfection efficiencies compared using Effectene. For this set of experiments, D4Z4-GFP constructs (Figures 2.3 and 2.4 and Table 3.2) that had been used by previous members of the group and found to transfect efficiently were used. In addition, the use of the GFP tag meant that transfection efficiency could be analysed without the need for immunocytochemistry. Four different reagents were trialled and cells were incubated for 24 or 48 hours. Transfection efficiency was calculated by number of cells expressing GFP/total number of nuclei. For each coverslip, cell counts were taken from five separate fields with a good spread of nuclei, the average transfection efficiency is shown in Table 3.2. Although the FuGene reagent gave excellent results for the TE671 cells, Effectene gave consistently good transfection efficiencies for both cell lines and was used for subsequent transfections. DNA that had been purified using the Sigma columns gave the best results and this kit was used for all subsequent high-purity minipreps.

Typically, a reduced transfection efficiency of 5-25% was seen when the cell lines were transfected with FlmD4Z4 V5 or FlhD4Z4 V5 (Table 3.1) and stained with anti-V5 antibodies (Figure 3.1). C2C12 cells had transfection efficiencies at the lower end of the range while TE671 cells typically had a higher number of nuclei showing expression.

Construct Name	Vector	Insert
GFP NtermD4Z4	pEGFP C3	Mouse Dux N-terminal region (homeodomains)
FlmD4Z4 GFP	pEGFP N1	Full length mouse Dux ORF
FlhD4Z4 V5	pcDNA 3.1 B	Full length human DUX4 ORF
FlmD4Z4 V5	pcDNA 3.1 A	Full length mouse Dux ORF

Table 3.1 Expression constructs used in this chapter

Constructs were created in Jane Hewitt's laboratory by Jannine Clapp and are shown schematically in Figures 2.1 to 2.4.

Cell Type	Reagent	Details	Time	Efficiency
C2C12	Effectene	GFP NtermD4Z4	24h	25%
	Effectene	pEGFP-N1 purified with Qiagen kit	24h	15%
	Effectene	pEGFP-N1 purified with Sigma kit	24h	25%
	Effectene	pEGFP-N1 purified with Qiagen kit without lyse blue	24h	20%
	FuGene	FlmD4Z4 GFP (3:2 ratio of FuGene:DNA)	24h	0%
	FuGene	FlmD4Z4 GFP (5:2 ratio of FuGene:DNA)	24h	0%
	FuGene	FlmD4Z4 GFP (7:2 ratio of FuGene:DNA)	24h	0%
	GeneJuice	GFP NtermD4Z4	24h	2%
	Transpass	FlmD4Z4 GFP (10µl Transpass)	24h	10%
	Transpass	FlmD4Z4 GFP (12µl Transpass)	24h	45%
	Transpass	FlmD4Z4 GFP (10µl Transpass)	48h	0%
	FuGene	FlmD4Z4 GFP (5:2 ratio of FuGene:DNA)	48h	0%
	GeneJuice	GFP NtermD4Z4	48h	5%
	Effectene	GFP NtermD4Z4	48h	10%
TE671	Effectene	GFP NtermD4Z4	24h	45%
	Effectene	pEGFP-N1 purified with Qiagen kit	24h	40%
	Effectene	pEGFP-N1 purified with Sigma kit	24h	50%
	Effectene	pEGFP-N1 purified with Qiagen kit without lyse blue	24h	45%
	FuGene	FlmD4Z4 GFP (3:2 ratio of FuGene:DNA)	24h	0%
	FuGene	FlmD4Z4 GFP (5:2 ratio of FuGene:DNA)	24h	0%
	FuGene	FlmD4Z4 GFP (7:2 ratio of FuGene:DNA)	24h	70%
	GeneJuice	GFP NtermD4Z4	24h	45%
	GeneJuice	GFP NtermD4Z4	48h	55%
	FuGene	FlmD4Z4 GFP (5:2 ratio of FuGene:DNA)	48h	2%
	Effectene	GFP NtermD4Z4	48h	25%

Table 3.2. Transfection efficiency of C2C12 and TE671 cells using different transfection reagents.

Transfection efficiency was calculated by number of cells expressing GFP/total number of nuclei. For each coverslip, cell counts were taken from five separate fields with a good spread of nuclei. The table shows an average of the five fields.

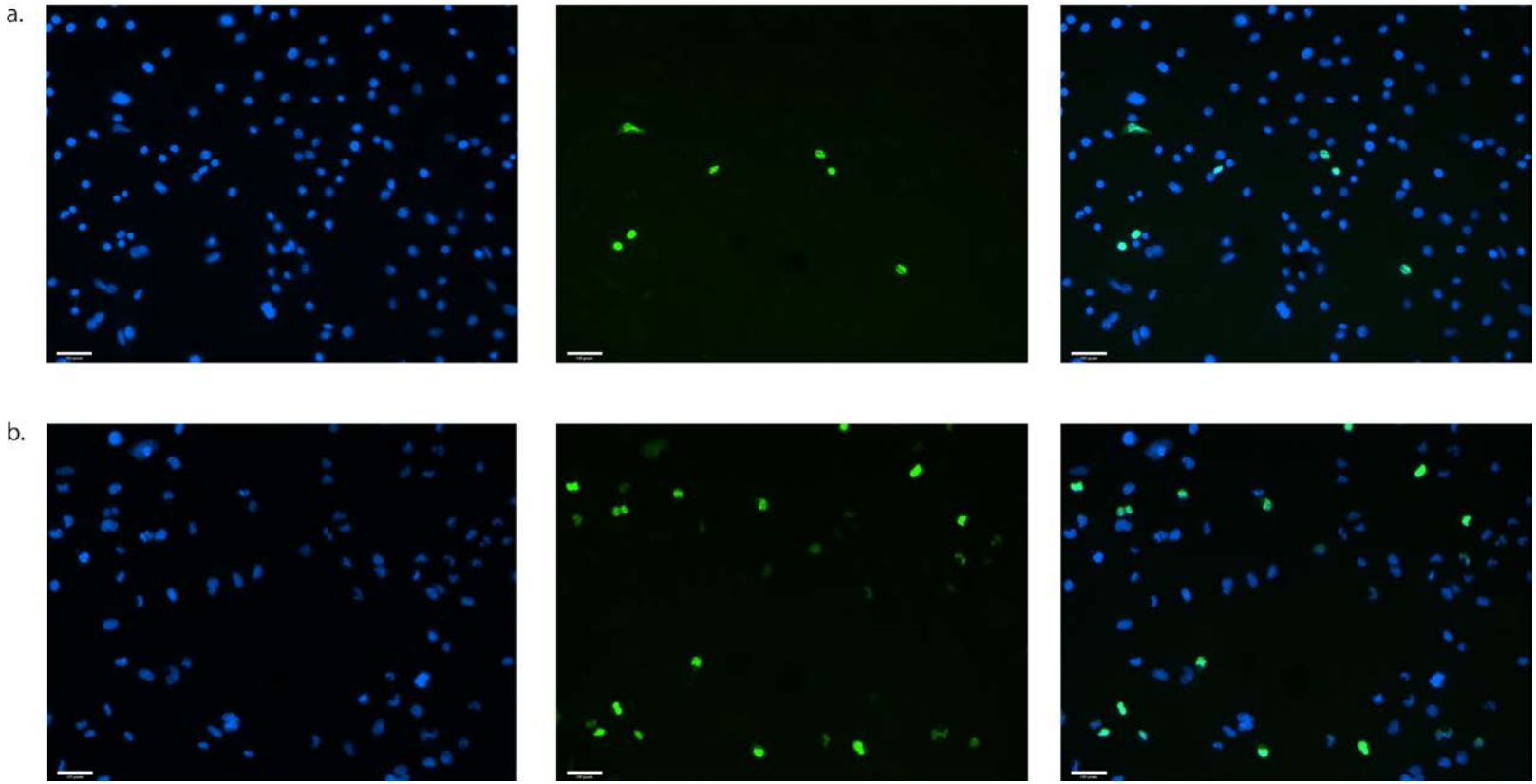


Figure 3.1 Expression of D4Z4 constructs

Expression of full length mD4Z4 in C2C12 cells (a) and hD4Z4 in TE671 cells (b). Transfections were carried out with Effectene reagent and immunocytochemistry performed after 24 hours. Nuclei are stained with DAPI. Expression of D4Z4 constructs was detected with an antibody to the V5 tag (see antibody table) and is shown in green.

3.2.2 Does DUX4 regulate PITX1 expression?

To confirm the finding that overexpression of DUX4 results in activation of the endogenous *PITX1* gene, cell lines were transiently transfected with constructs encoding DUX4 and the expression of PITX1 was then examined by immunostaining. In order to establish whether the mouse Dux protein has the same function, similar experiments were performed using expression constructs for mouse Dux.

3.2.2.i Generation of PITX1 expression constructs

Before testing whether DUX4 overexpression induces PITX1, PITX1 expression constructs were generated to confirm the specificity and sensitivity of a PITX1 antibody obtained from Dr Yi-Wen Chen, (Children's Hospital, Boston USA). Constructs encoding V5-tagged human DUX4 and mouse Dux proteins had been generated previously (Clapp *et al.*, 2007). The myc epitope was chosen to tag PITX1 so that co-expression of both proteins could be visualised by immunocytochemistry.

In order to produce a myc tagged expression vector, a myc tag was engineered into pcDNA3.1. The myc epitope was first amplified by creating primer dimers with primers mycSfuF and MycAgeR (Figure 3.2a), which were engineered to contain *Sfu*I and *Age*I restriction sites respectively. After digestion with *Sfu*I and *Age*I, this myc tag was then cloned in-frame into the vectors pcDNA3.1/V5 His A, B and C, replacing their V5 tags (Figure 3.2b).

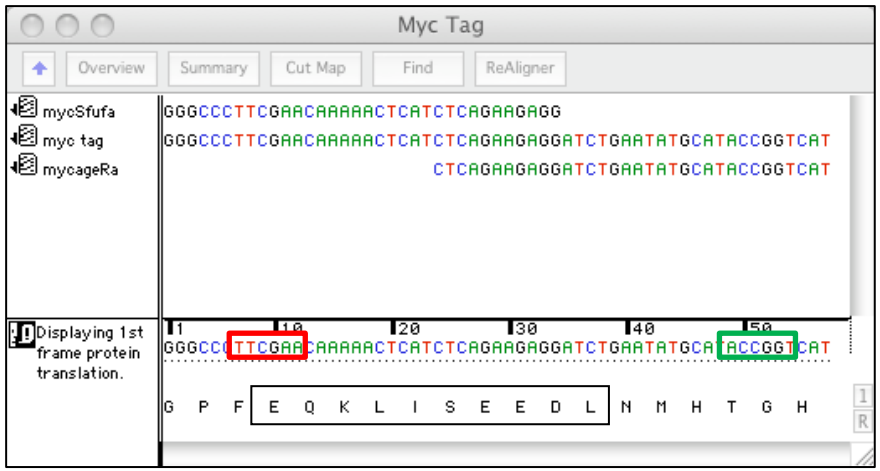
Primers for mouse and human PITX1 were engineered to contain restriction sites for *Eco*RI and *Sfu*I (Tables 2.4 and 2.5). The coding region for mouse *Pitx1* was amplified by RT-PCR from Swiss Webster mouse midterm embryo RNA with primers MPITX1F and MPITX1RSfuI, using the Improm reverse-transcription kit and KOD polymerase. The human *PITX1* coding region was similarly amplified from testis RNA, although Primestar polymerase was used instead of KOD. The RT-PCR products were gel purified, digested with *Eco*RI and *Sfu*I, and cloned into the myc tagged vector pcDNA3.1/myc HisA.

The mouse *PITX1* sequence contains an interrupted polyA tract ($A_2GA_3GA_9GA_{12}GA$) between position 743 and 786; the majority of the clones (9 of the 11) were found to contain deletions within this tract. However, two clones contained the correct sequence and one of these (pcDNA3.1/mPITX1) was used in subsequent experiments.

Despite using a proofreading polymerase, 5 of the 7 human clones had a number of consistent changes (86 nucleotide positions in 947bp) compared to the reference sequence in GenBank (NM_002653.4), it is possible that these represent amplification of a pseudogene. No identical matches to this sequence were found in the human genome using BLASTN so these variant clones are likely to come from an unsequenced part of the genome. However, two of the human *PITX1* clones conformed to the GenBank sequence (NM_002653.4) apart from a G-C substitution at position 896, changing an glycine to an alanine (Figure 3.3a). A BLASTN search using the *PITX1* reference sequence against the human EST trace archives identified other sequences with the same base change, suggesting that it is likely to be a polymorphism (Figure 3.3b). One of these clones (pcDNA3.1/hPITX1) was used in subsequent experiments.

Diagrams of both the mouse and human PITX1 constructs are shown in Figure 3.4.

a.



b.

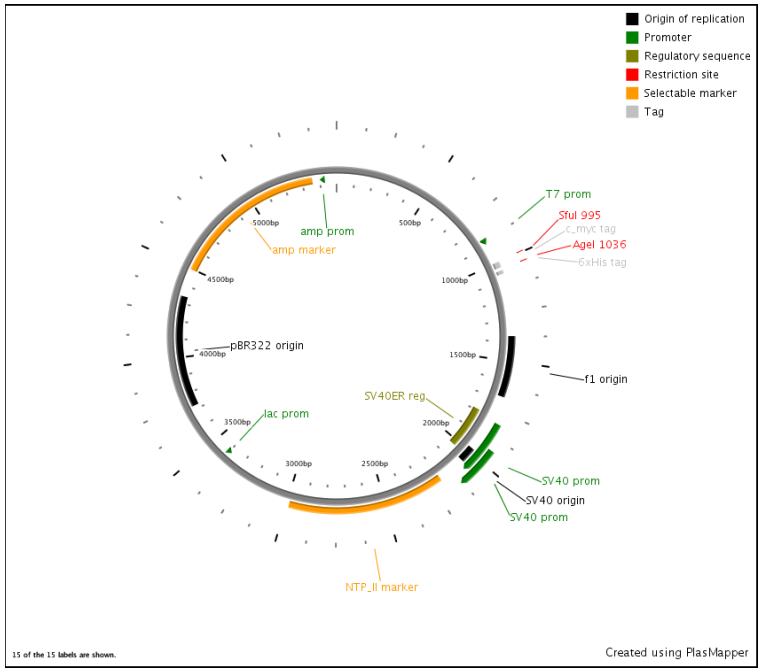
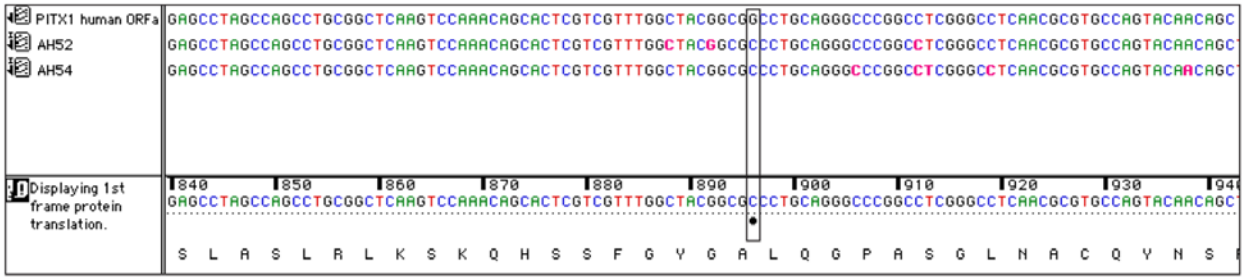


Figure 3.2 Engineering a Myc tagged expression vector

a) A screen shot of the Sequencher alignment of the mycSfuF and mycAgeR primers used to produce the 57bp Myc insert. The amino acid sequence is shown in the bottom frame and the box indicates the sequence of the Myc epitope. The *Sfu*I site is boxed in red and the *Age*I site is boxed in green. b) Map of the pcDNA3.1/myc His construct. Myc and His epitopes are shown in grey. The positions of the restriction sites used for removing the V5 tag and replacing it with the Myc insert are shown in red. Selective markers, promoters and tags are identified.

a.



b.

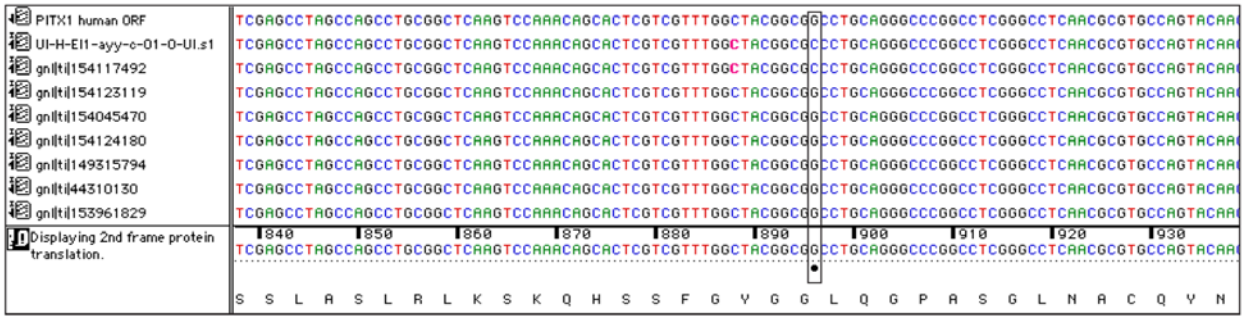
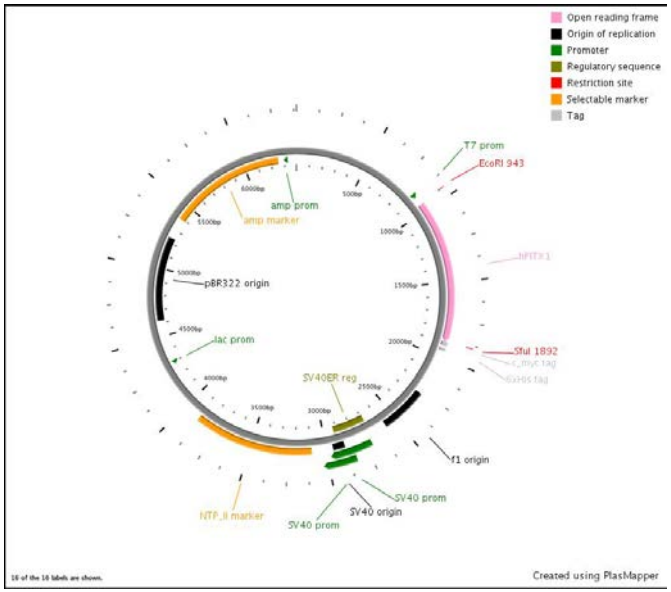


Figure 3.3 Human PITX1 sequence variants

a) A screenshot of the Sequencher alignment of the two correct hPITX1 sequences with the Genbank sequence NM_002653.4. The Sequencher programme indicates base ambiguities with a black dot underneath the alignment. The amino acid sequence is shown in the bottom frame. The position of the G-C base variation is highlighted at position 896. b) Alignment of NM_002653.4 with ESTs from the trace archives to show additional sequences with the G-C base change (highlighted).

a.



b.

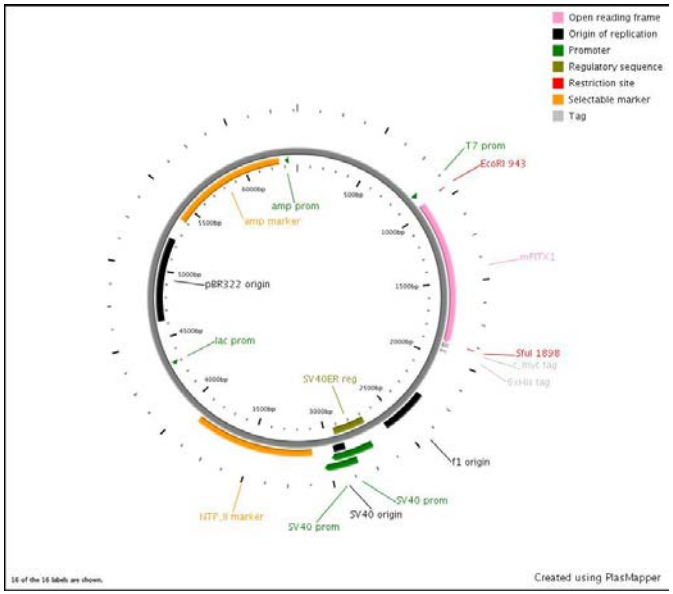


Figure 3.4 Human (a) and mouse (b) PITX1 myc constructs.

Map of the pcDNA3.1/myc HisA construct, engineered to contain human (a) or mouse (b) *PITX1* coding sequences. Myc and His epitopes are shown in grey. PITX1 inserts are highlighted in pink. The positions of the restriction enzymes EcoRI and SfuI used for cloning are indicated in red. Selective markers, promoters and tags are identified.

3.2.2.ii Validation of the PITX1 antibody

The PITX1.N2 polyclonal antibody (which apparently cross-hybridises to the mouse Pitx1 protein, Y-W Chen, personal communication) was raised against the N-terminal region of the human *PITX1* sequence. Alignment of the mouse and human PITX1 peptides is shown in Figure 3.5. The two proteins are highly conserved with 96% amino acid identity. The antibody has been previously described to identify endogenous PITX1 in C2C12 cells (Dixit *et al.*, 2007). To validate the antibody, C2C12 myoblast or TE671 rhabdomyosarcoma cells were transfected with the PITX1 constructs, and then immunocytochemistry was performed after 24 hours.

Dixit et al (2007) identified a PEST motif in PITX1. This motif is known to target proteins for degradation so they performed their transfection experiments in the presence of a proteosome inhibitor because if PITX1 is being degraded rapidly in C2C12 or TE671 cells expression may not be seen (Dixit *et al.*, 2007). Therefore, in the experiments described here, 24h after transfection the synthetic proteosome inhibitor, Z-Leu-Leu-Leu-al, was added to some of the transfections at a concentration of 500µg/ml. The cells were incubated for a further 5 hours before immunocytochemistry was performed.

Following transfection with hPITX1 constructs, hPITX1 expression could be seen in 1% cells using the anti-myc antibody after 48h, no expression could be seen after 24h. After addition of Z-Leu-Leu-Leu-al, a similar proportion of transfected cells could be seen after 24h. Co-staining with the anti-PITX1 polyclonal antibody and a monoclonal antibody against myc confirmed that the PITX1.N2 antibody is able to detect the human PITX1 protein (Figure 3.6).

Following transfection of the mPitx1 construct, the protein was identified in <1% of cells with the anti-myc antibody (Figure 3.6b). However, no positive nuclei were ever seen using the PITX1.N2 antibody. This suggests that, in contrast to the information supplied, PITX1.N2 does not cross react with mouse Pitx1, at least in immunocytochemistry. Therefore, it was not possible to investigate the effect expression of mouse Dux has on the endogenous mouse *Pitx1* gene. However, it was possible to test whether the mDux gene could activate *PITX1* expression in human cells.

```
Human    MDAFKGGMSLERLPEGLRPPPPPPHDMGPAFHLARPADPREPLENSASESSDTELPEKER 60
Mouse    MDAFKGGMSLERLPEGLRPPPPPPHDMGPSFHLARAADPREPLENSASESSDADLPDKER 60
         ******************************:*****.***************::**:***

Human    GGEPKGPEDSGAGGTGCGG-ADDPAKKKKQRRQRTHFTSQQLQELEATFQRNRYPDMSMR 119
Mouse    GGEAKGPEDGGAGSAGCGGGAEDPAKKKKQRRQRTHFTSQQLQELEATFQRNRYPDMSMR 120
         ***.*****.***.:**** *:************************************

Human    EEIAVWTNLTEPRVRVWFKNRRAKWRKRERNQQLDLCKGGYVPQFSGLVQPYEDVYAAGY 179
Mouse    EEIAVWTNLTEPRVRVWFKNRRAKWRKRERNQQLDLCKGGYVPQFSGLVQPYEDVYAAGY 180
         ************************************************************

Human    SYNNWAAKSLAPAPLSTKSFTFFNSMSPLSSQSMFSAPSSISSMTMPSSMGPGAVPGMPN 239
Mouse    SYNNWAAKSLAPAPLSTKSFTFFNSMSPLSSQSMFSAPSSISSMTMPSSMGPGAVPGMPN 240
         ************************************************************

Human    SGLNNINNLTGSSLNSAMSPGACPYGTPASPYSVYRDTCNSSLASLRLKSKQHSSFGYGG 299
Mouse    SGLNNINNLTGSSLNSAMSPGACPYGTPASPYSVYRDTCNSSLASLRLKSKQHSSFGYGG 300
         ************************************************************

Human    LQGPASGLNACQYNS 314
Mouse    LQGPASGLNACQYNS 315
         ***************
```

Figure 3.5 Alignment of human and mouse PITX1 amino acid sequences

Identical amino acids are highlighted with an asterisk. Within the first 82 amino acids there are 10 differences between the two sequences, ":" indicate conserved substitutions, and "." indicate semi-conserved substitutions. There is a single base pair insertion/deletion at position 80. The peptides are identical from position 83 onwards.

3.2.2.iii Does transfection with DUX4 constructs upregulate PITX1 expression?

While transfection with the human PITX1 construct was successful at the first attempt, transfection with the mouse PITX1 construct was initially unsuccessful. As such, co-immunofluorescence staining for PITX1 and DUX4 was initially performed on both C2C12 and TE671 cell lines after transfection with the FlmD4Z4 V5 (containing the full mouse *Dux* ORF) or FlhD4Z4 V5 (containing the full human *DUX4* ORF) constructs (Figures 2.1 and 2.2). Twenty-four hours after transfection the synthetic proteosome inhibitor, Z-Leu-Leu-Leu-al, was added to the transfections. The cells were incubated for a further 5 hours before immunocytochemistry was performed. Once it was established that the PITX.N1 antibody did not detect the mouse PITX1 protein, co –immunofluorescence experiments were restricted to TE671 cells.

Expression of the DUX constructs (FlmD4Z4 V5 and FlhD4Z4 V5) was seen in approximately 5% C2C12 myoblast cells and 25% TE671 cells (Figure 3.1). Co-immunofluorescence staining for PITX1 and hDUX4 was performed four times using TE671 cells and once in C2C12 cells. Cells transfected with the human PITX1 construct were used as a positive control for the PITX1.N2 antibody. Co-immunofluorescence staining for PITX1 and mDUX4 was performed three times using C2C12 cells and twice with TE671 cells. However, no expression of PITX1 was identified (Figure 3.7).

Thus, these experiments did not support the previous study by Dixit *et al.* proposing that overexpression of DUX4 results in expression of endogenous PITX1. In addition, no evidence that the PITX1.N1 antibody recognises the mouse protein was seen. Therefore, it was decided not to pursue this part of the project further.

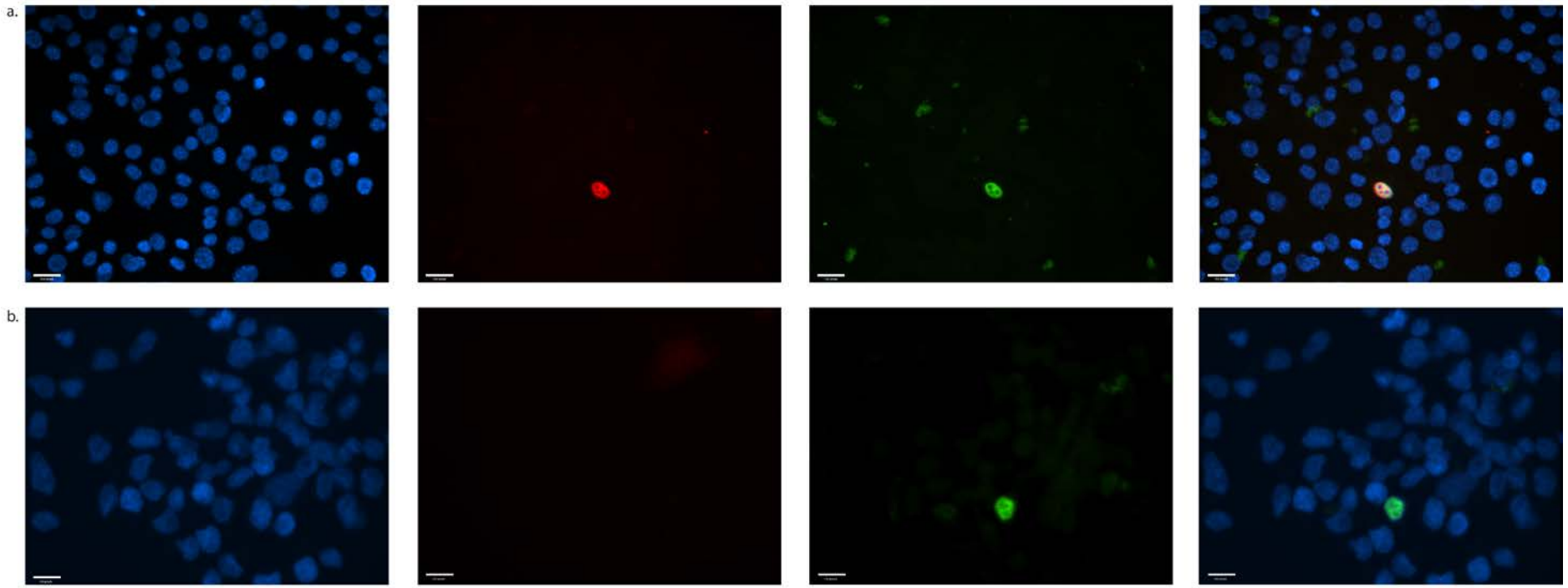


Figure 3.6 Validation of the PITX1 antibody

C2C12 cells 24h after transfection with a) Human PITX1 constructs and b) mouse PITX1 constructs. Cells were incubated for 5h with 500µg/ml proteasome inhibitor before analysis. Immunocytochemistry was performed with anti-PITX1 at a 1:1000 dilution (red) and anti-Myc at a 1:800 dilution (green) followed by appropriate secondary antibodies (see Table 2.1). Nuclei are stained with DAPI (blue).

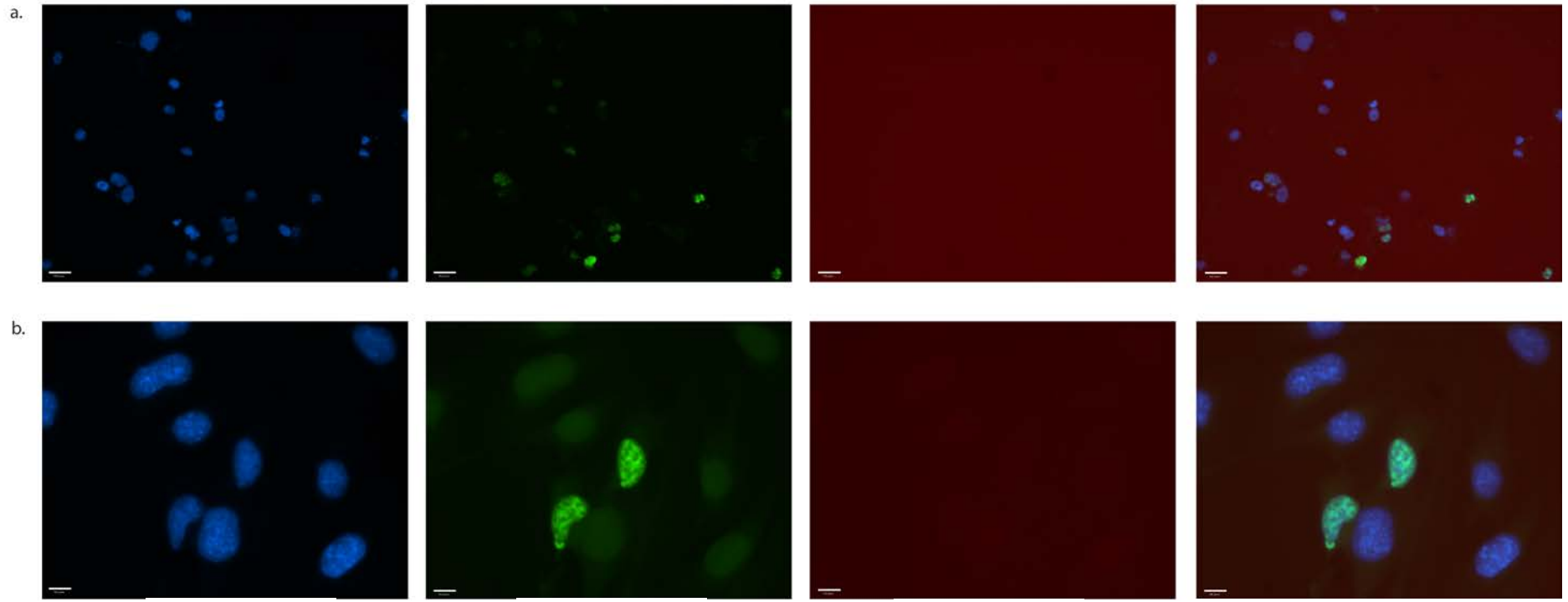


Figure 3.7 Expression of human or mouse DUX4 did not induce PITX1 expression

TE671 cell 24h after transfection with a) FlhD4Z4 V5 and b) FlmD4Z4. Cells were incubated for 5h with 500µg/ml proteasome inhibitor before analysis. Immunocytochemistry was performed with anti-PITX1 at a 1:1000 dilution (red) and anti-V5 at a 1:200 dilution (green) followed by appropriate secondary antibodies (see Table 2.1). Nuclei are stained with DAPI (blue).

3.3 Discussion

In 2007 Dixit *et al.* reported that PITX1 is specifically upregulated in the muscle of FSHD patients and proposed that DUX4 was an upstream regulator of this gene (Dixit, et al., 2007). In order to confirm this observation, and to see whether it held true for the mouse Dux gene, C2C12 and TE671 cells were transfected with the mD4Z4 or hD4Z4 constructs and the expression of PITX1 detected by immunocytochemistry.

Initially the PITX1 antibody obtained from Dr Yi-Wen Chen was tested on PITX1 expression vectors. The antibody was able to detect human PITX1 protein but not the mouse. In agreement with Dixit *et al.,* PITX1 expression could be detected earlier if a proteasome inhibitor was used.

The data suggesting that the PITX1 antibody did not identify the mouse PITX1 protein was unexpected, as it had previously been described to identify endogenous PITX1 in C2C12 cells (Dixit *et al.*, 2007). The peptide sequences of the two PITX1 proteins show 96% identity, however there are 10 variant positions. If a peptide including variant amino acids was used to raise the antibody, it is possible the antibody would not recognise both the human and mouse proteins. Dixit *et al.* did not provide any information about how they tested the antibody on the mouse PITX1 protein, more information on the antibody would be required before any further experiments were performed.

In co-immunofluorescence experiments, no PITX1 expression could be identified after transfection with mD4Z4 or hD4Z4 vectors. Despite the relatively high proportion of cells transfected with DUX4 (25% in TE671 cells), no PITX1 expression was identified in cells transfected with DUX4. This is in contrast to data published by Dixit *et al.* (2007) whose DUX4 construct activated expression of the endogenous PITX1 in C2C12 cells. The difference between these data may be as a result of variation in levels of expression of DUX4 between the two studies, in order to compare expression levels, quantitative studies would need to be performed.

It was not possible to make any conclusions about the effect expression of the mouse Dux gene has on endogenous mouse PITX1, as the PITX1 antibody supplied by Dr Yi-Wen Chen was

not shown to bind to the protein produced from the mouse PITX1 expression vector. Despite endogenous PITX1 being identified from C2C12 cells by Dixit and colleagues, no endogenous PITX1 expression was identified in this study so this line of work was no longer pursued.

In support of the data provided in this chapter, Bosnakovski *et al.* used an inducible cassette exchange in C2C12 cells to express DUX4 and studied the effect on gene expression. Despite identifying 1011 differentially expressed genes, no changes in PITX1 expression were seen. In addition, Tsumagari *et al.* have recently used exon-based microarrays to determine the expression profiles of myoblasts and myotubes derived from FSHD patients and control individuals. Differential expression of 295 genes in FSHD myoblasts and 797 genes in FSHD myotubes was determined compared with control cells, however no change in PITX1 expression was observed (Tsumagari *et al.*, 2011).

Chapter 4. Analysis of human DUX4 expression

4.1 Introduction

The FSHD phenotype is mostly restricted to the skeletal muscle, although some non-muscle tissues are involved in a subset of patients (Section 1.1). Muscle cells are therefore an obvious choice for study when looking for differences between FSHD patients and control individuals. The isolation of myoblasts from muscle biopsies is well established, however, this results in a heterogeneous mixture of myoblast and fibroblast cells. These contaminating fibroblasts may have an influence on the behaviour of the myoblasts so the proportion of each should be determined before studies can be accurately compared. The reported appearance of cells derived from muscle biopsies of FSHD patients varies between different studies. Winoker *et al.* (2003a) observed FSHD cells with swollen cytoplasm and large vacuoles and a higher percentage of necrotic cells than in control muscle samples. However, other groups found no morphological differences in FSHD myoblasts compared with controls and observed no necrotic cells (Morosetti *et al.*, 2007; Barro *et al.*, 2010). This variation in phenotype between FSHD samples suggests that there is likely to also be natural variation in expression levels of genes. Some variation in expression levels of genes between different muscle types has also been observed (Winokur *et al.*, 2003b; Dixit *et al.*, 2007; Ansseau *et al.*, 2009). Thus, the origin of the samples must also be taken into account when comparing studies using different patient cohorts.

In FSHD studies, myoblast cell lines from patients and controls have been used to search for differences in expression levels of genes between affected and unaffected individuals (Rijkers *et al.*, 2004; Alexiadis *et al.*, 2007; Ansseau *et al.*, 2009; Bodega *et al.*, 2009; Masny *et al.*, 2010), and also between affected and unaffected tissues in an individual patient (Winokur *et al.*, 2003b; Laoudj-Chenivesse *et al.*, 2005; Dixit *et al.*, 2007; Ansseau *et al.*, 2009). The differentiation of myoblasts into myotubes is well understood and this has enabled

observation of differences in how the cells behave upon differentiation into myotubes (Rijkers *et al.*, 2004; Bodega *et al.*, 2009; Masny *et al.*, 2010). In order to investigate the expression of DUX4 in FSHD, myoblast cell lines from three patients were used in this study.

At the start of the work described in this chapter transcription from the D4Z4 array in FSHD myoblast cell cultures had been published by two groups (Dixit *et al.*, 2007; Kowaljow *et al.*, 2007) and previous work in this laboratory by Dr Jannine Clapp had showed by RT-PCR that DUX4 is transcribed in human testis and fetal skeletal muscle (Figure 4.1a).

Using RT-PCR, Kowaljow *et al*. (2007) amplified a 1477bp fragment from RNA of proliferating and differentiating myoblast cell lines from five FSHD patients, which was absent in three controls. They were unable to identify any DUX4 protein in these cell lines. They did, however, report endogenous expression of a 50kDa protein in RD and RMS13 rhabdomyosarcoma cell lines using a DUX4 antibody on western blot.

Dixit *et al*. (2007) used Rapid Amplification of cDNA ends (RACE) to characterise the ends of the *DUX4* mRNAs from C2C12 cells transfected with genomic D4Z4 repeats and RT-PCR products from primary patient myoblast cells. The 3' RACE identified a 136bp intron located between the stop codon and the end of the D4Z4 unit, and a second 345bp intron located in the pLAM region. Whilst the second intron was removed from all transcripts, they reported alternative splicing of the first intron, resulting in two variant transcripts (Figure 1.7). A reverse primer 109bp from the end of the second intron was used in RT-PCR to amplify a 1700bp fragment from FSHD myoblasts but not control myoblasts (Dixit *et al.*, 2007). The group also identified a non-canonical polyA signal ATTAAA in the pLAM region, at position 8046 of the GenBank sequence FJ439133. Both Dixit *et al.* and Kowaljow *et al.* reported that all the transcripts they amplified showed 100% sequence identity, suggesting that only one of the repeats is expressed. However, they did not publish sequence data to confirm this.

Initially, the aim of the work described in this chapter was to test whether *DUX4* mRNA was expressed in normal tissues or cells and confirm expression in FSHD myoblast cell lines by RT-PCR. Any transcripts identified would be sequenced and compared with genomic sequence data from other projects in the laboratory aiming to map sequence variants within D4Z4

arrays. If possible, such comparison of variants might show whether FSHD transcripts originated from the deleted allele.

While this work was being carried out further publications reported expression from D4Z4. Although Snider *et al*. had difficulty in amplifying the full length transcripts that had been described previously, they were able to identify smaller fragments from the 3' and 5' ends in both FSHD and control myoblasts (Snider *et al.*, 2009). Using strand-specific RT-PCR they amplified both sense and antisense transcripts, in agreement with the unpublished data from Jannine Clapp. However, they were unable to amplify transcription from the central portion of the repeat (positions 2270-2398 of the first ORF in GenBank sequence FJ439133). In most cases their transcripts matched the last repeat, however there were some variant bases suggestive of transcription from internal D4Z4 repeats. By using 3'RACE on RNA isolated from FSHD myotubes they also confirmed the two splice forms reported previously (Dixit *et al.*, 2007).

As well as variant 3' ends, this group were also able to identify heterogeneous 5' ends. Re-sequencing of a phage clone (ƛ-42) representing a patient-derived, deleted *EcoRI* fragment containing two D4Z4 repeats identified a number of potential ORFs which are shown in Figure 4.1b. Using 5'RACE they were able to amplify transcripts originating from an MQG ORF (position 7050 to 7280 of GenBank sequence FJ439133) which falls within the previously identified MAL ORF (position 6006 to 7280 of GenBank sequence FJ439133) (see Figure 4.1). These 5' truncated products were uncapped, suggesting that they are processed from a larger transcript. Capped transcripts all originated from the standard MAL ORF. They also reported other splice variants which removed the MQG ORF, or which appeared to join internal repeats to the last repeat of the array. For a summary of these splice forms see Figure 4.1b.

Although amplifying full length transcripts had proved difficult in myoblasts, Snider and colleagues were able to amplify across the whole region in embryonic stem cell (ESC) and mesenchymal stem cell (MSC) lines. Therefore, analysis of a human embryonal carcinoma (hEC) cell line was subsequently included in the work described in this chapter.

In 2010, as the work described in this chapter was being completed, Snider *et al.* reported successful amplification of two transcripts (DUX4-fl and DUX4-s) from the final D4Z4 repeat in

muscle biopsies from FSHD patients and non-affected individuals (Snider *et al.*, 2010) using a nested RT-PCR approach (Figure 4.1c). The transcripts begin at the MAL ORF (position 6006 of FJ439133) and end immediately downstream of the poly A signal (position 8077 of FJ439133). The DUX4-fl transcript has one or both of the two introns identified by Dixit *et al.* spliced out, and was detected in 5/10 FSHD samples and none of the controls. The DUX4-s transcript uses a cryptic splice donor site 19bp downstream of the second homeobox, removing the distal end of the ORF (Figure 4.1c). DUX4-s was detected in all of the control samples and some of the FSHD samples. The group were also able to amplify the DUX4-fl fragment from human testis RNA (Snider *et al.*, 2010).

Thus, the first aim of the work in this chapter was to grow and characterise FSHD myoblast cell lines in order to determine if there was expression from the D4Z4 array. Expression of the DUX4 gene was then investigated by RT-PCR on myoblast cell lines, as well as the hEC cell line, N-TERA2, the rhabdomyosarcoma cell lines, RD and RMS13, and the germ cell tumour line GCT27. The final aim of this chapter was to compare sequence variants identified in any expressed transcripts with genomic sequence variants previously identified in the lab.

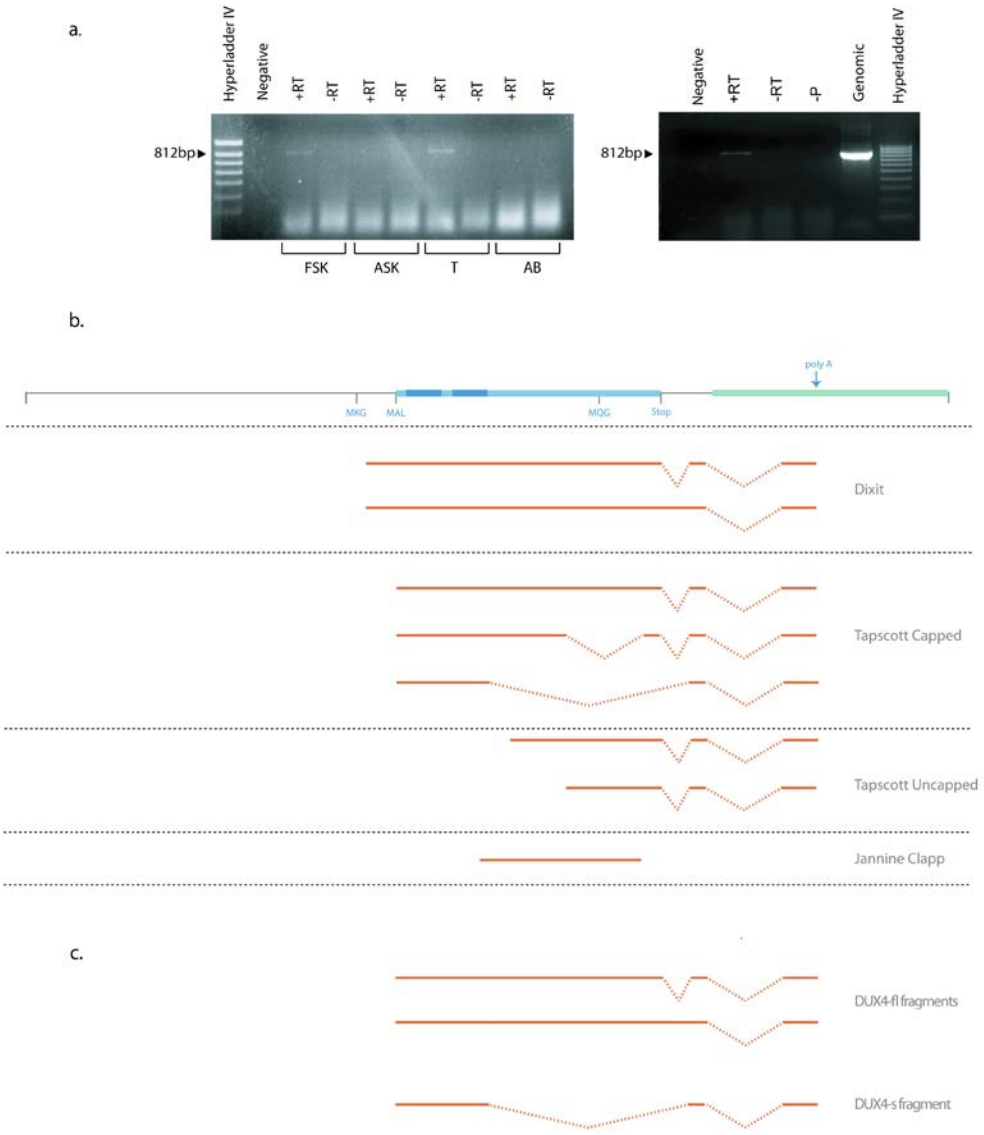


Figure 4.1 Expressed transcripts from D4Z4.

a) RT-PCR on fetal skeletal muscle (FSK), adult skeletal muscle (ASK), testis (T) and adult brain (AB) RNA (Ambion). Primers HDUX_F1 and HDUX_R1 amplify an 812bp fragment within the ORF (for sequences see table 2.5). Experiments performed by Dr Jannine Clapp. b) A schematic representation of the final repeat of the D4Z4 array with the transcripts identified by Dixit *et al.* (2007), Snider *et al*. (2009) and Jannine Clapp shown underneath. The DUX4 ORF is shown in light blue, with the homeodomains highlighted in dark blue. The pLAM region is shown in green and the polyA signal predicted by Dixit *et al.* is indicated. The start and stop sites of the MKG, MAL and MQG ORF's, identified by Snider *et al.* (2009), are shown underneath. For the transcripts, exons are represented by an orange line and introns are indicated with dotted lines. c) Representation of the transcripts referred to as DUX4-fl and DUX4-s in the Snider *et al.* (2010) publication.

4.2 Results

4.2.1 Maintenance of human cell lines

For this study a number of cell lines were grown in the lab. Three primary myoblast cell lines (GM17731, GM17869, GM17940) were obtained from the Corriell repository. These lines were from muscle biopsies of FSHD patients; GM17731 and GM17940 have three copies of the D4Z4 repeat unit on the disease allele whilst GM17869 has 5 repeat units (Table 4.1). For GM17869 and GM17940, corresponding lymphoblastoid cell lines were also obtained. These were used for haplotype and sequence analysis of the D4Z4 locus by an MRes student in the laboratory (see Section 4.2.2i).

The two rhabdomyosarcoma cell lines identified by Kowaljow and colleagues were also obtained from the ATCC. RD is a fetal rhabdomyosarcoma cell line that has been shown to be at least parental, if not identical, to TE671 cells (Stratton *et al.*, 1989). The RMS13 cell line was established from the bone marrow of a child with rhabdomyosarcoma (http://www.lgcstandards-atcc.org); these cells show ultrastructural elements of primitive skeletal muscle differentiation.

4.2.2 Analysis of myoblast cell lines

The myoblast cell lines were grown as detailed in Section 2.1. For verification of the GM17731, GM17940 and GM19869 cell lines, RT-PCR was used to confirm expression of the myoblast-specific marker desmin (Figure 4.2a). Myoblast cultures are a mixture of myoblasts and fibroblasts, and although myoblasts predominate on recovery from liquid nitrogen the fibroblast cells take over with increasing passage. In order to reduce the number of fibroblasts present, all myoblasts used for the analyses described in this thesis were limited to 5 passages. To determine the proportion of myoblasts in the cultures, cells were seeded onto coverslips and immunocytochemistry performed using antibodies to the myoblast-specific marker desmin. At these passage numbers typically 100% of the cells were stained with desmin (Figure 4.2b).

Myoblasts were differentiated in flasks coated with matrigel as described in Section 2.1.4. On reduction of serum in the media, myoblast cells fuse together to form myotubes. For verification that these cells have differentiated, immunocytochemistry was performed using the myotube-specific marker MF-20 (Figure 4.2c) and RT-PCR confirmed expression of the myosin heavy chain gene, which is specific to differentiating muscle cells (Figure 4.2d).

The fusion index of the myoblasts was calculated 6 days after differentiation. The index was calculated by dividing the number of DAPI-positive nuclei within myotubes (MF-20 positive cells) by the total number of nuclei in the field. The fusion index was averaged over 10 fields and was 71% for GM17869 and 83% for GM17940. The fusion index for the GM17731 cell line was much lower at 46%.

Coriell ID		Muscle Origin	Age	Sex	Chromosome 4		Chromosome 10	
Lymphoblast	Myoblast				Allele 1	Allele 2	Allele 1	Allele 2
	GM17731	Rhomboid muscle	15	Male	3	25	13	15
GM17868	GM17869	Scapular and subscapularis muscle	21	Female	5	31	5	39
GM17939	GM17940	Not defined	13	Male	3	33	15	26

Table 4.1 Myoblast cell lines

Information on age, sex of the donor and the muscle of origin (where known) of the myoblast cell lines from the Corelli institute. Matched lymphoblast lines were available for two of the myoblast lines, GM17869 and GM17940. The predicted number of D4Z4 repeats on each allele is indicated for both chromosome 4 and 10.

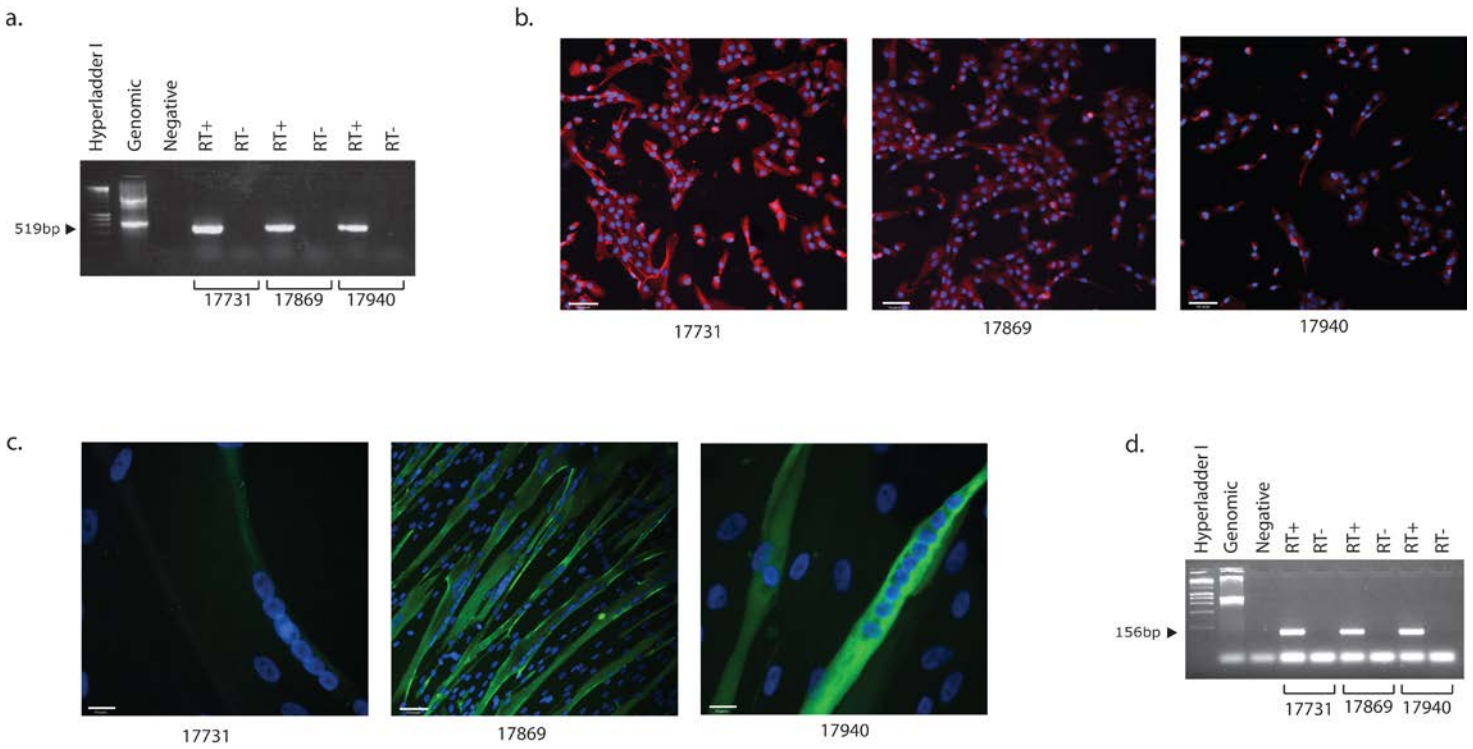


Figure 4.2 Characterisation of myoblast cell lines

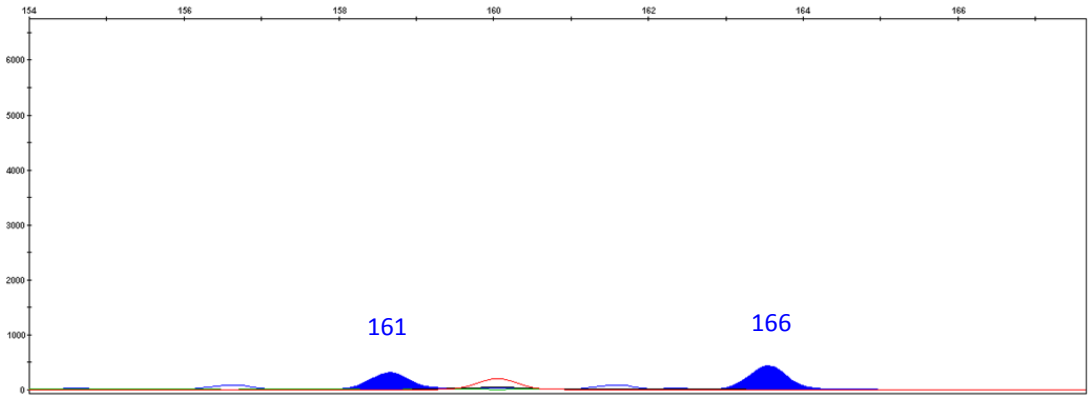
a) RT-PCR to show expression of desmin in myoblasts. For primers see table 2.5. b) Immunocytochemistry on myoblasts using an antibody to desmin (Table 2.1), shown in red. DAPI staining of the nucleus is shown in blue. c) Immunocytochemistry on myoblasts using the antibody MF20 (Table 2.1), shown in green. DAPI staining of the nucleus is shown in blue. d) RT-PCR to show expression of myosin in cell lines GM17731, GM17869 and GM17940 on differentiation into myotubes. For primers see Table 2.5.

4.2.2.i Haplotyping of myoblast cell lines

As the myoblast cell lines were derived from FSHD patients all were expected to contain the 4qA161 allele. Each of the cell lines were haplotyped to confirm the presence of the permissive allele and to determine the other alleles present. The haplotypes are listed in Figure 4.3b. The cell lines GM17869 and GM17940 were haplotyped by Joanne Pollington according to Lemmers *et al.* (2007) using sequence information from the amplification of the SSLP and p13E-11 region. DNA from the corresponding lymphoblast cell lines was used for these analyses. Haplotypes were assigned according to sequence variants (Lemmers *et al.* 2007, 2010), however, some haplotypes may not have been sampled due to PCR and cloning bias and so the complete set of haplotypes may not have been identified. As expected, each cell line contained at least one 4qA161 allele.

The GM17731 cell line was haplotyped using a modified version of the SSLP genotyping reaction described in Lemmers *et al.* (2010b). The products of this reaction were analysed by capillary gel electrophoresis (Figure 4.3a). This technique is sensitive enough to distinguish between single base differences in size. However, as the samples do not run true to size, they are compared with samples of known alleles as size standards. This cell line carries only 4qA161 and 10qA166 alleles and is presumably homozygous for each allele.

a.



b.

Coriell ID		Haplotypes
Lymphoblast	Myoblast	
	GM17731	4A161 and 10A166
GM17868	GM17869	4A161 and 10A166
GM17939	GM17940	4A161, 4B163, 10A166

Figure 4.3 Characterisation of myoblast cell lines

a) Capillary gel electrophoresis results from the 17731 myoblast cell line. Comparison with known size standards showed that samples were running at approximately 2.5 bp smaller than their true size so two allele sizes, 161 and 166, are predicted. A similar peak height between the two alleles suggests that two of each allele size is present in the cell line. b) Table to summarise haplotype and fusion index information. Haplotyping of GM17731 was by capillary gel electrophoresis as described in Section 2.9.2. Haplotyping of GM17869 and GM17940 were by PCR of lymphoblast cell lines by Joanne Pollington.

4.2.3 Initial attempts to amplify DUX4 from myoblast and rhabdomyosarcoma cell lines

Initially, three different primer pairs were used for RT-PCR, their positions and sequences are shown in Figure 4.4a and Table 2.5, respectively. HDUX4-F1 and HDUX4-R1 were designed previously by Dr Jannine Clapp to amplify an 812bp fragment corresponding to the distal end of the ORF and including the final 67bp of the second homeobox (Figure 4.1a). Primers #222 and #219 were used by Kowaljow *et al.* (2007) to amplify a 1477bp fragment encompassing the full ORF. Primers #222 and #407 were used by Dixit *et al.* (2007) to amplify a 1700bp fragment, with the reverse primer annealing in the pLAM region downstream of the putative stop codon.

As previous work by Dr Jannine Clapp had shown that the OneStep kit (Qiagen) is the most reliable reagent for amplifying the mouse Dux array (Clapp *et al.*, 2007), this kit was tested for amplification of the human sequences. The kit contains two reverse transcriptases and a HotStarTaq DNA polymerase that allows reverse transcription and amplification to take place in one reaction. Both reverse transcriptases have a higher affinity for RNA templates than other reverse transcriptases and one is optimised for use on small amounts of template which makes it useful for amplifying transcripts with low expression levels. Q-solution can also be included in the reaction, which helps amplification of difficult transcripts by modifying the melting behaviour of the nucleic acids (Qiagen OneStep RT-PCR Kit Handbook).

Using the OneStep kit with HDUX4 primers, no amplification was seen using RNA from either of the rhabdomyosarcoma cell lines (Figure 4.4b), or from the three myoblast cell lines in either proliferation or differentiation stages (Figure 4.4c). No amplification was seen from the testis or fetal skeletal muscle RNA used previously by Jannine Clapp. However, this product was not seen in all of her amplification attempts (Jane Hewitt, personal communication)

A two-step reaction was then performed with each of the primer pairs (HDUX4_F1/HDUX4_R1, #222/#219 and #222/#407). The Improm reverse transcription kit was used to make cDNA and the PCR step was performed with KAPAHiFi polymerase using a

separate buffer for GC rich templates. The reactions were performed using both the manufacturer's suggested cycling conditions, and those described in both the Dixit and Kowaljow publications (Dixit *et al.*, 2007; Kowaljow *et al.*, 2007). However, amplification was not seen from any of the RNA samples.

Finally, a two-step reaction was performed using the Superscript III reverse transcriptase for the initial step, and Platinum taq for the PCR step, as described by Dixit *et al*. and Kowlajow *et al.* However, no amplification of DUX4 was seen from any of the RNA samples used in this study (Figure 4.4d).

4.2.4 Amplification from human embryonal carcinoma cells

In 2009, Snider *et al*. showed amplification of 'full length' transcripts (see Figure 4.1c) from the D4Z4 array in human embryonal carcinoma cells (hEC) and human mesenchymal stem cells (hMSC). Human embryonal carcinoma cells are isolated from teratocarcinomas, a subset of germ cell tumours that form in the ovaries or testis. The testicular hEC cell line, NTERA-2, was cultured as described in Section 2.1. RT-PCR on RNA extracted from this cell line was performed using the OneStep kit and HDUX4 primer pair (Figure 4.5a). In contrast to the results with myoblast RNA, an 812bp product was amplified from both total (Figure 4.5b) and Poly $(A)^+$ RNA (Figure 4.5c). Although this product does not cover the full ORF, it does cover the central region that Snider *et al.* were unable to amplify from myoblasts. Thus, this data supports their report that longer transcripts are expressed in the hEC cell line. The RT-PCR products were sequenced and analysed as described in Section 4.2.6.

As both sense and antisense transcripts are expressed from the mouse *Dux* locus (Clapp *et al.*, 2007) and preliminary data from Dr Jannine Clapp had shown both sense and antisense transcripts for DUX4 (Figure 4.1a), the amplification protocol was modified to test for antisense RNA. When using the OneStep kit, both primers are present during the RT and PCR steps, so the amplified product could originate from either strand. Therefore, the reaction was modified so that only one of the primers was present in the RT step. The second primer was added after the reverse transcriptase had been inactivated. A product was

amplified from hEC RNA when either primer was used in the RT step (Figure 4.5d), indicating that transcription from this region occurs from both strands.

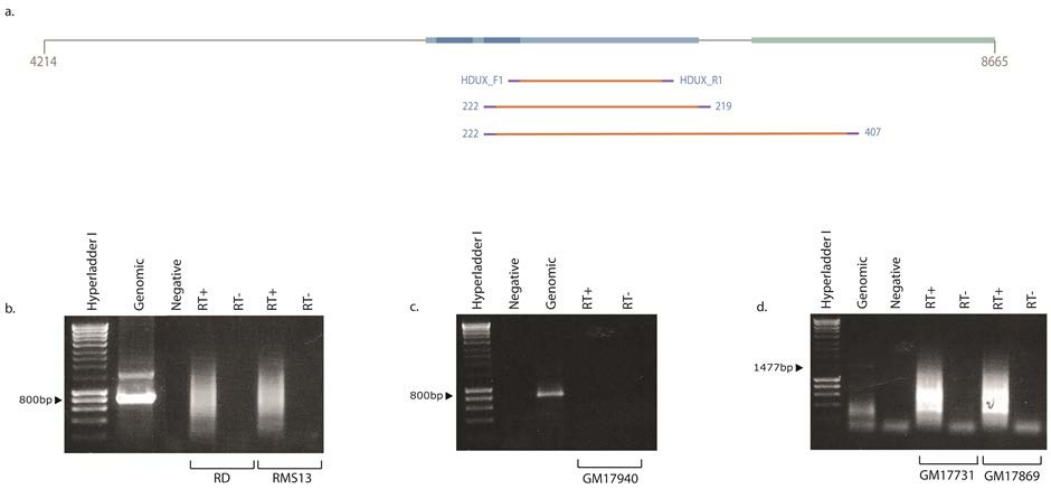


Figure 4.4 DUX4 RNA could not be detected in myoblast cells by RT-PCR

a) Schematic representation of the fragments amplified by primer pairs used in the initial amplification attempts. The DUX4 ORF is shown in light blue with the homeodomains highlighted in dark blue. The pLAM region is shown in green. Positions of the primer pairs are shown underneath. HDUX4_F1 and HDUX4_R1 were designed by Dr Jannine Clapp, primers #222 and #219 are from Kowaljow *et al.* and primers #222 and #407 were used in Dixit *et al.* b) RT-PCR of RNA extracted from the human rhabdomyosarcoma cell lines RD and RMS13 using the HDUX4 primer pair. c) RT-PCR of RNA extracted from myoblast cell line GM17940, using the HDUX4 primer pair. The same results were seen using the #222 and #407 primers used in the Dixit *et al.* publication. d) RT-PCR of RNA extracted from myoblast cell lines GM17731 and GM17869 using the #222 and #219 primers from Kowaljow *et al.*

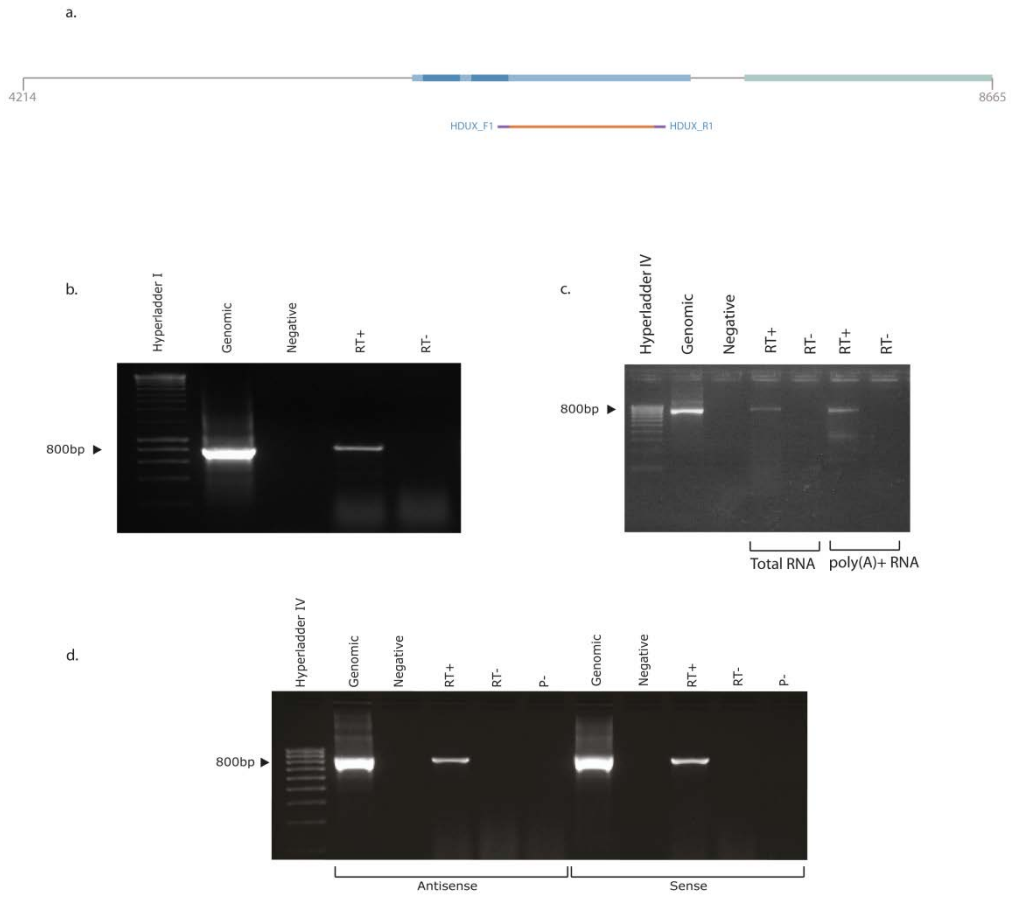


Figure 4.5 DUX4 expression in hEC cells

a) The DUX4 ORF is shown in pale blue with homeodomains highlighted in dark blue. The pLAM region is shown in green. The HDUX4 primers amplify an 812bp fragment, the position of this is shown underneath. b, c) RT-PCR of hEC RNA using the HDUX4 primer pair. d) Strand-specific RT-PCR of hEC RNA using the HDUX4 primer pair. p- indicates that only the primer used for the RT reaction was included. The RT-PCR band appears larger than the genomic, however, sequencing genomic fragments confirmed that this is an artefact of the electrophoresis.

4.2.5 Haplotyping of the hEC cell line

FSHD results from deletions on specific haplotypes only. Therefore, the hEC cell line was haplotyped to determine whether the expressed transcripts originated from an FSHD permissive allele. The SSLP and p13E-11 regions were amplified and sequenced according to Lemmers *et al.* (2007) (See section 1.14). Haplotypes were assigned based on 15 SNPs within the p13E-11 region (Figure 4.6). Of the 10 fragments sequenced, 4 were 10qA166 and the remaining 6 all 4qB163.

To confirm this data, a modified version of the SSLP genotyping reaction described in Lemmers *et al.* was subsequently performed and the products were analysed by capillary gel electrophoresis as described previously (section 4.2.2i). Two allele sizes, 163 and 166, were identified (Figure 4.7) consistent with the p13E-11 sequence data. The peak heights are approximately equal. The most likely constitution for this cell line is homozygous for both 4qB163 and 10qA166. Therefore, the expressed transcripts do not appear to come from an FSHD permissive allele and cannot provide information about variants that may be involved in the FSHD phenotype. However, as multiple D4Z4 alleles have been conserved and amplification of RT-PCR products from this cell line shows there is transcription from these non-permissive alleles, they are likely to have a role in normal development. Thus, data from this cell line can provide information on repeat variants that are expressed during normal development.

4.2.6 Sequence analysis of hEC transcripts

From sequence data being produced by others in Jane Hewitt's laboratory, it was clear that there is substantive nucleotide variation within D4Z4 arrays. Therefore, in order to identify sequence variants between D4Z4 repeats, the hDUX4 RT-PCR products were cloned into pSMART and 88 clones were sequenced to determine whether all the transcripts were the same sequence. These primers cover 812bp of the ORF and include the final 67bp of the second homeobox. From the 88 sequences, 17 distinct groups were identified based on 10

variant positions and a 6bp indel. Where a variant was present in only one of the 88 sequences it was not used to define a group as it was not possible to exclude that these were due to PCR errors. The 17 groups were compared to the sequence of the most distal D4Z4 repeat within GenBank sequence FJ439133 and as all of the clones vary from FJ439133 at position 6779 an additional variant was identified. Variants are shown in Table 4.2 and Figure 4.8. This data indicates that multiple repeats are expressed in the hEC cell line. Although the data are consistent with the sequences coming from chromosome 4, it is not possible to rule out that they come from chromosome 10. Two RT-PCR products from the poly(A)$^+$ RNA were also sequenced, these matched the variant groups 6 and 15.

Some of the sequences appear to be more common (for example 19 clones contained the same variants in group 6), but this could be due to biases in amplification and/or cloning rather than transcript abundance. All the clones contain a silent C-G change at position 6779 and therefore none of the sequences are 100% identical to the repeat in FJ439133. This is not unexpected, as the cell line does not contain a 4qA161 allele from which the FJ439133 allele originates and it is likely that this distal repeat varies between haplotypes. For 19 of the sequences this C-G change is the only difference to FJ439133. Importantly, no nonsense or frameshift mutations were seen and 6/11 of the base changes are silent. There is a 50% indel in 50% of the clones, alignment with sequence from the chimp suggests that a single copy is ancestral (Jane Hewitt, personal communication).

None of these variants fall within the 67bp of homeobox sequence at the start of the product. The remaining 744bp sequence cover the C-terminal domain which may be less constrained than the homeodomains and could account for the high number of variants in this stretch. Previous work in Jane Hewitt's laboratory had identified highly conserved positions in this region (Figure 4.9). The C-T variation at position 7122 falls within a highly conserved leucine residue, however this base change is silent. At position 7147, the C-T variation identified in 6 sequences results in a proline to leucine change. This proline residue is conserved between human, chimp, orangutan, macaque, marmoset, elephant and treeshrew (Figure 4.9).

```
4qB168     1 AGCCAGACTTCATCAAATTTCTGCAGCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA
10qA166    1 AGCCAGACTTCATCAAATTTCTGCAGCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA
4qA166     1 AGCCAGACTTCATCAAATTTCTGCAGCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA
4qA161     1 AGCCAGACTTCATCAAATTTCTGCACCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA
4qB163     1 AGCCAGACTTCATCAAATTTCTGCACCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA

hEC 1      1 AGCCAGACTTCATCAAATTTCTGCACCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA
hEC 2      1 AGCCAGACTTCATCAAATTTCTGCAGCAATGAAAAAAAAATTTACAAGAGAAAAACAAAA
                                       *

4qB168    61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCCGCTC
10qA166   61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCCGCTC
4qA166    61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCCGCTC
4qA161    61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCTGCTC
4qB163    61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCTGCTC

hEC 1     61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCTGCTC
hEC 2     61 AACCCTATTAAACGTCACGGACAAGGCCAGAGTTTGAATATACTGTGGTCATCTCCGCTC
                                                                     *

4qB168   121 CAGTGCAAACTGTTTTCAGAAAGCCTGCTTTTATTTTCCTTGCTGTAACAGAGGAACATT
10qA166  121 CAGTGCAAACTGTTTTCAGAAAGCCTGCTTTTATTTTCCTTGCTGTAACAGAGGAACATT
4qA166   121 CAGTGCAAACTGTTTTCAGAAAGCCTGCTTTTATTTTCCTTGCTGTAACAGAGGAACATT
4qA161   121 CAGTGCAAACTGTTTCCAGAAAGCCTACTTCTATTTTCCTTGCTGTAACAGAGGAACATT
4qB163   121 CAGTGCAAACTGTTTTCAGAAAGCCTACTTCTATTTTCCTTGCTGTAACAGAGGAACATT

hEC 1    121 CAGTGCAAACTGTTTTCAGAAAGCCTACTTCTATTTTCCTTGCTGTAACAGAGGAACATT
hEC 2    121 CAGTGCAAACTGTTTTCAGAAAGCCTGCTTTTATTTTCCTTGCTGTAACAGAGGAACATT
                           *          *   *

4qB168   181 TCCTGTCTTATGCTTATTCTACTCTGCAGTCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA
10qA166  181 TCCTGTCTTATGCTTATTCTACTCTGCAGTCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA
4qA166   181 TCCTGTCTTATGCTTATTCTACTCTGCAGTCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA
4qA161   181 TCCTGTCTTATGTTTATTCTACTCTGCAATCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA
4qB163   181 TCCTGTCTTATGCTTATTCTACTCTGCAATCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA

hEC 1    181 TCCTGTCTTATGCTTATTCTACTCTGCAATCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA
hEC 2    181 TCCTGTCTTATGCTTATTCTACTCTGCAGTCCCCTAAGGCTTTTTCTCTCCCTCCCAGAA
                         *                 *

4qB168   241 TCTTAAAGTGCATTCGAACGCACAGGCAAAATCCTCCCAGAAACTTGTGAAAACATAAAT
10qA166  241 TCTTAAAGTGCATTCGAACGCACAGGCAAAATCCTCCCAGAAACTTGTGAAAACATAAAT
4qA166   241 TCTTAAAGTGCATTCGAACGCACAGGCAAAATCCTCCCAGAAACTTGTGAAAACATAAAT
4qA161   241 TCTTAAAGTGCATTCGAACTCACAGGCAAAATCCTCCCAGAATCTTGTGAGAACATAAAT
4qB163   241 TCTTAAAGTGCATTCGAACTCACAGGCAAAATCCTCCCAGAATCTTGTGAGAACATAAAT

hEC 1    241 TCTTAAAGTGCATTCGAACTCACAGGCAAAATCCTCCCAGAATCTTGTGAGAACATAAAT
hEC 2    241 TCTTAAAGTGCATTCGAACGCACAGGCAAAATCCTCCCAGAAACTTGTGAAAACATAAAT
                               *                   *       *

4qB168   301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG
10qA166  301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG
4qA166   301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG
4qA161   301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG
4qB163   301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG

hEC 1    301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG
hEC 2    301 GATCTGACTAGTTTGGCATTGCTTTTGGGGATCTGGGAAAATCTGTGCACACTTCTGGAG



4qB168   361 ACCCTTGTCAGGCCATTTTTTATAAATCTATTGTGCCTCAAGTCAGAAGTGTCTGAGGGG
10qA166  361 ACCCTTGTCAGGCCATTTTTTATAAATCTATTGTGCCTCAAGTCAGCAGTGTGTGAGGGG
4qA166   361 ACCCTTGTCAGGCCATTTTTTATAAATCTATTGTGCCTCAAGTCAGAAGTGTGTGAGGGG
4qA161   361 ACCCTTGTCATGCCATTTTTTATAAATCTATTGTGCCTCAAGTCAGAAGTGTGTGAGGGG
4qB163   361 ACCCTTGTCATGCCATTATTTATAAATCTATTGTGCCTCAAGTCAGAAGTGTGTGAGGGG

hEC 1    361 ACCCTTGTCATGCCATTATTTATAAATCTATTGTGCCTCAAGTCAGAAGTGTGTGAGGGG
hEC 2    361 ACCCTTGTCAGGCCATTTTTTATAAATCTATTGTGCCTCAAGTCAGCAGTGTGTGAGGGG
                       *      *                                *     *
```

```
4qB168   421 AGATGGGGAGACATTGGGATGTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC
10qA166  421 AGATGGGGAGACATTGGGATGCGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC
4qA166   421 AGATGGGGAGACATTGGGATGCGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC
4qA161   421 AGATGGGGAGACATTGGGATGCGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC
4qB163   421 AGATGGGGAGACATTGGGATGCGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC

hEC 1    421 AGATGGGGAGACATTGGGATGCGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC
hEC 2    421 AGATGGGGAGACATTGGGATGCGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTC
                                      *
```

Figure 4.6 Assignment of p13E11 sequences from the hEC cell line

Alignment of the p13E-11 sequences from Lemmers *et al.* (2007) used to assign haplotypes to the hEC cell line. The 15 variant positions are highlighted with an asterisk. The sequences of two clones corresponding to the 4qB163 and 10qA166 alleles are shown.

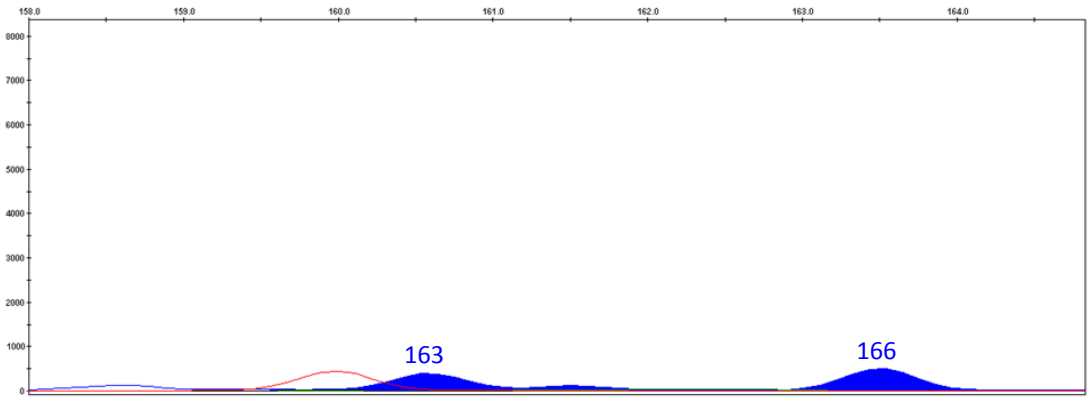


Figure 4.7 Haplotyping of the hEC cell line.

Capillary gel electrophoresis results from the hEC cell line. Comparison with known size standards showed that samples were running at approximately 2.5 bp smaller than their true size so two allele sizes, 163 and 166, are predicted. A similar peak height between the two alleles suggests that two of each allele size is present in the cell line.

```
Group 1           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 2           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 3           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 4           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 5           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 6           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 7           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 8           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 9           GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 10          GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 11
Group 12          GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 13          GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 14          GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 15          GGGCTTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 16          GGGCCTCCCGGAGTCCAGGATTCGGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
Group 17          GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG

FJ439133    6398  GGGCCTCCCGGAGTCCAGGATTCAGATCTGGTTTCAGAATCGAAGGGCCAGGCACCCGGG
                    G  L  P  E  S  R  I  Q  I  W  F  Q  N  R  R  A  R  H  P  G
                       *                 *

Group 1           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 2           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 3           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 4           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 5           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 6           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 7           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 8           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 9           ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 10          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 11
Group 12          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 13          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 14          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 15          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 16          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
Group 17          ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG

FJ439133    6458  ACAGGGTGGCAGGGCGCCCGCGCAGGCAGGCGGCCTGTGCAGCGCGGCCCCCGGCGGGGG
                    Q  G  G  R  A  P  A  Q  A  G  G  L  C  S  A  A  P  G  G  G


Group 1           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGGGTGGGGAACGGGGCTTCC
Group 2           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGGGTGGGGAACGGGGCTTCC
Group 3           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 4           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 5           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 6           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 7           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 8           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGAGTGGGGAACGGGGCTTCC
Group 9           TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 10          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 11
Group 12          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 13          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 14          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGAGTGGGGAACGGGGCTTCC
Group 15          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGAGTGGGGAACGGGGCTTCC
Group 16          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
Group 17          TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGAGTGGGGAACGGGGCTTCC

FJ439133    6518  TCACCCTGCTCCCTCGTGGGTCGCCTTCGCCCACACCGGCGCGTGGGGAACGGGGCTTCC
                    H  P  A  P  S  W  V  A  F  A  H  T  G  A  W  G  T  G  L  P
                                                      *
```

```
Group 1          CGCACCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 2          CGCACCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 3          CGCACCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 4          CGCACCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 5          CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 6          CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 7          CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 8          CGCGCCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 9          CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 10         CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCCCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 11                                   GGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 12         CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 13         CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 14         CGCACCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 15         CGCACCGCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 16         CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
Group 17         CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC

FJ439133    6578 CGCACCCCACGTGCCCTGCGCGCCTGGGGCTCTCCCACAGGGGGCTTTCGTGAGCCAGGC
                  A  P  H  V  P  C  A  P  G  A  L  P  Q  G  A  F  V  S  Q  A
                    *  *                          *

Group 1          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 2          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 3          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 4          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 5          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
Group 6          AGCAAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
Group 7          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
Group 8          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 9          AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
Group 10         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 11         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
Group 12         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGGTCTCCCA
Group 13         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGGTCTCCCA
Group 14         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 15         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA
Group 16         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
Group 17         AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGACGGCAGAGGGGGTCTCCCA

FJ439133    6638 AGCGAGGGCCGCCCCCGCGCTGCAGCCCAGCCAGGCCGCGCCGGCAGAGGGGATCTCCCA
                  A  R  A  A  P  A  L  Q  P  S  Q  A  A  P  A  E  G  I  S  Q
                    *                                   *           *

Group 1          ACCTGCCCCGGCGCGCCGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 2          ACCTGCCCCGGCGCGCCGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 3          ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 4          ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 5          ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 6          ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 7          ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 8          ACCTGCCCCGGCGCGCCGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGGCGGGGCGCT
Group 9          ACCTGCACCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 10         ACCTGCCCCGGCGCGCCGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 11         ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 12         ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 13         ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 14         ACCTGCCCCGGCGCGCCGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 15         ACCTGCCCCGGCGCGCCGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 16         ACCTGCACCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
Group 17         ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT

FJ439133    6698 ACCTGCCCCGGCGCGCGGGGATTTCGCCTACGCCGCCCCGGCTCCTCCGGACGGGGCGCT
                  P  A  P  A  R  G  D  F  A  Y  A  A  P  A  P  P  D  G  A  L
                        *        *                                 *
```

```
Group 1           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 2           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 3           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 4           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 5           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGTCGGGAGGACCGGGA
Group 6           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 7           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 8           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 9           CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 10          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 11          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 12          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 13          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 14          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 15          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 16          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
Group 17          CTCCCACCCTCAGGCTCCTCGGTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA

FJ439133    6758  CTCCCACCCTCAGGCTCCTCGCTGGCCTCCGCACCCGGGCAAAAGCCGGGAGGACCGGGA
                   S  H  P  Q  A  P  R  W  P  P  H  P  G  K  S  R  E  D  R  D
                                        *                    *

Group 1           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 2           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 3           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 4           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 5           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 6           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 7           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 8           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 9           CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 10          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTG:CACAGCCTGGGCCCGCTCAAGC
Group 11          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 12          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 13          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 14          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 15          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 16          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
Group 17          CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC

FJ439133    6818  CCCGCAGCGCGACGGCCTGCCGGGCCCCTGCGCGGTGGCACAGCCTGGGCCCGCTCAAGC
                   P  Q  R  D  G  L  P  G  P  C  A  V  A  Q  P  G  P  A  Q  A
                                                     *

Group 1           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 2           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 3           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCGCGTCCCAGGGGAGTCCGTGGTGGGG
Group 4           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCGCGTCCCAGGGGAGTCCGTGGTGGGG
Group 5           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 6           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 7           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 8           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 9           GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 10          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 11          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 12          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGG:
Group 13          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 14          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 15          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 16          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
Group 17          GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG

FJ439133    6878  GGGGCCGCAGGGCCAAGGGGTGCTTGCGCCACCCACGTCCCAGGGGAGTCCGTGGTGGGG
                   G  P  Q  G  Q  G  V  L  A  P  P  T  S  Q  G  S  P  W  W  G
                                                     *                    *
```

```
Group 1           CTGGGGCCGGGGTCCGCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 2           CTGGGGCCGGGGTCCGCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 3           CTGGGGCCGGGGTCCGCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 4           CTGGGGCCGGGGTCCGCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 5           CTGGGGCCGGGGTCCGCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 6           CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 7           CTGGGGCCGGGGTCCTCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 8           CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 9           CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 10          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 11          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 12          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 13          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 14          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 15          CTGGGGCCGGGGTCCGCAGGTCGCTGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 16          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
Group 17          CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC

FJ439133    6938  CTGGGGCCGGGGTCCCCAGGTCGCCGGGGCGGCGTGGGAACCCCAAGCCGGGGCAGCTCC
                   W  G  R  G  P  Q  V  A  G  A  A  W  E  P  Q  A  G  A  A  P
                               *        *
```

```
Group 1           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 2           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 3           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 4           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 5           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 6           ACCCCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 7           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 8           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG
Group 9           ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG
Group 10          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCACCTCCGCGCGGCAGGGGCAGATGCAAGG
Group 11          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG
Group 12          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG
Group 13          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG
Group 14          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 15          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 16          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCC::::::GCGCGGCAGGGGCAGATGCAAGG
Group 17          ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG

FJ439133    6998  ACCTCCCCAGCCCGCGCCCCCGGACGCCTCCGCCTCCGCGCGGCAGGGGCAGATGCAAGG
                   P  P  Q  P  A  P  P  D  A  S  A  S  A  R  Q  G  Q  M  Q  G
                     *                          ******
```

```
Group 1           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 2           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 3           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 4           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 5           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 6           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 7           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 8           CATCCCGGCGCCCTCCCAGGCGCTCCGGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 9           CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 10          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 11          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 12          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 13          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 14          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 15          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACCCCCCTGCGG
Group 16          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
Group 17          CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG

FJ439133    7058  CATCCCGGCGCCCTCCCAGGCGCTCCAGGAGCCGGCGCCCTGGTCTGCACTCCCCTGCGG
                   I  P  A  P  S  Q  A  L  Q  E  P  A  P  W  S  A  L  P  C  G
                                          *                    *
```

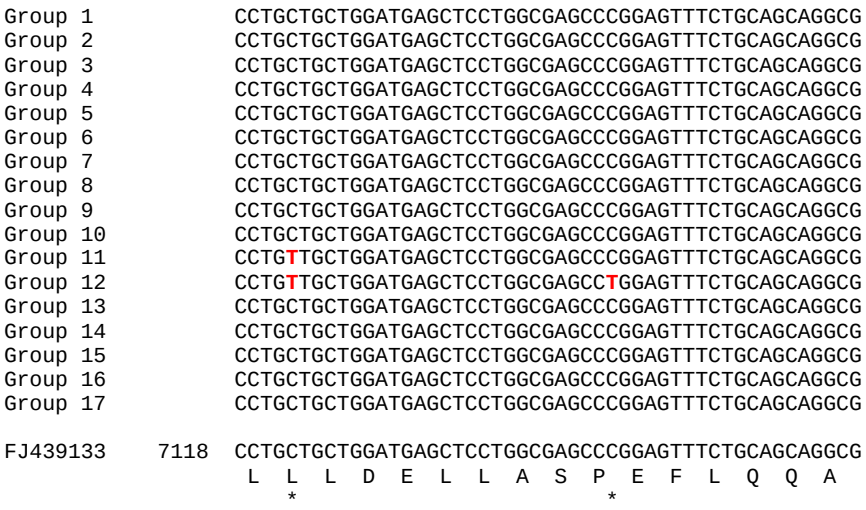


Figure 4.8 Alignment of the 17 distinct groups of expressed sequences

Alignment of the 17 distinct groups identified in the expressed RT-PCR products. DNA sequences were edited manually to correct base-calling errors. Primer sequences were trimmed and sequences aligned using ClustalW. Sequences are compared to the equivalent region of the most distal D4Z4 repeat unit in the FJ439133 (GenBank) reference sequence. Group 11 is represented by a single clone that failed to sequence from the 5' end. The translation is shown underneath and base changes from the consensus are highlighted with an asterisk. For an overview of the variants see Table 4.2.

	Position	6559	6584	6678	6690	6704	6714	6779	6953	6963	7031	7122	7147
Variant Group	**Number of Clones**	**C**	**C**	**C**	**A**	**C**	**G**	**C**	**C**	**C**	**(GCCTCC)n**	**C**	**C**
1	4	G	G	A	G	C	C	G	G	T	1	C	C
2	1	A	G	A	G	C	C	G	C	T	1	C	C
15	2	A	G	A	G	C	C	G	G	T	1	C	C
14	1	A	G	A	G	C	C	G	C	C	1	C	C
3	2	C	G	A	G	C	G	G	G	T	1	C	C
8	11	A	G	A	G	C	G	G	C	C	2	C	C
17	3	A	C	A	G	C	G	G	C	C	2	C	C
10	1	C	C	A	G	C	C	G	C	C	2	C	C
4	3	C	C	C	G	C	G	G	C	C	1	C	C
5	1	C	C	C	A	C	G	G	G	T	1	C	C
6	19	C	C	C	A	C	G	G	C	C	1	C	C
7	11	C	C	C	A	C	G	G	C	T	1	C	C
9	7	C	C	C	A	A	G	G	C	C	2	C	C
11	1	nd	nd	C	A	C	G	G	C	C	2	T	C
12	6	C	C	C	G	C	G	G	C	C	2	T	T
13	14	C	C	C	G	C	G	G	C	C	2	C	C
16	1	C	C	C	A	A	G	G	C	C	1	C	C
Coding change		* A - E A - G		* P - T	* I - V		* G - R						* P - L

Table 4.2 Summary of variant groups identified from hEC RT-PCR products

The FJ439133 sequence is shown at the top with sequence position identified. Sequence variants for each of the 17 haplotypes are shown, blue blocks are the same base as the FJ439133 sequence, orange blocks indicate a change. All variants fall within the ORF, however, none are within the homeobox region. Coding changes are highlighted with an asterisk and the amino acid change is shown underneath. There is a 6bp indel in 50% of clones, 1 indicates a single copy of the GCCTCC sequence, 2 indicates there are two copies present in the variant.

Figure 4.9 Clustal alignment of the conserved C-terminal domain

Taken from Clapp *et al.* (2007). Clustal alignments of the C-terminal regions of DUXc, DUX4 and the mouse and rat DUX proteins. Residues that are invariant in all species are highlighted in black. Residues that are conserved in at least 60% sequences are highlighted in dark grey, with conservation substitutions in light grey. The proline residue which is changed in 6 of the hEC RT-PCR products is indicated with an asterisk.

4.2.6.i Comparison of expressed variant groups with genomic sequence data

As described in Section 4.2.5, it is likely that both of the chromosome 4 alleles in the hEC cell line are 4qB163 alleles. The HHW416 somatic cell hybrid (donated by Dr. Sara Winokur) was haplotyped by an MRes student in the laboratory using sequence variants identified by Lemmers *et al.* (2007), and was shown to contain a 4qB163 allele. Amplification of overlapping fragments that covered the repeat unit had been carried out on genomic DNA from the cell line. 89 PCR products had been cloned and sequenced. As both cell lines contain the 4qB163 allele, the expressed hEC sequences were compared with the genomic HHW416 sequences, to determine whether they shared variants and might provide information about which repeats in the array are expressed.

Like the expressed sequences, the HHW416 sequences were compared with the GenBank sequence FJ439133. The 104 genomic sequences were split into 26 distinct groups which were identified by 20 variant positions (Table 4.3). Of the 14 variant positions identified in the hEC sequences, 10 of these were also seen in the HHW416 sequences. However, of the 17 expressed hEC groups, only 4 were identical to a HHW416 genomic group (Table 4.4). The genomic sequences are longer than the expressed hEC transcripts, as such they contain more variant positions and based on this extra information they could be separated into more distinct groups than the hEC transcripts. This has resulted in more than one of the genomic sequences being a potential match for the expressed hEC groups, 13 and 17 (see Table 4.4).

Work in Jane Hewitt's laboratory has also identified variant positions in 4qA161 and 10qA166 genomic sequences, the number of genomic clones containing each of the variants identified in this work have been included in Table 4.3. All 31 of the 10qA166 clones contained the C-G variant at position 6779, this variant was not seen in any of the 4qA161 clones. In addition, 4 4qA161 and 1 10qA166 alleles contained the GCCTCC duplication, and a further 3 variants were seen in single 4qA161 sequences.

4.2.6.ii Comparison of sense and antisense data

In order to establish whether any of the expressed haplotypes were specific to sense or antisense expression, sequences were grouped according to the strand from which they were transcribed. Where both primers were present during the reverse transcription reaction, sequences were recorded as unknown. The data is summarised in Table 4.5 and although there are some haplotypes that contain only one type of transcript, there are no large preferences for one or the other strand in any of the haplotypes.

	Position	6023	6559	6584	6678	6690	6714	6779	6793	6909	6910	6953	6962	6991	7023	7031	7147	7246	7322	7428	7631
Variant Group	Number of Clones	C	C	C	C	A	G	C	C	C	C	C	C	C	G	(GCCTCC)n	C	A	C	T	C
1	4	C	C	C	C	G	G	G	C	C	C	C	C	C	G	2	C	A	G	T	C
2	12	C	C	C	C	G	G	G	C	C	C	C	C	C	G	2	C	A	G	C	C
3	2	C	C	C	C	G	G	G	C	C	C	C	C	C	A	2	C	A	G	C	C
4	15	C	C	C	C	G	G	G	C	C	C	C	C	C	G	2	C	A	·	T	C
5	1	C	C	C	C	G	G	G	C	C	C	C	C	C	G	2	T	A	·	T	C
6	1	T	C	C	C	G	G	G	C	C	C	C	C	C	G	2	T	A	·	T	C
7	8	T	C	C	C	G	G	G	C	C	C	C	C	C	G	2	C	A	·	T	C
8	1	C	A	C	A	G	G	G	C	C	C	C	C	C	G	2	C	A	G	C	C
9	24	C	A	C	A	G	G	G	C	C	C	C	C	C	G	2	C	A	G	T	A
10	4	C	A	C	A	G	G	G	C	C	C	C	C	C	G	2	C	A	·	T	C
11	1	C	A	G	A	G	G	G	C	C	C	C	C	C	G	2	C	A	·	T	C
12	1	T	C	C	C	A	G	G	C	C	T	C	C	C	G	2	C	A	·	T	C
13	1	T	C	C	C	A	G	C	C	C	T	C	C	C	G	2	C	A	·	T	C
14	1	C	C	C	C	A	G	G	T	C	C	C	C	C	G	2	C	A	G	T	C
15	1	C	C	C	C	G	G	C	T	C	C	C	T	A	G	1	C	A	G	T	C
16	3	C	C	C	C	A	G	G	C	C	C	C	T	A	G	1	C	A	G	T	C
17	1	T	C	C	C	A	G	G	C	C	C	C	T	A	G	1	C	A	·	T	C
18	1	T	C	C	G	A	G	G	C	C	C	C	T	C	G	1	C	A	·	T	C
19	1	T	C	C	C	A	G	G	C	T	C	C	T	C	G	1	C	A	·	T	C
20	1	C	C	C	C	A	G	G	C	C	C	C	T	C	G	1	C	A	·	T	C
21	3	C	C	C	G	A	G	G	C	C	C	C	T	C	G	1	C	A	G	T	C
22	1	C	C	C	C	G	G	G	C	C	C	C	C	C	G	1	C	A	G	T	C
23	1	C	A	G	A	G	G	G	C	C	C	G	T	C	G	1	C	A	G	T	A
24	10	C	A	G	A	G	G	G	C	C	C	G	T	C	G	1	C	A	G	T	C
25	2	C	A	G	A	G	C	G	C	C	C	G	T	C	G	1	C	A	·	T	C
26	3	C	A	G	A	G	C	G	C	C	C	G	T	C	G	1	C	T	G	T	C
Coding change			* A - E		* P - T	* I - V	* G - R		* P - L	* P - S	* P - L			* A - E	* A - T		* P - L	* E - V	Downstream of ORF		
hEC RT-PCR products		nd	✓	✓	✓	✓	✓	✓				✓	✓			✓	✓		nd	nd	nd
4qA161 genomic sequences		0	1/30	0	1/30	1/30	0	0	0	0	0	0	0	0	0	4/30	0	0	nd	nd	nd
10qA166 genomic sequences		nd	0	0	0	0	0	31/31	0	0	0	0	0	0	0	1/31	0	0	nd	nd	nd

Table 4.3 Distinct sequence groups amplified from the HHW416 cell line

The FJ439133 sequence is shown on the top line. Sequence variants for each of the groups are shown, blue blocks are the same base as the FJ439133 sequence, orange blocks indicate a change from the consensus. Coding changes are highlighted with an asterisk. The in-frame duplication at position 7031 in 50% of the groups, 1 indicates a single copy of the GCCTCC sequence, 2 indicates two copies present. Variants that are also present in the hEC RT-PCR products (see Table 4.1) are indicated with a ✓. Work in Jane Hewitt's laboratory has also identified variant positions in 4qA161 and 10qA166 genomic sequences; the number of genomic clones containing each base change is indicated for each of these alleles.

Expressed group	Genomic Variant Group
1	No match
2	No match
3	No match
4	No match
5	No match
6	No match
7	20
8	11
9	No match
10	No match
11	No match
12	No match
13	1,2,4,7
14	No match
15	No match
16	No match
17	8,9,10

Table 4.4 Comparison of RT-PCR products and genomic sequence

Sequences of transcripts amplified from hEC RNA were compared with genomic sequence data from the HHW416 cell line. Genomic haplotypes which potentially match the expressed groups are shown in the right hand column.

Group	Sense	Antisense	Unknown	Total
1		1	3	4
2		1		1
3	1	1		2
4	3			3
5			1	1
6	3		16	19
7	5	4	2	11
8	2	7	2	11
9	3		4	7
10	1			1
11	1			1
12		5	1	6
13	7	4	3	14
14		1		1
15		1	1	2
16	1			1
17	2	1		3

Table 4.5 Comparison of group between sense and antisense transcripts

The distribution of sense and antisense transcripts across the 17 expressed groups identified from the hEC cell RNA. Where both primers were present during the reverse transcription reaction the orientation of transcripts is listed as unknown.

4.2.7 Expression at the 3'and 5' ends of the ORF

In their 2009 publication, Snider *et al.* also had difficulty amplifying full length transcripts from D4Z4 however they were able to amplify smaller transcripts from the 3' and 5' ends of the ORF. The RT-PCRs they describe in the paper (Figure 4.10a) were repeated on RNA extracted from the myoblast cell lines established in this lab. However, no amplification was seen from either proliferating or differentiating stages (Figure 4.10d).

It was possible to amplify transcripts at the 5' end of the ORF (primer pairs 1797/1906 and 1797/2235) from hEC cell RNA (Figure 4.10c); although attempts to amplify products from the 3' end of the ORF (2672/2952) were unsuccessful (Figure 4.10c). The fragments from primers 1797-2235 were cloned and 11 were sequenced, 8 of which were identical (Figure 4.11). In addition, the final 3 sequences had only a single base pair change from FJ439133, this was in a different position in each sequence so could not be excluded as a PCR error. As this PCR product covers the homeobox sequences it would be expected to be highly constrained and so a lack of variant positions in this region is expected.

This fragment overlaps with the first 96bp of the HDUX4 transcript to cover the whole of the ORF (Figure 4.10b). However, the overlap does not contain sufficient variants to link unambiguously these RT-PCR products to those amplified with the HDUX4 primers.

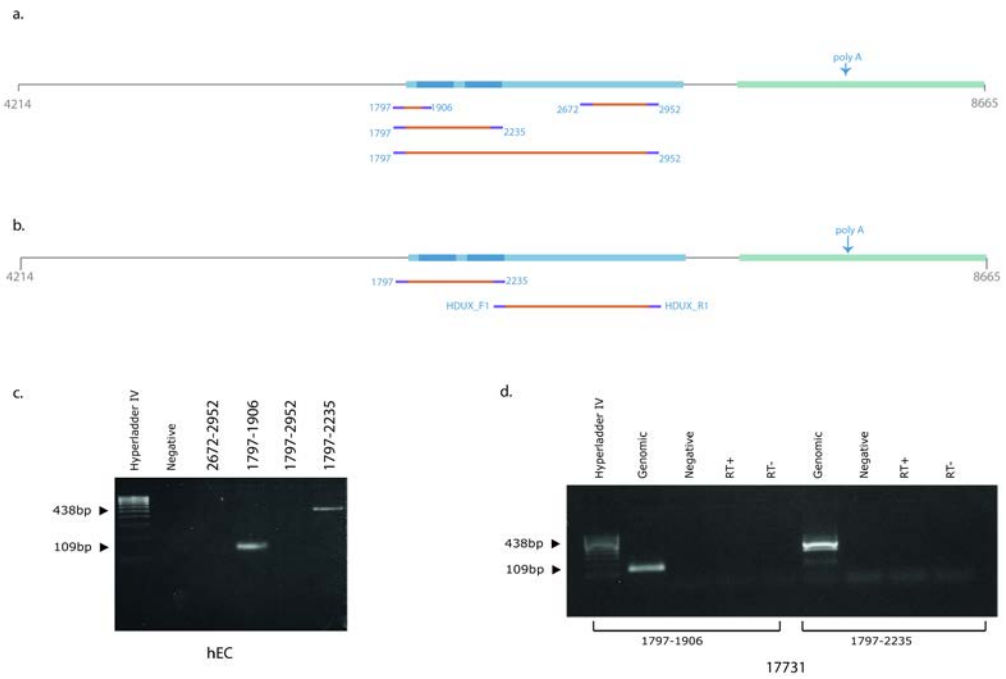


Figure 4.10 Attempts to amplify fragments from the ORF using primers from Snider *et al.* (2009)

a) Location of the primer pairs used by Snider *et al.* (2009). The ORF is shown in pale blue with homeodomains highlighted in dark blue. The pLAM region is shown in green. Predicted transcripts are shown underneath in orange. b) Schematic representation of the overlap between the HDUX4 and 1797/2235. c) RT-PCR on RNA extracted from hEC cells using Tapscott primer pairs shown in a. d) RT-PCR on RNA extracted from GM17731 cells using primer pairs 1797/1906 and 1797/2235. The integrity of all RNA samples was checked using an RT-PCR for β-actin, as described in Section 2.4.1 and Figure 2.5.

```
FJ439133 6001   CCGCGATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 3                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGGG
hEC 4                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 7                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 13               ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 12               ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 11               ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 8                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 6                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 9                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 10               ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
hEC 5                ATGGCCCTCCCGACACCCTCGGACAGCACCCTCCCCGCGGAAGCCCGGGGACGAG
                   A  M  A  L  P  T  P  S  D  S  T  L  P  A  E  A  R  G  R  G
                                                                        *

FJ439133 6061   GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 3           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 4           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 7           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 13          GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 12          GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 11          GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 8           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 6           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 9           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 10          GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
hEC 5           GACGGCGACGGAGACTCGTTTGGACCCCGAGCCAAAGCGAGGCCCTGCGAGCCTGCTTTG
                  R  R  R  R  L  V  W  T  P  S  Q  S  E  A  L  R  A  C  F  E

FJ439133 6121   AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 3           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 4           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 7           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 13          AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 12          AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 11          AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 8           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 6           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 9           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 10          AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
hEC 5           AGCGGAACCCGTACCCGGGCATCGCCACCAGAGAACGGCTGGCCCAGGCCATCGGCATTC
                  R  N  P  Y  P  G  I  A  T  R  E  R  L  A  Q  A  I  G  I  P

FJ439133 6181   CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 3           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 4           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 7           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 13          CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 12          CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 11          CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 8           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 6           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 9           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 10          CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
hEC 5           CGGAGCCCAGGGTCCAGATTTGGTTTCAGAATGAGAGGTCACGCCAGCTGAGGCAGCACC
                  E  P  R  V  Q  I  W  F  Q  N  E  R  S  R  Q  L  R  Q  H  R


FJ439133 6241   GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 3           GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 4           GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 7           GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 13          GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 12          GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 11          GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 8           GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 6           GGCGGGAATCTCGGCCCTGGCCCGGGAGACACGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 9           GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 10          GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
hEC 5           GGCGGGAATCTCGGCCCTGGCCCGGGAGACGCGGCCCGCCAGAAGGCCGGCGAAAGCGGA
                   R  E  S  R  P  W  P  G  R  R  G  P  P  E  G  R  R  K  R  T
                                             *

FJ439133 6301   CCGCCGTCACCGGATCCCAGACCGCCCTGCTCCTCCGAGCCTTTGAGAAGGATCGCTTTC
```

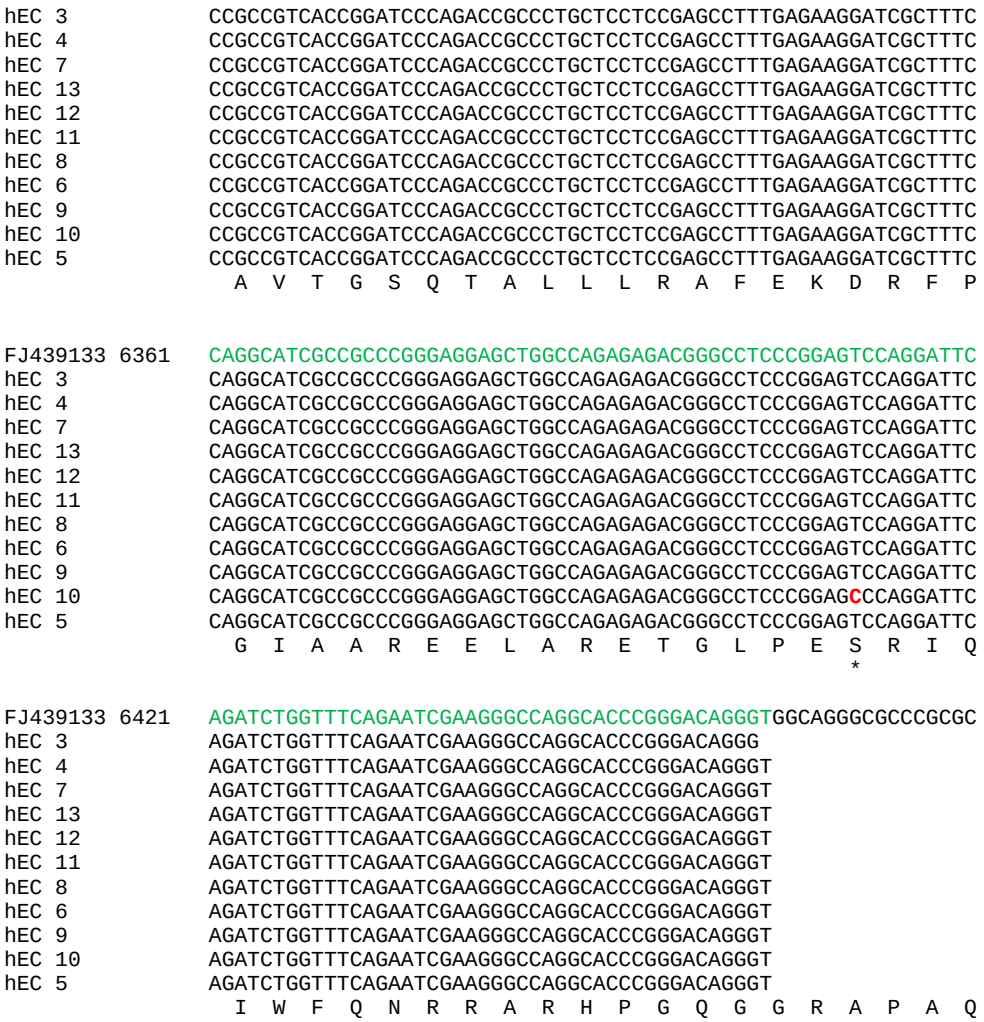


Figure 4.11 Alignment of the RT-PCR products using the primer pair 1797/2235

DNA sequences were edited manually to correct base-calling errors. Primer sequences were trimmer and the sequences aligned using Clustal. Sequences are compared to the equivalent region of the most distal D4Z4 repeat unit in the FJ439133 (GenBank) reference sequence. The homeobox sequences are highlighted in green. Three sequences have a single base change from FJ439133, these are highlighted with an asterisk.

4.2.8 Transcription outside of the DUX4 ORF

In order investigate whether transcription was restricted to the predicted ORF, primers were designed to amplify fragments upstream of the ORF but within the repeat unit (Figure 4.12a). Using a Onestep RT-PCR on hEC RNA a transcript was amplified with the 4q7 primer pair at the 5' end of the repeat (Figure 4.12b), but not with the 4q8 primer pair immediately upstream of the ORF. The 4q7 fragment was cloned and sequenced to confirm it originated from the repeat. No transcripts were amplified from myoblast or myotube RNA (Figure 4.12b).

Although some potential promoter regions upstream of the D4Z4 array have been suggested, the region has not been thoroughly investigated for control sequences. The TACAA promoter region identified by Dixit *et al.* (2007) is 5822bp downstream of this probe region however it is possible that there is a promoter sequence near to the probe region which could account for this transcription.

4.2.9 Amplification of transcripts from the most distal D4Z4 repeat

Snider *et al.* used a nested RT-PCR to amplify full-length DUX4 transcripts that originated from the most distal D4Z4 repeat using RNA from testis and FSHD muscle biopsies. They also identified a smaller transcript (DUX4-s, Figure 4.1b) in some of the FSHD muscle samples and in all of the control muscle samples. Thus, these nested conditions were used on RNA extracted from the three cultured myoblast cell lines established in this laboratory, primer positions are shown in Figure 4.13a.

No full length (DUX4-fl, Figure 4.13b, c) transcripts were amplified from the myoblast cell lines, in either proliferating or differentiated states. Smaller products were amplified from RNA extracted from the GM17869 (Figure 4.13b) and GM17940 (Figure 4.13c) cell lines after differentiation into myotubes. However, these products appear to be too small to represent the DUX4-s transcripts. Despite a number of attempts, cloning the fragments was unsuccessful and therefore it was not possible to confirm their origin. No amplification was

seen when the nested RT-PCRs were repeated on RNA that was extracted from the hEC cell line (Figure 4.13d).

As Snider *et al.* had noted expression from two germ cell tumour lines (Snider *et al.*, 2010); the nested RT-PCR was repeated on RNA extracted from the germ cell tumour line GCT27 (donated by Dr. Paul Scotting). Both the nested RT-PCR and the haplotyping of this cell line were performed by Amy Prosser.

In muscle cells, Snider *et al.* had been able to amplify the full length transcript in myoblasts only from those that contained a 4qA allele, however, they were able to amplify transcripts from 4qA, 4qB and 10q alleles from testis RNA. They did not provide information on the chromosomal origin of transcripts from the germ cell tumour lines (Snider *et al.*, 2010). Using the SSLP genotyping reaction described in Section 4.2.2.i, two 161 size alleles, one 159 size allele and one 166 size allele were detected in the GCT27 cell line. Thus, the two alleles on chromosome 4 are likely to be 4qA159 and 4qA161. The nested RT-PCR reaction on the GCT27 cell line was performed using primers 14A and 174 for the primary reaction and 15A and 175 for the secondary reaction. A product of a similar size to the DUX4-fl transcript was amplified (Figure 4.14a); this was cloned and sequenced to confirm its origin.

A summary of the RT-PCRs performed on the D4Z4 array is provided in Table 4.6.

4.2.10 Analysis of DUX4 transcripts from the GCT27 cell line

The product amplified from the GCT27 cell line was cloned into pGEM-T Easy and 16 clones were sequenced. Of these, 8 were sequenced across the full length of the transcript. These sequences were compared with GenBank sequence FJ439133 and base changes from this consensus are shown in Figure 4.14b. Variants that only appear in a single clone have not been included in the analysis as it cannot be excluded that these are PCR errors.

On the basis of the changes identified, 11 of the sequences can be split into 3 groups (Figure 4.14b). The remaining 5 have failed to sequence through the whole of the product and are missing sequence information at the variant positions used to group the previous 11 sequences. A 10bp deletion of the sequence GTACCAGCAG was identified in 6 of the clones.

This deletion is located immediately proximal to the start of the second intron (See Figure 4.14c), so is likely to be variation in splice site usage. Of the clones with sequence data in this region, all those that did not have the 10bp deletion contained a C-G change at position 7322 within the first intron, which is also present on the 4qB163 allele in the HHW419 cell line. Thus, none of these sequences were identical to the FJ439133 sequence, which is from a 4qA161 allele and are therefore likely to originate from one of the other GCT27 alleles. It is possible that the remaining 5 sequences that could not be grouped match the FJ439133 4qA161 sequence, these will require re-sequencing.

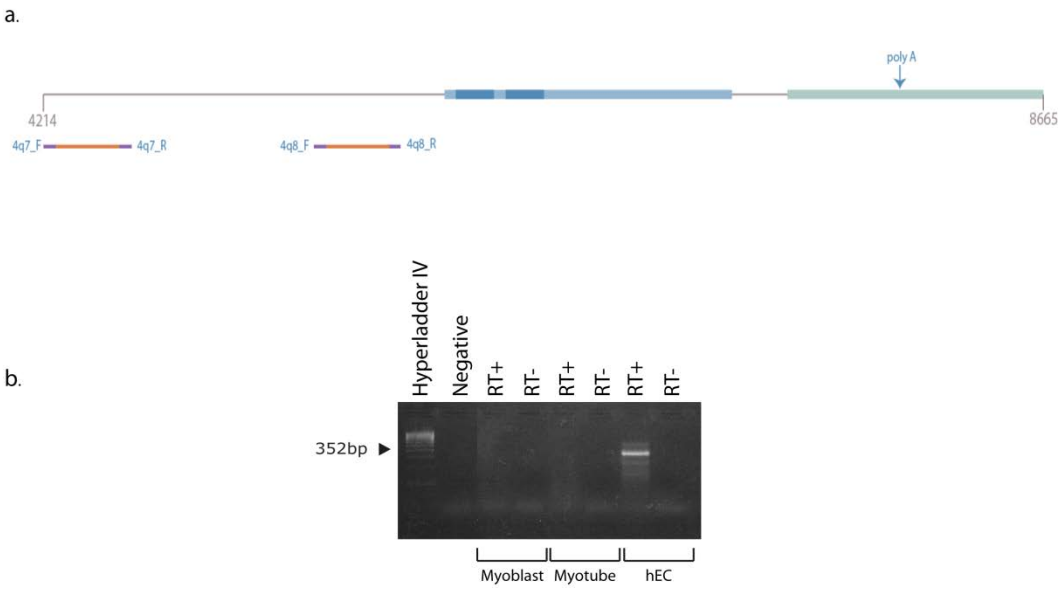


Figure 4.12 Expression from outside of the DUX4 ORF

a) Position of primers to amplify within the D4Z4 repeat, outside of the ORF. The ORF is shown in pale blue with the homeodomains highlighted in dark blue. pLAM is shown in green. b) Amplification of 4q7 transcripts by RT-PCR from RNA extracted from 17869 myoblast, myotube and hEC cell lines.

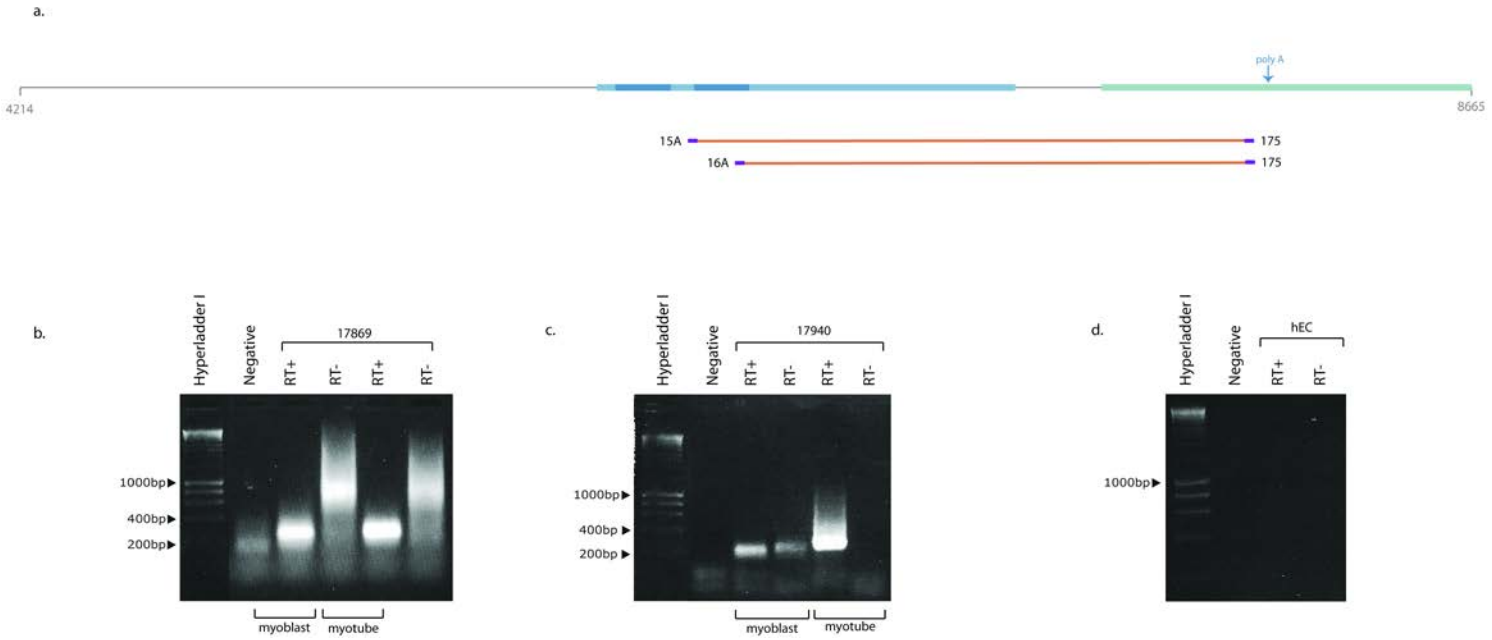
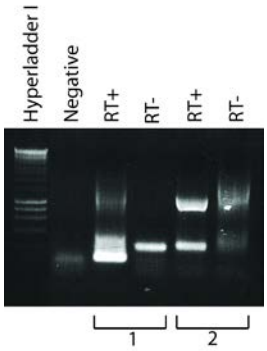


Figure 4.13 Attempts to amplify transcripts from the most distal D4Z4 repeat

a) The DUX4 ORF is shown in pale blue with homeodomains highlighted in dark blue. The pLAM region is shown in green. The positions of primers used in the final round of the nested PCR are shown underneath. RT-PCR results are shown for the primer pair 15A and 175, which amplify either a 1263bp fragment (both introns spliced out) or a 1399bp fragment (only intron 2 is spliced out). Results for the primer pair 16A and 175 were similar. b) Nested RT-PCR on RNA extracted from GM17869 at both proliferating and differentiated (myotube) stages. c) Nested RT-PCR on RNA extracted from GM17940 at both proliferating and differentiated (myotube) stages. d) Nested RT-PCR on RNA extracted from the hEC cell line.

a



b

Variant Group	Position	6690	7071	7322	7507
	FJ439133	C	T	C	GTACCAGCAG
1	GCT27 1	T	C	G	GTACCAGCAG
	GCT27 2	T	C	G	GTACCAGCAG
2	GCT27 3	C	T	C	del
	GCT27 4	C	T	C	del
	GCT27 5	C	T	C	del
	GCT27 6	C	T	C	del
	GCT27 8	C	T	C	del
	GCT27 22	nd	T	C	del
3	GCT27 12	nd	T	G	GTACCAGCAG
	GCT27 16	C	T	G	GTACCAGCAG
	GCT27 10	C	T	G	nd
Ungrouped	GCT27 17	C	nd	nd	nd
	GCT27 18	C	T	nd	nd
	GCT27 19	C	nd	nd	nd
	GCT27 23	C	nd	nd	nd
	GCT27 11	nd	nd	nd	nd

c

```
FJ439133  1541  GCCCAGGTACCAGCAGGTGGGCCGCCTACTGCGCACGCGCGGGTTTGCGGGCAGCCGCCT
GCT27 16        GCCCAGGTACCAGCAG::::::::::::::::::::::::::::::::::::::::::::
GCT27 5         GCCCAG::::::::::::::::::::::::::::::::::::::::::::::::::::::
                 P  G  T  S  R  W  A  A  Y  C  A  R  A  G  L  R  A  A  A  W
```

Figure 4.14 Analysis of GCT27 transcripts

a) Nested RT-PCR of RNA extracted from the GCT27 cell line (performed by Amy Prosser). b) Table to show base changes from the FJ439133 consensus sequence (highlighted in orange). Sequences are split into three groups based on these changes. nd = no data. Sequences without data at the variant positions could not be grouped. c) Sequence alignment of GCT27 16 (without 10bp deletion) and GCT27 5 (with 10bp deletion). FJ439133 exon sequence is highlighted in blue and the intron sequence is highlighted in green. Both sequences end with the splice donor signal, AG, so the deletion is likely due to a variation in splice site usage.

			Myoblasts			Myotubes								
Primers	**RT-PCR reagents**	**hEC**	**17731**	**17869**	**17940**	**17731**	**17869**	**17940**	**RMS13**	**RD**	**Fskm**	**Testis**	**ABr**	**GCT27**
HDUX4	Plat Taq	x	x											
	Improm/Kapa							x				x		
	Onestep	✓✓✓	x	x	x	x	x	xxx	xx	xx	xxxx	xx		
4q7	Biomix				x									
	Onestep	✓✓	x			x								
4q8	Onestep	x												
1797/1906	Plat Taq	✓✓	xx	x	x								xx	
1797/2235	Plat Taq	✓✓	xxx	x	x								x	
2672/2952	Plat Taq	x												
222/219	Improm/Kapa		x	x	x	x	x	x	xx	xx		xxx		
222/407	Plat Taq	x	x	x			xx			x				
	Improm/Kapa		x	x	x	x	x	x	x	x		x		
15A/175	Plat Taq	xx	x	xxx	xx		xx	x				xxx		x✓xx

Table 4.6 RT-PCRs for amplification of DUX4

A positive amplification is indicated by a ✓ and a negative amplification is indicated with an x. Orange boxes show where amplification from a cell line/primer pair combination was successful and blue boxes indicate unsuccessful amplification attempts. Where there was no amplification attempt boxes are left blank. Fskm = Fetal skeletal muscle. ABr = Adult Brain.

4.2.11 Protein expression from DUX4

Proteins were extracted from the patient myoblast cell lines in order to test for DUX4 expression as previously reported by Dixit *et al.* (2007). Western blotting was carried out on these protein extracts using the 9A12 antibody created by Dixit *et al*. (kindly provided by Prof Alexandra Belayew). As a control for antibody binding, TE671 cells were transfected with a construct overexpressing DUX4. The DUX4 protein runs at approximately 60KDa, in contrast to published data which shows the protein at 52kDa (Dixit *et al.*, 2007).

No bands of the same size as the DUX4 protein seen in the transfected cells were seen in any of the myoblast cell lines (Figure 4.15). A number of other bands are seen which are likely due to non-specific binding on the 9A12 antibody. As the antibody does not seem to be specific to DUX4 it is possible that the protein band identified by Dixit *et al.* is also due to nonspecific binding, especially since the DUX4 protein in the transfected cells appears to run at a different size to that reported by Dixit *et al.* Proteomic analysis of proteins identified by this antibody will be important in order to confirm they are *bone fide* DUX4 proteins.

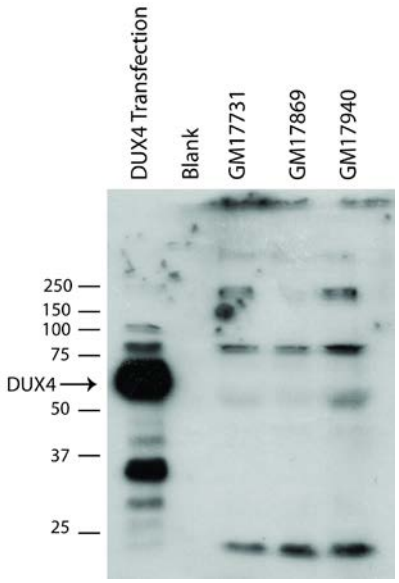


Figure 4.15 Western Blot of myoblast cell lines

Western blotting of TE671 cells transfected with a construct overexpressing DUX4 (lane 1) and myoblast cell lines (lines 3, 4 and 5) using the 9A12 antibody (Dixit *et al.*, 2007) (1:1000 dilution) followed by a Goat anti-Rabbit IgG (1:100000 dilution). The DUX4 protein in the transfected cells is indicated at around 60kDa, there is a smaller, strong band at around 35kD which is likely to be due to protein degradation. Additional bands are likely due to non-specific binding of the 9A12 antibody. No DUX4 protein was identified in the three myoblast cell lines (GM17731, GM17869 and GM17940).

4.2.12 Expression of gene loci proximal to the D4Z4 array

A number of studies have compared expression of sequences upstream of the D4Z4 array in FSHD and control cells (for a summary see Table 1.1). The data produced is conflicting. In order to investigate expression from this region in the cell lines established in this lab, RT-PCR was performed with primers that amplify the DUX4C and FRG2 ORFs. If this region has a more open chromatin structure in FSHD and germ-line cells then there could potentially be non-coding transcription in the region. In order to test for transcription from regions that do not contain a functional ORF, primers for the p13E-11 probe region were also used for RT-PCR. Primer positions and sequences are shown in Figure 4.16a and Table 2.5 respectively. RNA extracted from hEC cells and myoblast cell lines in both their proliferating and differentiated (myotube) stages, was used for RT-PCR, as described in Section 2.4.1.

A summary of the RT-PCRs performed on sequences upstream of the D4Z4 array is provided in Table 4.7.

4.2.12.i p13E-11

Using primers for the p13E-11 probe region, transcripts were amplified from hEC RNA using the OneStep kit (Figure 4.16d) but were not detected when using a two-step RT-PCR. No amplification for this locus was seen from RNA extracted from any of the myoblast cell lines.

Although some potential promoter regions upstream of the D4Z4 array have been suggested, the region has not been thoroughly investigated for control sequences and so it's possible there is a promoter sequence near to the probe region which could account for this transcription. The hEC cell line is thought to recapitulate many of the events that occur early in embryogenesis, when chromatin is more open and expression from CpG promoters is more likely to occur (Almstrup *et al.*, 2004; Andrews *et al.*, 2005; Schwartz *et al.*, 2005).

4.2.12.ii DUX4C

It has been reported that low levels of *DUX4C* expression can be detected in both control and FSHD myoblasts, in both proliferating and differentiated stages (Ansseau *et al.*, 2009). However, no expression of *DUX4C* was detected in RNA extracted from any of the cell lines tested, despite following the published reaction conditions (Figure 4.16b). If transcription from *DUX4C* does occur in these cell lines, the difficulty in amplifying the transcripts suggests that they have a low abundance, or are degraded rapidly after production.

4.2.12.iii FRG2

Rijkers *et al.* amplified *FRG2* mRNA from differentiating myoblast cultures and these transcripts were sequenced and shown to originate predominantly from chromosome 10 (Rijkers *et al.*, 2004). Expression of *FRG2* sequences has also been detected in some non-FSHD myopathies; however, these were derived from either chromosomes 3 or 22 (Rijkers *et al.*, 2004), which have identical sequences over this region.

For this study, reverse transcription was carried out using the Superscript III reverse transcriptase with oligo dT primers and then amplified using Biotaq polymerase and PCR primers EX2BF and SSC8 (according to Rijkers *et al.*, 2004). Expression of *FRG2* was detected in RNA from all cell lines investigated (Figure 4.16c)

The RT-PCR products from the GM17731 myoblast RNA were cloned and sequenced. To determine the chromosome of origin, these sequences were then compared with *FRG2* mRNA sequences in GenBank (Chromosome 3 or 22, NM_001124759.1; Chromosome 4, AY714545.1 and Chromosome 10, NM_001080998.1). Of the 18 transcripts sequenced, 8 originated from chromosome 10, 7 from chromosomes 3 or 22 and 3 from chromosome 4 (Figure 4.17) . In agreement with published data, a higher number of transcripts originated from chromosome 10 than chromosome 4. However, Rijkers *et al.* only identified transcripts from chromosomes 3 or 22 in patients with a non-FSHD myopathy, while this study amplified these sequences

from FSHD myoblast cells. As most of the transcripts originated from chromosomes other than 4, it is unlikely these play a causative role in FSHD.

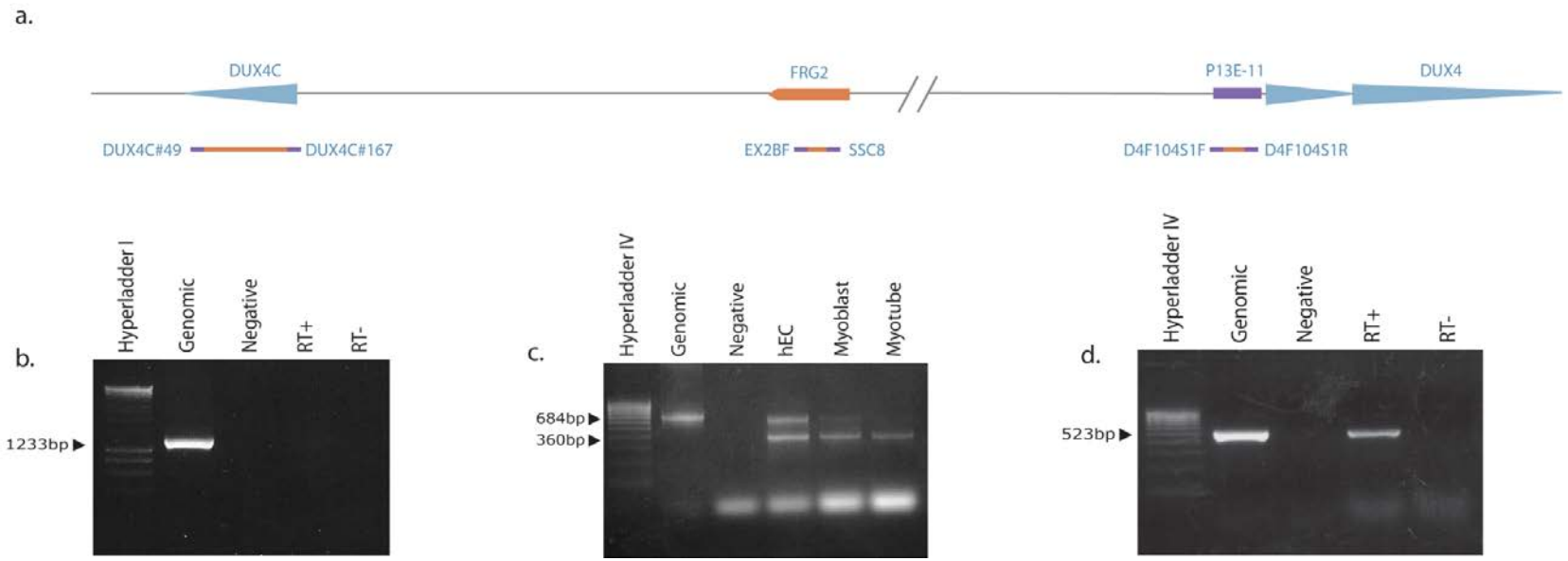


Figure 4.16 Expression of sequences proximal to D4Z4

a) Schematic representation of the products expected by RT-PCRs of loci upstream of the D4Z4 array. DUX4 repeats are indicated with blue arrows and DUX4C is represented by an inverted arrow upstream of the array. The p13E-11 probe region is shown by a purple block. The FRG2 gene is shown in orange with the arrowhead indicating the direction of transcription. Primer sequences are provided in table 2.5. b) RT-PCR of the DUX4C ORF on RNA extracted from the hEC cells. c) RT-PCR of the FRG2 ORF in RNA extracted from hEC cells and from the myoblast cell line (GM17731) in both its proliferating and differentiated myotube form. d) RT-PCR of the p13E-11 region from RNA extracted from hEC cells.

		Myoblasts			Myotubes			
	hEC	**17731**	**17869**	**17940**	**17731**	**17869**	**17940**	**Fskm**
P13E11								
Plat Taq	x							
Onestep	✓✓	xx					x	
Dux4C								
Biomix	x							
Plat Taq	x				x	x	x	
Onestep								
FRG2								
Plat Taq								
Superscript/kapa		x						
Superscript/biotaq	✓x✓	x✓✓✓	✓✓	✓	x✓	✓✓	✓	xxx
Onestep								
FRG1								
Abgene		✓	✓	✓				
Biomix				✓	✓			

Table 4.7 RT-PCRs for amplification of upstream sequences

A positive amplification is indicated by a ✓ and a negative amplification is indicated with an x. Orange boxes show where amplification from a cell line/primer pair combination was successful and blue boxes indicate unsuccessful amplification attempts. Fskm = Fetal skeletal muscle. ABr = Adult Brain.

```
AH550    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCCGACATAGCTCGCACA 60
AH552    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCCGACATAGCTCGCACA 60
AH583    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCCGACATAGCTCGCACA 60
AH590    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCCGACATAGCTCGCACA 60
AH526    ------------------TGGGCTAGGTCTTGATAAACAGCCTCCGACATAGCTCGCACA 42
AH587    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH555    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH591    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH589    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH586    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH580    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH551    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH582    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH581    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH553    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH588    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCCTCTGACATAGCTCGCACA 60
AH584    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCTTCCGACATAGCTCGCACA 60
AH585    ATCTGCTGTGCCCACACCTGGGCTAGGTCTTGATAAACAGCTTCCGACATAGCTCGCACA 60
                                                  *  *

AH550    GAGGTCACCAAGCTTTTTCGAAGTGACGGTGTTTGGACTTCTAGGGCCCGAGACCTATGC 120
AH552    GAGGTCACCAAGCCTTTTCGAAGTGACGGTGTTTGGACTTCTAGGGCCCGAGACCTATGC 120
AH583    GAGGTCACCAAGCTTTTTCGAAGTGACGGTGTTTGGACTTCTAGGGCCCGAGACCTATGC 120
AH590    GAGGTCACCAAGCTTTTTCGAAGTGACGGTGTTTGGACTTCTAGGGCCCGAGACCTATGC 120
AH526    GAGGTCACCAAGCTTTTTCGAAGTGACGGTGTTTGGACTTCTAGGGCCCGAGACCTATGC 102
AH587    GAGGTCGCCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH555    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH591    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH589    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH586    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH580    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH551    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH582    GAGGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH581    GAAGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH553    GAAGTCACCAAGCTTTTTCGAATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH588    GAAGTCACCAAGCTTTTTCGGATTGACGGTGTTTGGACTCCTAGGGCCCGAGACCTATGC 120
AH584    GAGGTCACCAAGCTTTTTCGAAGTGACGGTGTTTGGACTCCTAGAGCCTGAGACT-ATGC 119
AH585    GAGGTCACCAAGCTTTTTCGAAGTGACGGTGTTTGGACTCCTAGAGCCTGAGACC-ATGC 119
           *   *       *      * *                 *    *   *     **

AH550    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACTTCCCCTATGGTGGACATCACAGGTCTCC 180
AH552    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACTTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH583    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACTTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH590    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACTTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH526    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACTTCCCCTATGGTGGGCATCACAGGTCTCC 162
AH587    CGCCTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH555    CGCTTGCTGCGCCCAGTGCAAGCCCCGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH591    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH589    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH586    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH580    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH551    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH582    CGCTTGCTGCGCCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH581    CGCTTGCTGTGCCCAGTGCAAGCCCTGGAATGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH553    CGCTTGCTGTGCCCAGTGCAAGCCCTGGAATGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH588    CGCTTGCTGTGCCCAGTGCAAGCCCTGGAATGTCCCCTATGGTGGGCATCACAGGTCTCC 180
AH584    TGCTTGCTGCACCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCACCACAGGTCTCC 179
AH585    TGCTTGCTGCACCCAGTGCAAGCCCTGGAACGTCCCCTATGGTGGGCACCACAGGTCTCC 179
         *  *     **              *    **             *  *
```

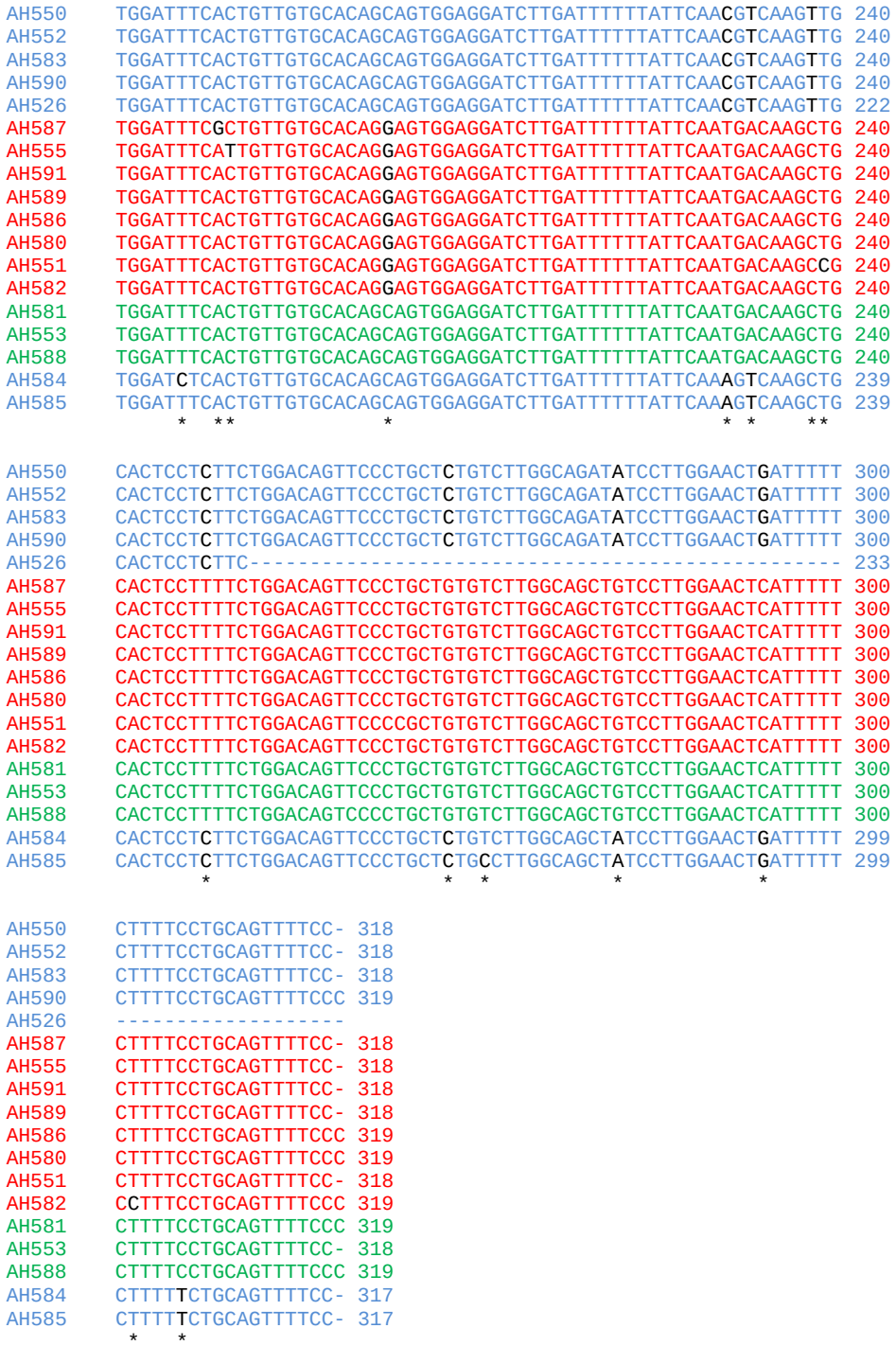


Figure 4.17 Clustal alignment of *FRG2* sequences

Clustal alignment of *FRG2* sequences. Identical bases are highlighted with an asterisk. Sequence positions are indicated on the right hand side. Chromosome 3 or 22 sequences are highlighted in blue. Chromosome 10 sequences are highlighted in red. Chromosome 4 sequences are highlighted in green.

4.3 Discussion

Despite the identification of transcription from the distal D4Z4 repeat unit, there has been no detailed information about whether internal repeats are expressed. The work described here provides evidence that multiple repeats are expressed from a human embryonal carcinoma cell line. In addition, this data supports the hypothesis of a role for normal transcription from the D4Z4 array in the germ-cell lineage.

4.3.1 Analysis of myoblast cell lines

Patient myoblast cell lines (GM17731, GM17869 and G17940) were established in the laboratory and expression of myoblast/myotube specific markers was used to verify the cultures. It was confirmed that at early passages, the cultures consist almost entirely of myoblast cells (determined by their expression of desmin) and so further analyses using these cell lines (see chapter 4) were carried out on cultures with a passage number of 5 or below. If these cells were to be used at later passages it would be necessary to repeat the analyses and determine if the percentage of myoblasts in the culture was still sufficient.

On differentiation of the patient cell lines, the fusion index was calculated after 6 days in differentiation media and ranged from 46%-83%, depending on the cell line. When analysing the results of further studies using the differentiated cells (see chapter 4), it will be important to consider the fact that the data will originate from a mixture of myoblasts and myotubes.

The three patient cell lines were haplotyped in order to confirm that they carry the permissive 4q161 allele that was identified by Lemmers *et al.* The GM17869 and GM17940 cell lines were haplotyped using sequence information from the P13E-11 and SSLP regions. With this technique, PCR and cloning bias may have meant that the complete set of haplotypes was not sampled. It is also possible that polymerase slippage during amplification of the microsatellite repeat at the SSLP region could also have caused a variation in copy number. Any variation in this region would have resulted in the incorrect assignment of a haplotype. The GM17731 cell line was analysed using the SSLP genotyping reaction described

in Lemmers *et al*. (2010). The results show two allele sizes, of 161 and 166, and suggest that there is an equal number of each size in the cell line. 4A161 is the most prevalent chromosome 4 allele in most populations (38% Europe, 40% CEU, YRT 20%) (Lemmers *et al.*, 2010b) while 10A166 is the most prevalent chromosome 10 allele in all populations that have been studied so far (Europe 86%, CEU 86%, YRT 77%) (Lemmers *et al.*, 2010b). The data would therefore predict that the GM17731 cell line contains two 4A161 and two 10A166 alleles. The finding that all of the cell lines carried a 4A161 allele was to be expected as it has previously been shown that at least one copy of this allele is required for FSHD pathogenicity.

4.3.2 Expression of upstream sequences

A number of studies have looked at expression of candidate genes upstream of the D4Z4 array and the results have been conflicting (see Table 1.1). In order to investigate expression of upstream sequences in the cell lines established in this lab, primers were designed to amplify from the p13E-11, FRG2 and DUX4c regions.

Two previous studies had been unable to amplify *DUX4C* transcripts (Alexiadis *et al.*, 2007; Osborne *et al.*, 2007). In 2009, Ansseau *et al.* identified low levels of mRNA expression of *DUX4C* in both patients and control cells and saw an increase in *DUX4C* protein in FSHD cells by western blot. In the work described here, the primers and conditions described by Ansseau *et al.* were repeated. However, no transcripts were amplified from RNA extracted from the three myoblast cell lines, or from the hEC cell line. If transcription from *DUX4C* does occur in these cell lines, the difficulty in amplifying the transcripts suggests that they have a low abundance, or are degraded rapidly after production. Some FSHD patients have been identified that have a deletion including the *DUX4C* repeat, which makes its involvement in the disease unlikely.

An increase in *FRG2* expression in FSHD cells compared with control cells was seen by both Gabellini *et al.* (2002) and Rijkers *et al.* (2004). However, in 2007 a microarray study by Osbourne *et al.* saw no significant change. In this work, expression of *FRG2* mRNA was detected in all of the three myoblast cell lines and in the hEC cell line. As the study was

restricted to FSHD myoblasts no interpretations about level of expression can be made. As *FRG2* has homologous copies on many other chromosomes, the amplified transcripts from GM17731 were sequenced to confirm their origin. Of the transcripts amplified, 44% originated from chromosome 10, 39% from chromosomes 3/22 and only 17% from chromosome 4. The region around *FRG2* is known to be deleted in some FSHD patients, and as the majority of expression comes from chromosomes other than 4, it is unlikely that these transcripts play a causative role in FSHD. It remains possible; however, that contraction of the array on chromosome 4 may have trans-acting effects on the expression of *FRG2* on these other chromosomes.

In this work, primers used to amplify the p13E-11 probe region were used in an RT-PCR on RNA extracted from hEC cells, and the myoblast cell lines. No transcripts were amplified from the myoblast cell lines, however expression from this region was detected in the hEC cell line. The hEC cell line is thought to recapitulate many of the events that occur early in embryogenesis, when chromatin is more open and CpG promoters are thought to be active (Almstrup *et al.*, 2004; Andrews *et al.*, 2005; Schwartz *et al.*, 2005). It is possible that the region proximal to the D4Z4 array has an open chromatin structure at this time, and that (presumably) non-coding transcription can occur. While the work described here was being carried out, genomic D4Z4 sequences incorporating this proximal region were amplified from FSHD and control alleles and sequenced by Joanne Pollington. Comparison of disease permissive and non-permissive alleles did not identify any polymorphisms that were specific to the FSHD permissive alleles so it seems unlikely that any transcription from this region plays an important role in the FSHD pathology.

As it seemed unlikely that any of these regions had a causative role in the FSHD phenotype, and as more expression data from DUX4 was being published, no further studies of expression from the upstream regions were performed.

4.3.3 Expression from the D4Z4 array in myoblast cell lines

While the work in this chapter was being carried out, transcription from the final repeat of the D4Z4 array was confirmed and an explanation for the haplotype specific nature of the disease was published. Lemmers *et al.* (2010a) identified a polymorphism in the pLAM region that creates a functional polyadenylation signal on disease permissive chromosomes. Non-permissive haplotypes have polymorphisms that deem this signal non-functional. The pLAM region is absent on 4qB chromosomes, which explains why these haplotypes are non-permissive. The group hypothesised that the polyA signal stabilises transcripts from the last repeat and that expression of these transcripts causes the FSHD phenotype.

In order to confirm their hypothesis, a functional polyA signal was added to non-permissive haplotypes and vice versa, the resulting constructs were transfected into C2C12 cells and the stability of the transcript was assessed by northern blot. Transcripts were only detected from haplotypes which contained a functional polyA signal; the use of the signal was confirmed with 3'RACE (Lemmers *et al.*, 2010a).

Lemmers *et al*. (2010a) also analysed a number of FSHD families with non-standard chromosomes. One of these families had the permissive polyA end on a chromosome 10 background, which shows that it is not the chromosome background which affects the permissiveness of the haplotype, but the sequence at the distal end of the array. This finding confirms the role of DUX4 transcripts in FSHD and provides further evidence that the upstream candidate genes do not play a causative role in the disease.

Snider *et al.* (2010) were able to amplify full length DUX4 transcripts from this final repeat in cell lines with alleles carrying the functional polyA signal. In cell lines with non-permissive alleles only smaller fragments could be amplified. In order to amplify transcripts, the group had to use a nested RT-PCR protocol which indicates a very low abundance of transcripts in the cell lines.

In this work, a variety of reverse transcription and amplification reagents and conditions have been used on RNA extracted from the FSHD myoblast cell lines, including a repeat of the nested RT-PCR conditions described in the Snider *et al.* (2010) paper. No amplification from

the myoblast cell lines was seen. The difficulty in amplifying transcripts from D4Z4 suggests that expression levels are very low. It is possible that the myoblast culturing conditions may have affected expression of DUX4 and this could explain the differences between data from the cell lines established in this lab and those from the published expression studies.

4.3.4 Expression from the D4Z4 array in hEC cell lines

All of the RT-PCRs were repeated on RNA extracted from the hEC cell line. This cell line has been shown to have a global gene expression profile similar to that of undifferentiated human embryonic stem cells (Schwartz *et al.*, 2005) and recapitulates many of the events that occur early in embryogenesis (Almstrup *et al.*, 2004; Andrews *et al.*, 2005; Schwartz *et al.*, 2005).

Similar to the myoblast cell lines, full length mRNA transcripts could not be amplified from the hEC cells. It was possible, however, to amplify smaller fragments that covered the majority of the ORF (1161 out of 1275 bases). In the nested RT-PCR described by Snider *et al.* (2010) the reverse primer falls within the pLAM region, as this is not present in the 4qB163 alleles found in the hEC cells, it would not be expected to amplify a product from this cell line.

The largest of the amplified fragments covered 812bp from within the second homeodomain to 114bp short of the stop codon. Sequencing of these fragments identified 10 variant positions and a 6bp indel that allowed the sequences to be divided into 17 different variant groups. This data indicates that multiple repeats are expressed in the hEC cell line studied. The TSPY and USP17 genes are other examples of protein coding genes that are organised in a tandem array, and it has been shown that multiple copies of these repeats are also expressed (Manz *et al.*, 1993; Saitoh *et al.*, 2000).

Of the 14 variants identified in this 812bp product, 6 resulted in coding changes, none of which fell within the homeodomain. RT-PCR products amplified using the primer pair 1797/2235 (from Snider *et al.* 2009) cover both of the homeobox sequences, 11 of these products were sequenced and no variant positions were identified. As the homeoboxes code for DNA binding domains, there is likely to be strong selection against variation in these

sequences. Single base pair changes that occurred in only one of the sequences were not included in the analysis. This means that the variants identified are likely to be *bone fide* base changes, rather than PCR errors. It should be noted that the primers used were based on the consensus sequence for DUX4 and could therefore be selecting for specific sequences, which may have led to a bias in amplification.

Strand-specific RT-PCR identified transcripts from both strands of the DUX4 sequence. There did not appear to be a preference for expression from either strand although the number of transcripts sequenced may be too small to identify any bias. Antisense transcripts are thought to control the expression of their sense counterparts (Katayama *et al.*, 2005; Xu *et al.*, 2011). As the sense and antisense transcripts overlap, it is possible that the antisense transcript binds to the sense copy and prevents protein binding. This may act as a mechanism for repression of the D4Z4 array, controlling expression from the repeats and preventing translation of the sense mRNA into a functional protein.

The hEC cell line was haplotyped and shown to contain two 4qB163 alleles at the D4Z4 locus. As these cells do not contain any FSHD permissive alleles, the sequence data from this cell line cannot provide information about variants that may be involved in the FSHD phenotype. However, amplification of RT-PCR products from this cell line shows that transcription can occur from these non-permissive alleles, and as multiple D4Z4 alleles have been conserved, the sequences would be expected to have a normal function outside of their role in FSHD. Therefore, data from this cell line can provide information on the repeat variants that are expressed during this normal role. Snider *et al.* have reported expression of DUX4 transcripts in both testis and germ-cell tumour lines, the hEC cell and GCT27 cell data described in this chapter supports the hypothesis that transcription from D4Z4 occurs in the germ-line lineage.

Data from the expressed hEC cell transcripts was compared with genomic sequence data from the HHW419 cell line. This somatic cell hybrid line had previously been haplotyped and shown to contain a 4qB163 allele. Despite the sequences coming from different genomes, they share the same D4Z4 haplotype and so the sequences would be expected to be highly similar.

Of the 17 variant groups identified in the RT-PCR products, only 4 matched genomic sequences, and there were 17 genomic sequences without an equivalent expressed transcript. The lack of coverage of these genomic sequences in the hEC products may be due to variation in sequence between the two genomes. Alternatively, the sequenced hEC transcripts may not cover the full range of the expressed repeats. It is also possible that not all of the D4Z4 repeats in the array are expressed, or that some are degraded more rapidly than others.

RT-PCR of the polyadenylated fraction of hEC RNA also amplified the HDUX4 product. Although the internal repeats do not appear to have a functional poly (A) signal, the presence of transcripts in this poly (A) fraction warrants further investigation. In order to confirm that the transcripts identified in this fraction are polyadenylated an oligo dT primer will need to be used for the RT-PCR step, however this was not possible with the OneStep RT-PCR kit used for amplification of these products. Further investigation into the poly (A) signals used in this cell line by 3' RACE would give more information about the normal expression of sequences from the D4Z4 arrays on a B-type chromosome.

Not only could expression from the ORF be identified in the hEC cells, but expression upstream of the ORF, within each repeat, could also be identified. There are currently no recognised promoter sequences which could explain this expression, however, as the repeat sequences have not yet been fully characterised it is not possible to rule out the possibility of control sequences in some of the repeats that could account for these transcripts.

4.3.5 Expression of the DUX4 protein

Western blotting of protein extract from the myoblast cell lines did not identify any DUX4 protein. As there were a number of other bands in these myoblast lines, likely due to nonspecific binding, the antibody was not used for any further investigations. The DUX4 protein in transfected cells ran at around 60KDa, while published data shows a mass of 52KDa. As the antibody is not specific to DUX4 it will be important to perform proteomic

analysis of any proteins identified by this antibody in order to confirm they are *bone fide* DUX4 proteins.

Chapter 5. Analysis of mouse *Dux* expression

5.1 Introduction

In 2007, Clapp *et al.* identified an homologous Dux array in the murine lineages. In contrast to primates, the mouse array is not located at a telomere and the flanking genes are different to the human array. Thus, although they are unlikely to be true orthologues, this is the only DUX gene array in the mouse and so may be functionally equivalent to D4Z4.

Work in Jane Hewitt's laboratory has been sequencing this region in the mouse. In C57BL/6J mice, there are approximately 40 repeats that are split into 3 separate Dux clusters (Figure 5.1a). Pulse Field Gel Electrophoresis (PFGE) using a probe to the repeat shows a similar organisation for other inbred strains of mice, however, outbred CD1 mice show a more complex pattern (Clapp *et al.*, 2007; Figure 5.1b). The mouse repeat unit is 4899bp long, with an ORF of 2025bp. The double homeodomain region has 45% amino acid identity between the human and mouse, and 31% identity at the C-terminal domain (Figure 5.2). The difference in length of the ORF between the two species can be accounted for by a repeated 276bp region within the C-terminal region which is present in a single copy in humans while there are 5 copies in the mouse repeat (Figure 5.3). There is less sequence conservation between the mouse and human repeats outside of the ORF.

Clapp *et al.* published evidence for expression of at least some copies of the mouse Dux gene. Although they were unable to amplify the full ORF, overlapping primers were used to show that the whole ORF is transcribed (Figure 5.4a) and this was confirmed using RNA-FISH and *in situ* hybridisation. This difficulty in amplifying full length transcripts can also be seen in at the human D4Z4 array, however work described in the previous chapter has shown it is possible to amplify smaller fragments from hEC cell RNA that cover 1161/1275bp of the ORF. Clapp and colleagues were also able to amplify both sense and antisense transcripts from the Dux array by RT-PCR. However, in contrast to the apparently limited expression of *DUX4*

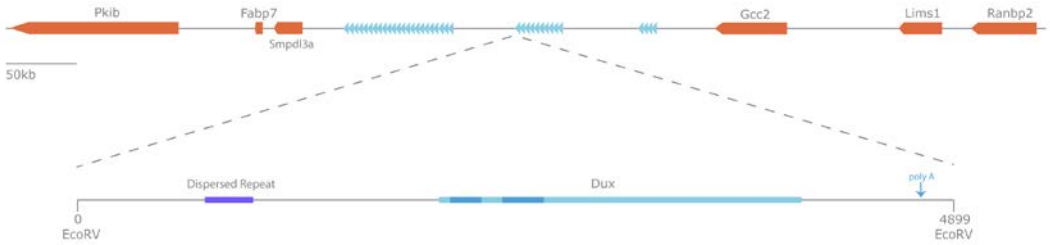
shown in human cell lines, expression in the mouse was identified in a number of different tissues, particularly the CNS (Figure 5.4b). Since the expression patterns seem to differ between species it will be important to establish whether human *DUX4* and the mouse *Dux* genes have equivalent functions.

Overexpression of DUX4 in cultured cells has been shown to have toxic effects (Gabriels *et al.*, 1999; Kowaljow *et al.*, 2007; Wallace *et al.*, 2011). In 2009, Bosnakovski *et al.* generated mouse Dux inducible C2C12 cell lines and used them to evaluate the toxicity of the mouse Dux protein. They found that 44% of cells were dead 24h after induction of Dux expression while another 30% showed characteristics of apoptosis (Bosnakovski *et al.*, 2009), and suggest that mouse Dux expression has a toxic effect on myoblasts similar to that seen with overexpression of DUX4 in TE671 cells (Kowaljow *et al.*, 2007). The group also showed that *Xenopus* embryos injected with mouse Dux at the 16 cell stage developed tail defects.

D4Z4 is GC-rich with a high frequency of CpG dinucleotides, around 290 in each repeat unit. In 2001, Tsien *et al.* used Southern blot analysis with several CpG methylation sensitive restriction enzymes to show that these sites are mostly methylated in the D4Z4 repeats in brain, lung, liver, heart and spleen. In 2003, Overveld *et al.* examined two methylation sensitive sites in the first proximal unit of the array and showed significant hypomethylation of D4Z4 in individuals with FSHD on the disease chromosome. Non-penetrant gene carriers were also hypomethylated on their contracted alleles (van Overveld *et al.*, 2003). In 2009, de Greef *et al.* used a CpoI digest to interrogate D4Z4 methylation on internal and proximal repeat units on 4q and 10q separately. They showed that hypomethylation of the repeats is observed below a threshold number of repeats, irrespective of the haplotype, and that a sharp increase in methylation levels is seen between 7 and 14 units. The level of hypomethylation on FSHD alleles was lower on the internal repeats (de Greef *et al.*, 2009).

In order to establish whether the mouse repeats have a similar methylation pattern to the human array, methylation sensitive restriction enzymes were used to investigate the methylation status of the mouse repeats.

a



b

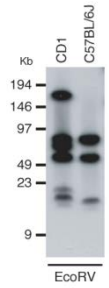


Figure 5.1 Mouse Dux array organisation

a) Schematic representation of the mouse Dux clusters and flanking genes. Flanking genes are shown in orange. There are three clusters of Dux repeats, which are indicated with a blue arrowhead. A representation of a single Dux repeat is shown underneath. There is a dispersed repeat upstream of the ORF, which is not found in the human repeat. The ORF is shown in pale blue with the homeodomains highlighted in dark blue. b) PFGE analysis of *Eco*RV-digested genomic DNA. The filter was hybridized with a ^{32}P-labeled Dux probe, was washed under high-stringency conditions, and was exposed for 6 h (Clapp *et al.* 2007).

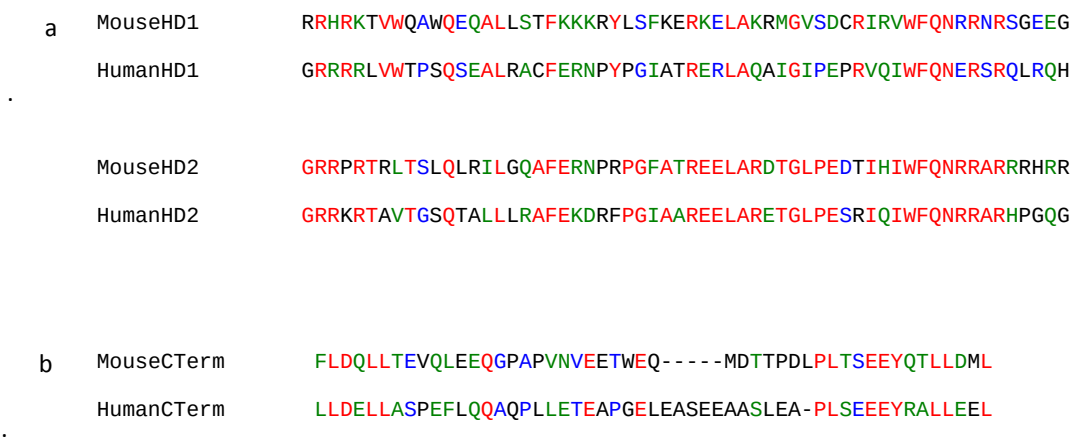


Figure 5.2 Clustal alignment of the mouse Dux and human DUX4 conserved regions

a) Alignment of the two homeodomain regions. The alignment shows several invariant or highly conserved amino acids. Identical residues are highlighted in red, amino acids with similar properties are highlighted in green and those with weakly similar properties in blue. b) Alignment of the C-terminal regions of mouse Dux and human DUX4 proteins. All sequences are in the 5'-3' orientation

a

b

```
MouseDux2       PGGRVHEGSQESFWPQE-EAGSTGMDTSSPSDSNSFCRESQPSQVAQPCGAGQEDARTQ
MouseDux4       PGGRVHEGSWESFWPQE-EAGSTGLDTSSPSDSNSFCRESQPSQVAQRRGAGQEDARTQ
MouseDux1       PEGREREGAQENLLPQE-EAGSTGMDTSSPSDLPSFCGESQPFQVAQPRGAGQQEAPTQ
MouseDux3       PGSREHEGSQDSLLPLE-EAVNSGMDTSIPSIWPTFCRESQPPQVAQPSGPGQAQAPTQ
MouseDux5       PGSMEHEGSQDSLLPLE-EAANSGRDTSIPSIWPAFCRKSQPPQVAQPSGPGQAQAPIQ
HumanDUX4       GVRGGWGRGGPQVAGAAWEPQAGAAPPPQPAPPDASASARQG-QMQGIPAPSQALQEPA
                                  *          *          *  *        *

MouseDux2       ADSTGPLELLLLDQLLDEVQ-KEEHVPVPLDWGRN 92
MouseDux4       ADSTGPLELLLFDQLLDEVQ-KEEHVPAPLDWGRN 92
MouseDux1       AGNAGSLEP-LLDQLLDEVQ-VEEPAPAPLNLDGD 91
MouseDux3       GGNTDPLE-LFLYQLLDEVQ-VEEHAPAPLNWDVD 91
MouseDux5       GGNTDPLE-LFLDQLLTEVQ-LEEQGPAPVNVEET 91
HumanDUX4       PWSALP-CGLLLDELLASPEFLQQAQPLLETEAPG 92
                       **          *
```

Figure 5.3 The mouse Dux ORF contains five copies of a repeat unit in the C-terminal region

a) Schematic representation of the five copies of the C-terminal repeated domains found in the mouse ORF. The ORF is shown in pale blue with the homeodomains highlighted in dark blue. The 5 repeats are indicated underneath. b) ClustalW alignment of the 5 mouse repeats and the single copy present in human DUX4. Sequences were edited manually to align the conserved LDQLL domain (highlighted). Invariant positions are indicated with an asterisk. All sequences are in the 5'-3' orientation

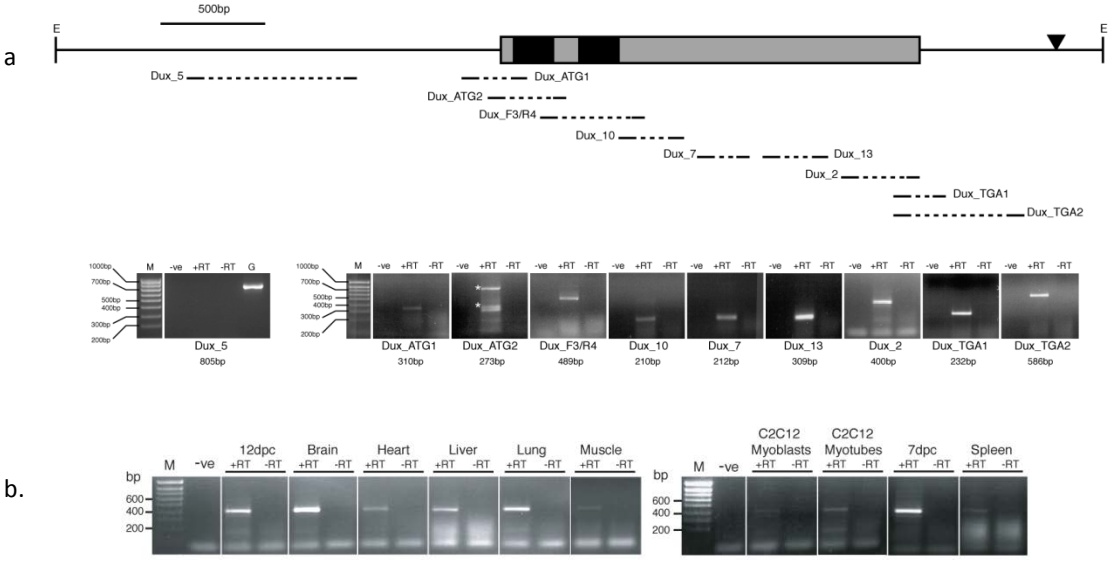


Figure 5.4 Evidence of transcription from the mouse Dux array. Taken from Clapp *et al.* (2007)

a) RT-PCR analysis of the mouse *Dux* repeat. Representative agarose gels of RT-PCR and genomic PCR products. M = molecular-weight ladder; -ve = no template; -RT = RNA added after inactivation of reverse transcriptase; +RT = RNA present throughout the OneStep reaction; G = genomic DNA template. The ORF is indicated by the gray rectangle, and the homeobox sequences by the black boxes. The putative polyA addition site is indicated by the black triangle. b) RT-PCR of mouse tissues by use of primers Dux_2f and Dux_2r, which should give a product of 400 bp. RT indicates a control reaction, where RNA was added after inactivation of the reverse transcriptase. dpc = days post coitum; C2C12 = mouse myoblast cell line. *Dux* transcripts were amplified from a range of tissues and embryonic stages. Detection of amplification was robust in the brain. In muscle cells (both in vivo and in vitro), amplification was weak but consistent. Sequencing of the products confirmed that they originated from the array. -ve = no template; -RT = control reaction where RNA was added after inactivation of the reverse transcriptase. +RT = RNA present throughout the OneStep reaction. (From Clapp *et al.* 2007.)

The main aim of this chapter was to investigate whether the sequence variation and expression pattern is similar between the mouse and human repeats. Sequence data has been gathered from the mouse repeat by RT-PCR and compared with genomic sequence data gathered by others in Jane Hewitt's laboratory to establish from which regions of the array the transcripts originate. In addition, to establish whether the mouse array resides in the same environment and has the same epigenetic markers as the human array, the methylation status of the repeats has been investigated.

5.2 Results

5.2.1 Amplification of mouse Dux transcripts

Mouse embryo RNA (Ambion) was treated with DnaseI and amplification was performed using the OneStep Reverse Transcription kit, as described in Section 2.13.2. Initial attempts to amplify across the full length of the ORF were unsuccessful. However, it was possible to amplify smaller fragments from the beginning (mATG primers) and end (mD4Z42 and mTGA primers) of the ORF, as well as a central region 25bp downstream of the second homeodomain (mDux10 primers; Figure 5.5a,b). Thogether, these cover 881bp of the Dux ORF as well as a further 63bp 5' and 470bp 3'. RT-PCR products were cloned into the pSMART vector and sequenced; analysis of the sequences is described in Section 5.2.2.

Evidence of production of antisense transcripts had been shown previously by Clapp *et al.* (2007). In order to confirm this and test whether antisense transcription occurred from the regions amplified in this study, a modified OneStep reaction was performed as described in Section 4.2.4. Antisense transcription was seen only with the mATG primer pair, not the mD4Z42 (Figure 5.6).

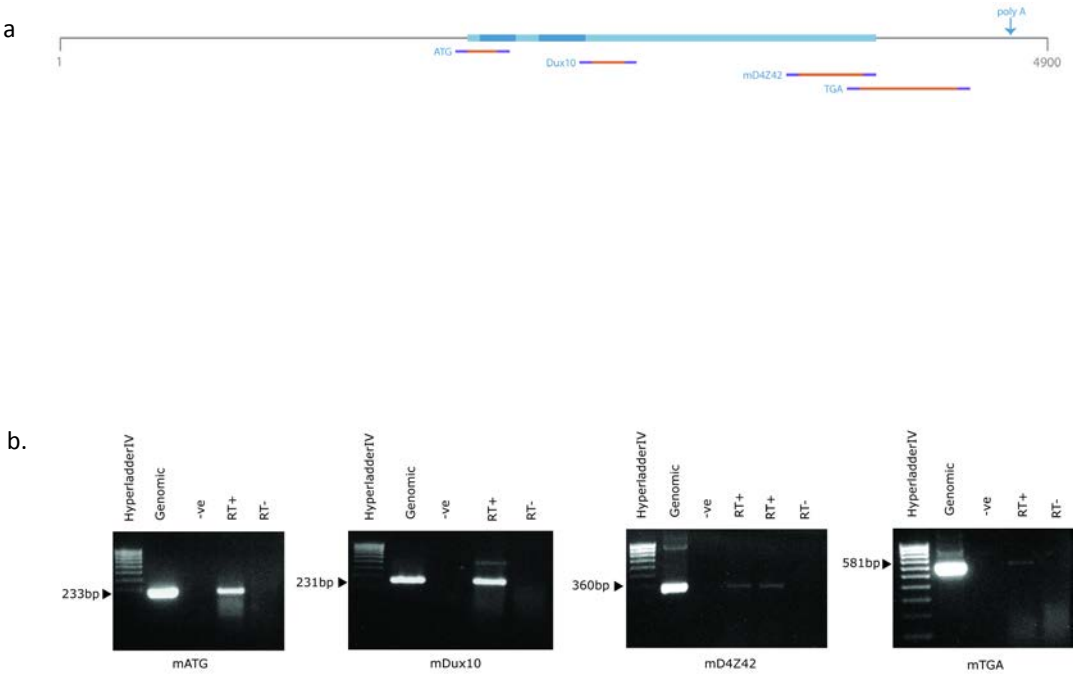


Figure 5.5 Amplification of mDux transcripts

a) Schematic representation of the mouse Dux fragments amplified by RT-PCR. b) Results of RT-PCR amplification using mATG, mDux10, mD4Z42 and mTGA primer pairs. The additional band in the mDux10 reaction was likely due to the primers annealing at multiple sites, the correct size band was extracted from the gel for sequencing

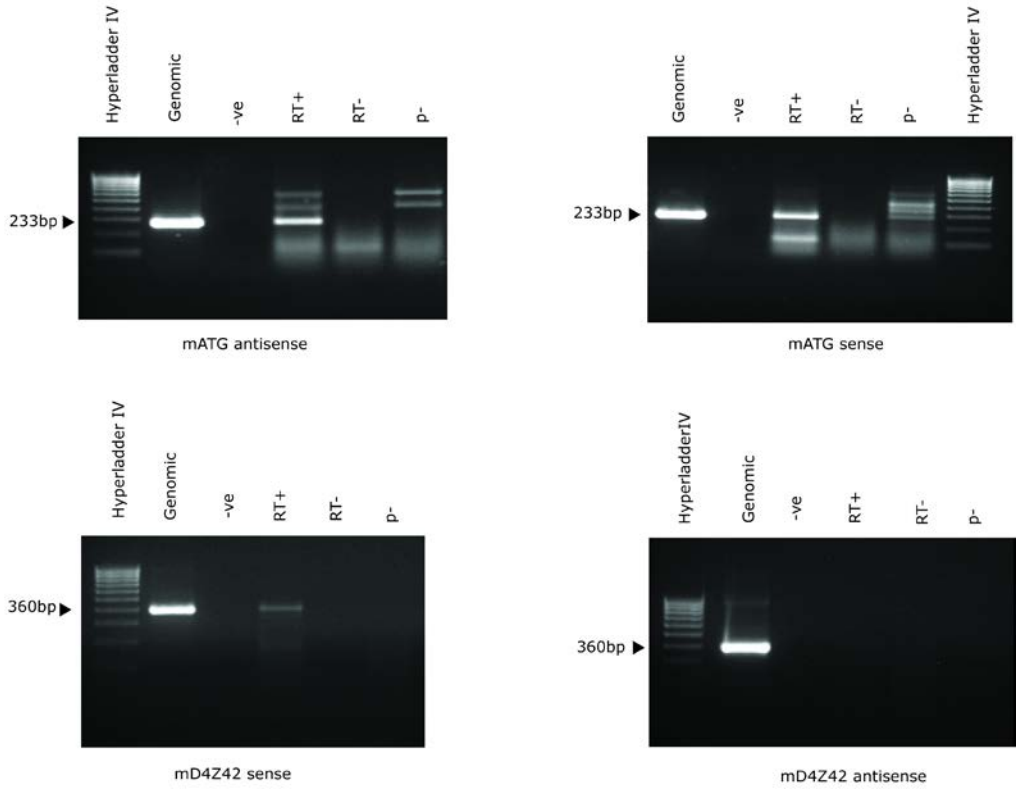


Figure 5.6 Strand-specific amplification of transcripts from the mDux array

Results of RT-PCR amplification using mATG and mD4Z4 primer pairs. In order to identify sense or antisense transcription only one of the primers was included in the reverse transcription step. Thus, for amplification of antisense transcripts, only the forward primer was included in the RT reaction. The p- controls did not have the second primer added. Additional bands in the mATG reactions were likely due to primers annealing at multiple sites, the correct size band was extracted from the gel for sequencing.

5.2.2 Sequence analysis of mouse Dux transcripts

As described in section 5.1, three closely linked but distinct Dux clusters had previously been identified in mice. The smallest of these clusters contains 2 full repeat units and 4 partial repeats, and is present in the fully sequenced BAC clone 142C15 (GenBank Accession number AC146701 Figure 5.7). Repeat units from the two larger clusters of the mouse Dux array had been amplified by PCR and sequenced by others in Jane Hewitt's laboratory and a consensus sequence had been compiled from these genomic sequences. The genomic sequence data and the 142C15 BAC clone contain sequence information from C57/Bl6 mice. The RNA used in this expression study was derived from Swiss Webster mice.

Initially, the clones containing the four RT-PCR products amplified during this work were compared to the genomic consensus sequence and any variant positions identified. Sequences that showed high similarity to the genomic consensus are included in Figure 5.8. For each primer pair there were several clones that had a significant number of variants compared to this consensus and showed much higher similarity to the repeats within the 142C15 BAC clone. Analysis of these sequences is described below.

5.2.2.i Sequence analysis of the mATG RT-PCR products

The mATG RT-PCR product covers 233bp around the predicted ATG start codon and includes 83bp at the start of the first homeobox. A total of 22 clones were sequenced and compared to the genomic consensus (Table 5.1). Five clones were identical to the consensus sequence and a further 11 clones contained only one or two substitutions and are likely to originate from the main Dux clusters. Four clones carry the same G>A transition resulting in an amino acid substation (S11N), while another six have the same R15L substitution. Neither of these substitution fall within the homeodomain.

Six of the clones contained the same 6 base changes from the consensus and when aligned with the BAC clone 142C15 were found to be identical to the first full repeat (repeat 3,

Figure 5.7) but all have a single nucleotide change from the other full repeat in this BAC. None of the variants identified in this RT-PCR product were present in the genomic sequences. Therefore, these clones are likely to represent transcripts from this genomic region.

a

b.

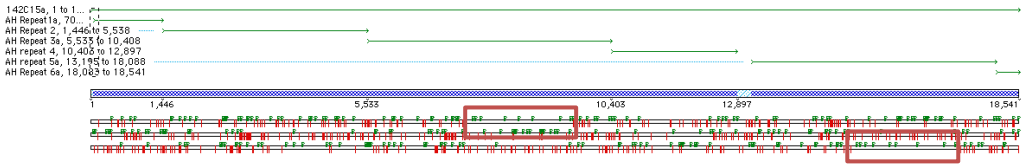


Figure 5.7 Schematic representation of the BAC clone 142C15

a) A schematic representation of the region of BAC clone 142C15 that contains mouse Dux repeats, indicated in orange. Regions 1 and 6 are partial; 3 and 5 are full length and contain complete Dux ORFs. However, repeats 2 and 4 lack full ORFs and are truncated to 4100bp and 2495bp respectively. b) A Sequencher alignment of the Dux repeats in the 142C15 BAC. Below the alignment is a schematic representation of the three coding frames, a green flag indicates a possible start codon and a red line indicates a stop codon. There is a full length ORF within repeats 3 and 5 (highlighted).

```
      1 CTTGACCCACATTGGAAGGAGACGGTATGTTTACCATTCTACAATGATCGACAATTCTAC
     61 AGAGAGCCTTATGGCAGGCCAGCAGGACAAAACAATCTCTCATTTGCTGGCCGTCACCTC
    121 AGGACTACTTATTTGAAGTGTCTCCAGTGTTCAAGGCTAACTCCAGAGATCTAAGAGCAC
    181 AGAACATACCGCCAGCTAACACAGCACATGCAGGAAGATGATCAACTCTTTTCTTCAACC
    241 TGCTCCATCGAAAGTGCACAACCTACTGGTGTCTCAAGCTTCCAGGCTCCTTTTCATACA
    301 GTCTGTGAAAGAAACACCTTGTGAGGTGTCTCCATCTCTCTCTGTCTCTGTCTATCTGTC
    361 TGTCTGTCTCTGTCTCTCTCTCTCTCCCTCCATTCCTCTTTTGCTCCCCCTCCCATTTCC
    421 CTCCTTGCCTCCATTTCACCATCTCTTCCACTCTCTGTCTCCATCCCCATCCTTCTACCC
    481 TCCCATATTCACTCCCCCCATCCACTTTCTACCTCCCTACTTCCCTATCTCTCTCTATCC
    541 ATTCTTCCCTTCCTTCTGCACTCTGTCACTCTCTCCCTACCACCCTCCACCCTCTGTCCC
    601 TAAATCCCTTCCCCCCTTCTCTCCACATCTGTGTTTGTCTCTCTCTTCGTGTCTTCCTCT
    661 GCCCCTAACCCCACCCATGGTCGTGACTTTATCTTCCCTTAGGATATTTGTGAGCATGAT
    721 GTGTGTGTGTTTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTGTGCTTGTGTGTGCCC
    781 GCATGTGTGCACGTGTTTGTGTGTGCGTGTGTGCGTGCATGCATGCAAATGTGTGTATGT
    841 GTGTGTTTGGTCTGAGGGTGTGCCTGTTCACAATTGTCTCTGTGTGTTGCTGCCAGGTGC
    901 CTAGGGGCTGTTTGGATTTTCATTTAATCTCAGTACAGGTGATGTTCCCTCTTGTCCTCA
    961 TAGCACAGTCAGACTTGGAAAGTCAAGGAAGGGGGTCTGAAACACTCTAGAGATAGGATG
   1021 GAGGTGGTGATGTCTTTGGATCTCAGACCATGATGTTGGGATCGTCAGTGTGTGGTCTTG
   1081 TCTTTGTAAGCTGATTGAATCCGGATGGGATGAACTGAGCATGGCTTCCATAGGGCTTGG
   1141 GATTCCTGGAAGGCTGAGTCCAATTCCCCAAGCTTTACTGAAGACTGCTCCCCTTCTCAT
   1201 AGGTGTCCTGGTACCTGTGCCACCTGCCAGTCAATGAATGATTTGGCTGATGGGAATGGC
   1261 GAGTCCTCTGACTCTTGTGTGCTCCCTGGGTGTGGGTCTAGACTGGCGACCCCGTGGCTT
   1321 GCCAGGGATGAGGAGCCTTTGGGGAGATTTTGCTGAGTGTCAGAGGACGCTTGAGGTCCG
   1381 CCTCCTGGTGCCCTGATGTCAGGTGCCAGGCGTGGCAGGCGTGGCGCTTCTTGGCGGCAC
   1441 CGCGAGGAAGGGGTAGGCATGTTCTGTAGCGCCTAACTGGTAGGTAGTGGGCGGGACTAC
   1501 CTGAGCAGAGGCAGAGGTATTTAAGGGGCAGTGGTCACAGCCACTCTGCTGGCAGTTGCT
mATG 1                                                 GCCACTCTGCTGGCAGTTGCT
mATG 2                                                 GCCACTCTGCTGGCAGTTGCT
mATG 3                                                 GCCACTCTGCTGGCAGTTGCT
mATG 4                                                 GCCACTCTGCTGGCAGTTGCT
mATG 5                                                 GCCACTCTGCTGGCAGTTGCT
mATG 6                                                 GCCACTCTGCTGGCAGTTGCT
mATG 7                                                 GCCACTCTGCTGGCAGTTGCT
mATG 8                                                 GCCACTCTGCTGGCAGTTGCT
mATG 9                                                 GCCACTCTGCTGGCAGTTGCT
mATG 10                                                GCCACTCTGCTGGCAGTTGCT
mATG 11                                                GCCACTCTGCTGGCAGTTGCT
mATG 12                                                GCCACTCTGCTGGCAGTTGCT
mATG 13                                                GCCACTCTGCTGGCAGTTGCT
mATG 14                                                GCCACTCTGCTGGCAGTTGCT
mATG 15                                                GCCACTCTGCTGGCAGTTGCT

   1561 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 1  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 2  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 3  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 4  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 5  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 6  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 7  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 8  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 9  GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 10 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 11 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 12 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 13 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 14 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 15 GCAGCTTGTGCTTGTTCTGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGGACTCCTGG
mATG 16                 TCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGG
```

```
   1621 GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
 mATG 1  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCCGTTGGTGGCAGTGGTGTGGCA
 mATG 2  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
 mATG 3  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
 mATG 4  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAATGGTGTGGCA
 mATG 5  GGGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAATGGTGTGGCA
 mATG 6  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
 mATG 7  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
 mATG 8  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAATGGTGTGGCA
 mATG 9   GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
mATG 10  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
mATG 11  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAATGGTGTGGCA
mATG 12  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
mATG 13  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
mATG 14  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
mATG 15  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAATGGTGTGGCA
mATG 16  GAGGCCGTCATTGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCA
                     M  A  E  A  G  S  P  V  G  G  S  G  V  A
          *                                 *        *

   1681 CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 1  CTAGAATCCCGGCGGCG
 mATG 2  CTGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 3  CTGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 4  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 5  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 6  CTGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 7  CTGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 8  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
 mATG 9  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 10  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 11  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 12  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 13  CTGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 14  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 15  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
mATG 16  CGGGAATCCCGGCGGCGCAGGAAGACGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTA
         R  E  S  R  R  R  R  K  T  V  W  Q  A  W  Q  E  Q  A  L  L
           **

   1741 TCAACTTTCAAGAAGAAGAGATACCTGAGCTTCAAGGAGAGGAAGGAGCTGGCCAAGCGA
 mATG 2  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 3  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 4  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 5  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 6  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 7  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 8  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
 mATG 9  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 10  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 11  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 12  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 13  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 14  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 15  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
mATG 16  TCAACTTTCAAGAAGAAGAGATACCTGAGCTT
         S  T  F  K  K  K  R  Y  L  S  F  K  E  R  K  E  L  A  K  R

   1801  ATGGGGGTCTCAGATTGCCGCATCCGCGTGTGGTTTCAGAACCGCAGGAATCGCAGTGGA
         M  G  V  S  D  C  R  I  R  V  W  F  Q  N  R  R  N  R  S  G

   1861  GAGGAGGGGCATGCCTCAAAGAGGTCCATCAGAGGCTCCAGGCGGCTAGCCTCGCCACAG
         E  E  G  H  A  S  K  R  S  I  R  G  S  R  R  L  A  S  P  Q

   1921  CTCCAGGAAGAGCTTGGATCCAGGCCACAGGGTAGAGGCATGCGCTCATCTGGCAGAAGG
         L  Q  E  E  L  G  S  R  P  Q  G  R  G  M  R  S  S  G  R  R

   1981  CCTCGCACTCGACTCACCTCGCTACAGCTCAGGATCCTAGGGCAAGCCTTTGAGAGGAAC
         P  R  T  R  L  T  S  L  Q  L  R  I  L  G  Q  A  F  E  R  N

   2041  CCACGACCAGGCTTTGCTACCAGGGAGGAGCTGGCGCGTGACACAGGGTTGCCCGAGGAC
         P  R  P  G  F  A  T  R  E  E  L  A  R  D  T  G  L  P  E  D

   2101  ACGATCCACATATGGTTTCAAAACCGAAGAGCTCGGCGGCGCCACAGGAGGGGCAGGCCC
         T  I  H  I  W  F  Q  N  R  R  A  R  R  R  H  R  R  G  R  P


   2161 ACAGCTCAAGATCAAGACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
```

```
 Dux10 1                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 2                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 3                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 4                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 5                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 6                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 7                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
 Dux10 8                ACTTGCTGGCGTCACAAGGGTGGGATGGGGCCCCTGCAGGTCCG
 Dux10 9                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
Dux10 10                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
Dux10 11                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
Dux10 12                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTACAGGTCCG
Dux10 13                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
Dux10 14                ACTTGCTGGCATCACAAGGGTCAGATGGGGCCCCTGCAGGTCCG
Dux10 15                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
Dux10 16                ACTTGCTGGCATCACAAGGGTCGGATGGGGCCCCTGCAGGTCCG
Dux10 17                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTACAGGTCCG
Dux10 18                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTACAGGTCCG
Dux10 19                ACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTACAGGTCCG
         T  A  Q  D  Q  D  L  L  A  S  Q  G  S  D  G  A  P  A  G  P
                                    *        **           *

    2221 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 1 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 2 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 3 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 4 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 5 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 6 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGGAGCAGGAAGT
 Dux10 7 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 8 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
 Dux10 9 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 10 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 11 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 12 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAGCTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 13 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 14 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 15 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 16 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAACTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 17 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAGCTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 18 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAGCTTGTTGCCACAGGAAGAAGCAGGAAGT
Dux10 19 GAAGGCAGAGAGCGTGAAGGTGCCCAGGAGAGCTTGTTGCCACAGGAAGAAGCAGGAAGT
         E  G  R  E  R  E  G  A  Q  E  N  L  L  P  Q  E  E  A  G  S
                                    *                 *

   2281  ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 1 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 2 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAATCCCAGCCT
 Dux10 3 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 4 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 5 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 6 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 7 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGGGTCCCAGCCT
 Dux10 8 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
 Dux10 9 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
Dux10 10 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
Dux10 11 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
Dux10 12 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCAGAGAGTCCCAGCCT
Dux10 13 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAATCCCAGCCT
Dux10 14 ACGGGGATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
Dux10 15 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCGGAGAGTCCCAGCCT
Dux10 16 ACGGGGATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGTTGGAGAGTCCCAGCCT
Dux10 17 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCAGAGAGTCCCAGCCT
Dux10 18 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCAGAGAGTCCCAGCCT
Dux10 19 ACGGGCATGGATACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCAGAGAGTCCCAGCCT
         T  G  M  D  T  S  S  P  S  D  L  P  S  F  C  G  E  S  Q  P
               *                                    **   **
```

```
2341  TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
 Dux10 1 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
 Dux10 2 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
 Dux10 3 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
 Dux10 4 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
 Dux10 5 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
 Dux10 6 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
 Dux10 7 TTCCAGGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
 Dux10 8 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGTCCCCACTCGAGCAGGCAAC
 Dux10 9 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCGACAAGAGGCCCCCACTCGAGCAGGCAAC
Dux10 10 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
Dux10 11 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
Dux10 12 TTCCAAGTGGCACAGCCCCGTGGGGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
Dux10 13 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
Dux10 14 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
Dux10 15 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCGAGCAGGCAAC
Dux10 16 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
Dux10 17 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
Dux10 18 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
Dux10 19 TTCCAAGTGGCACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAAC
         F  Q  V  A  Q  P  R  G  A  G  Q  Q  E  A  P  T  R  A  G  N
               *                 *        *        *        *

   2401  GCAGGCTCTCTGGAACCCCTCCTTGATCAGCTGCTGGATGAAGTCCAAGTAGAAGAGCCT
 Dux10 1 GCAGGCT
 Dux10 2 GCAGGCT
 Dux10 3 GCAGGCT
 Dux10 4 GCAGGCT
 Dux10 5 GCAGGCT
 Dux10 6 GCAGGCT
 Dux10 7 GCAGGCT
 Dux10 8 GCAGGCT
 Dux10 9 GCAGGCT
Dux10 10 GCAGGCT
Dux10 11 GCAGGCT
Dux10 12 GCAGGCT
Dux10 13 GCAGGCT
Dux10 14 GCAGGCT
Dux10 15 GCAGGCT
Dux10 16 GCAGGCT
Dux10 17 GCAGGCT
Dux10 18 GCAGGCT
Dux10 19 GCAGGCT
      A  G  S  L  E  P  L  L  D  Q  L  L  D  E  V  Q  V  E  E  P


   2461  GCTCCAGCCCCTCTGAATTTGGATGGAGACCCTGGTGGCAGGGTGCATGAAGGTTCCCAG
         A  P  A  P  L  N  L  D  G  D  P  G  G  R  V  H  E  G  S  Q

   2521  GAGAGCTTTTGGCCACAGGAAGAAGCAGGAAGTACAGGCATGGATACTTCTAGCCCCAGC
         E  S  F  W  P  Q  E  E  A  G  S  T  G  M  D  T  S S  P  S

   2581  GACTCAAACTCCTTCTGCAGAGAGTCCCAGCCTTCCCAAGTGGCACAGCCCTGTGGAGCG
         D  S  N  S  F  C  R  E  S  Q  P  S  Q  V  A  Q  P  C  G  A

   2641  GGCCAAGAAGATGCCCGCACTCAAGCAGACAGCACAGGCCCTCTGGAACTCCTCCTCCTT
         G  Q  E  D  A  R  T  Q  A  D  S  T  G  P  L  E  L  L  L  L

   2701  GATCAACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGTCCCACTGGATTGGGGT
         D  Q  L  L  D  E  V  Q  K  E  E  H  V  P  V  P  L  D  W  G

   2761  AGAAATCCTGGCAGCAGGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGAGGAA
         R  N  P  G  S  R  E  H  E  G  S  Q  D  S  L  L  P  L  E  E

   2821  GCAGTAAATTCGGGCATGGATACCTCGATCCCTAGCATCTGGCCAACCTTCTGCAGAGAA
         A  V  N  S  G  M  D  T  S  I  P  S  I  W  P  T  F  C  R  E

   2881  TCCCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCCCCCACTCAA
         S  Q  P  P  Q  V  A  Q  P  S  G  P  G  Q  A  Q  A  P  T  Q

   2941  GGTGGGAACACGGACCCCCTGGAGCTCTTCCTCTATCAACTGTTGGATGAAGTCCAAGTA
         G  G  N  T  D  P  L  E  L  F  L  Y  Q  L  L  D  E  V  Q  V

   3001  GAAGAGCATGCTCCAGCCCCTCTGAATTGGGATGTAGATCCTGGTGGCAGGGTGCATGAA
         E  E  H  A  P  A  P  L  N  W  D  V  D  P  G  G  R  V  H  E
```

```
   3061  GGTTCGTGGGAGAGCTTTTGGCCACAGGAAGAAGCAGGAAGTACAGGCCTGGATACTTCA
         G  S  W  E  S  F  W  P  Q  E  E  A  G  S  T  G  L  D  T  S

   3121  AGCCCCAGCGACTCAAACTCCTTCTTCAGAGAGTCCAAGCCTTCCCAAGTGGCACAGCGC
         S  P  S  D  S  N  S  F  F  R  E  S  K  P  S  Q  V  A  Q  R

   3181  CGTGGAGCGGGCCAAGAAGATGCCCGCACTCAAGCAGACAGCACAGGCCCTCTGGAACTC
MD4Z4 1                                                 GGCCCTCTGGAACTC
MD4Z4 2                                                 GGCCCTCTGGAACTC
         R  G  A  G  Q  E  D  A  R  T  Q  A  D  S  T  G  P  L  E  L


   3241  CTCCTCTTTGATCAACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTG
MD4Z4 1  CTCCTCTTTGATCAACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTG
MD4Z4 2  CTCCTCTTTGATCAACTGCTGGGCGAAGTCCAAAAG:AAGAGCATGTGCCAGCCCCACTG
         L  L  F  D  Q  L  L  D  E  V  Q  K  E  E  H  V  P  A  P  L
                              *           *

   3301  GATTGGGGTAGAAATCCTGGCAGCATGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCC
MD4Z4 1  GATTGGGGTAGAAATCCTGGCAGCATGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCC
MD4Z4 2  GATTGGGGTAGAAATCCTGGCAGCATGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCC
         D  W  G  R  N  P  G  S  M  E  H  E  G  S  Q  D  S  L  L  P


   3361  CTGGAGGAAGCAGCAAATTCGGGCAGGGATACCTCGATCCCTAGCATCTGGCCAGCCTTC
MD4Z4 1  CTGGAGAAAGCAGCAAATTCGGGCAGGGATACCTCGATCCCTAGCATCTGGCCAGCCTTC
MD4Z4 2  CTGGAGGAAGCAGCAAATTCGGGCAGGGATACCTCGATCCCTAGCATCTGGCCAGCCTTC
         L  E  E  A  A  N  S  G  R  D  T  S  I  P  S  I  W  P  A  F
               *

   3421  TGCAGAAAATCCCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCC
MD4Z4 1  TGCAGAAAATCCCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCC
MD4Z4 2  TGCAGAAAATCCCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCC
         C  R  K  S  Q  P  P  Q  V  A  Q  P  S  G  P  G  Q  A  Q  A


3481     CCCATTCAAGGTGGGAACACGGACCCCCTGGAGCTCTTCCTTGATCAACTGCTGACCGAA
MD4Z4 1  CCCATTCAAGGTGGGAACACGGACCCCCTGGAGCTCTTCCTTGATCAACTGCTGACCGAA
MD4Z4 2  CCCATTCAAAGTGGGAACACGGACCCCCTGGAGCTCTTCCTTGATCAACTGCTGACCGAA
         P  I  Q  G  G  N  T  D  P  L  E  L  F  L  D  Q  L  L  T  E
                  *

   3541  GTCCAACTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mD4Z4 1  GTCCAACTTGAGGAGCAGGGGCCTGCCCCTGTGGATGTGGAGGAA
mD4Z4 2  GTCCAACTTGAGGAGCAGGGGCCTGCCTCTGTGAATGTGGAGGAA
mTGA 1       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 2       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 3       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 4       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 5       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 6       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 7       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 8       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
mTGA 9       CTTGAGGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATG
      V  Q  L  E  E  Q  G  P  A  P  V  N  V  E  E  T  W  E  Q  M
                                    *     *

 3601  GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 1 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 2 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 3 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 4 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 5 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 6 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 7 GACACAACATCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 8 GACACAACACCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
mTGA 9 GACACAACATCTGATCTGCCTCTCACTTCAGAAGAATATCAGACTCTTCTAGATATGCTC
      D  T  T  P  D  L  P  L  T  S  E  E  Y  Q  T  L  L  D  M  L
               *
```

```
 3661  TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 1 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 2 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 3 TGACCCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 4 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 5 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 6 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 7 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 8 TGACTCCCCGACAGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
mTGA 9 TGACTCCCCGACGGTACCCCTTGCTTCTAGAAACCCGAGAGGCCAAAGTCCTGAAGAGAC
           *       *

 3721  CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 1 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 2 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 3 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 4 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 5 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 6 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 7 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 8 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC
mTGA 9 CCGATTTGGAACTGGAGAAGGGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTC


 3781  CAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 1 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCCAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 2 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 3 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 4 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 5 CAGTGACTTCCCGAAATCCAAGGTGTCTCCCCAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 6 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCCAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 7 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATCGCAGGTGGGGA
mTGA 8 CAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGA
mTGA 9 CAGTGACTTCCCGAAATACAAGGTGTCTCGCTAACTATAAGGATTAATTGCAGGTGGGGA
                         *           * *             *   *

 3841  TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 1 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 2 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 3 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 4 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGAGAGAG
mTGA 5 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 6 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 7 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 8 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
mTGA 9 TAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAG
                                                              *

 3901  AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 1 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 2 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 3 AGTGACTCTGCTACAGGGACAGGTGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 4 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 5 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 6 AGAGACTCTACTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 7 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
mTGA 8 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCGACAACAAGGAGC
mTGA 9 AGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGC
         *       *             *                          *

 3961  CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 1 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACCGGCTCCAGTGGACCCCAGGAA
mTGA 2 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 3 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 4 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 5 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 6 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 7 CTAAAGATAATCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 8 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
mTGA 9 CTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAA
                *                           *
```

Figure 5.8 Alignment of RT-PCR clone sequences with the Dux repeat consensus

The sequences of the RT-PCR were aligned to the Dux consensus sequence using Sequencher. The predicted amino acid sequence of the ORF is shown underneath. Nucleotide changes from the consensus are highlighted in red and potential coding changes are indicated with an asterisk.

						Homeodomain		
Position	**1571**	**1597**	**1609**	**1670**	**1682**	**1706**	**1749**	**1753**
No. of clones	**C**	**T**	**G**	**G**	**G**	**C**	**C**	**A**
5	**C**	T	G	G	G	C	C	A
6	**C**	T	G	G	T	C	C	A
4	**C**	T	G	A	G	C	C	A
1	**C**	G	A	G	G	C	C	A
6	**T**	G	A	G	G	T	T	G
Coding Change	A-V	F-V	E-K	S-N	R-L	T-M		K-E

Table 5.1 Summary of nucleotide variants in mATG clones

Sequence variants are shown in orange, blue blocks are the same base as the genomic consensus sequence. The number of clones containing each set of variants is shown. Three variants fall within the homeodomain sequence. Coding changes are shown underneath.

5.2.2.ii Sequence analysis of the mDux10 RT-PCR products

The mDux10 RT-PCR product covers 231bp, beginning 25bp downstream of the second homeodomain (Figure 5.8). Twenty six mDux10 clones were sequenced and compared with the genomic consensus (Table 5.2). Seven of the variants identified in this RT-PCR product were also seen in the genomic sequences. None of the clones were identical to the consensus sequence, 19 sequences had between 1 and 3 variations from the consensus and 4 of these variant positions result in a coding change.

There were 7 highly variant clones, which contained 31 base changes from the consensus. These sequences were aligned to the BAC clone 142C15 and were almost identical to repeat two, which is a truncated repeat lacking an ORF (Figure 5.7). The sequences did not match any of the other repeats in the BAC clone.

5.2.2.iii mD4Z42

The mD4Z4 RT-PCR product is 360bp long and ends 78bp upstream of the stop codon (Figure 5.8). Twenty eight clones were sequenced and compared with the genomic consensus (Table 5.3). One clone was identical to the consensus sequence, while 1 sequence had only a single base change. Five clones had a T-C change at position 3247 that was also seen in 19 of the 69 genomic sequences. In addition, these 5 sequences all had a further 7 or 8 base changes from the consensus that were not present in the genomic sequence data, this suggests it is unlikely that these 5 clones originated from these particular repeats. Of the 11 variant positions identified in these 5 clones, 9 result in coding changes.

The final 21 sequences had 31 or 32 base changes from the consensus sequence as well as two deletions of 3bp and 2bp respectively. These highly variant sequences were aligned with the BAC clone 142C15. 19 of the sequences were identical to the final repeat except for 7 single base pair changes that were present in only one of the sequences and are likely to be PCR or cloning errors. A further 2 sequences had a single T-C change at position 103679, causing a phenylalanine to serine substitution. These sequences aligned with repeat one, the

first partial repeat. This data suggests that most of the RT-PCR products using these primers originate from repeats represented in the BAC clone 142C15.

Finally, the 5 sequences that had 6-9 changes from the genomic consensus were aligned with the BAC sequence. Two of the clones varied at 2 positions from this sequence (aligned with repeat 3) and a single base change each which is likely due to PCR or sequencing error. One sequence aligned with repeat 4, this had 3 changes from the sequence. The remaining 2 sequences were highly variable compared with this BAC clone. It is likely that at least the last 3 sequences come from genomic repeats that have not yet been sampled.

5.2.2.iv mTGA

The mTGA RT-PCR product is 581bp long, including the final 117bp of the ORF and a further 464bp downstream of the stop codon. 31 clones were sequenced and compared with the consensus (Table 5.4). Four clones were identical to the consensus sequence, while a further 5 clones had only one or two base changes. 22 clones contained a 4bp deletion of GAGA at position 1496, 8 of these clones had a further 8-10 base changes from the consensus as well as a single base pair deletion at position 3895 in 2 of the clones and position 3960 in the other 6. The remaining 14 clones contained 24 base changes from the consensus sequence as well as two single base pair deletions. None of the variants identified in these RT-PCR products are present in the genomic sequences.

The 14 highly variant clones were compared with the BAC clone 142C15 and all were found to be identical to repeat 2 apart from 15 single base pair changes that were present in only one of the sequences and are likely to be PCR or cloning errors. Of the remaining 8 clones, 6 aligned with repeat number 5 of the BAC clone with 2 variants present in all sequences and a further 6 single base pair changes in individual sequences. The final 2 clones aligned at repeat number 6 and contained 8 variants compared with the BAC clone. These 8 sequences are likely to originate from repeats that have not yet been sequenced.

Position	2187	2195	2199	2201	2212	2222	2229	2231	2232	2234	2240	2244	2252	2255	2256	2286	2295	2296	2298	2302	2304	2308	2312	2319	2326	2331	2337	2339	2342	2350	2374	2381	2390	2402	2407
No. of clones	G	G	G	A	G	A	A	A	G	G	G	C	A	T	G	C	C	T	G	C	T	G	T	C	G	G	G	C	T	G	C	C	A	C	T
11	G	G	G	A	G	A	A	A	G	G	G	C	A	T	G	C	C	T	G	C	T	G	T	C	G	G	G	C	T	G	C	C	G	C	T
1	G	G	G	A	G	A	A	A	G	G	G	C	A	T	G	C	C	T	G	C	T	G	T	C	G	A	G	C	T	G	C	C	G	C	T
1	G	G	G	A	G	A	A	A	G	G	G	C	A	T	G	C	C	T	G	C	T	G	T	C	G	A	G	C	T	G	C	C	A	C	T
4	G	G	G	A	A	A	A	A	G	G	G	C	G	T	G	C	C	T	G	C	T	G	T	C	A	G	G	C	T	G	C	C	A	C	T
2	A	G	G	A	G	A	A	A	G	G	G	C	A	T	G	G	C	T	G	C	T	G	T	C	G	G	G	C	T	G	C	C	A	C	T
7	C	T	C	G	G	T	G	G	C	A	T	A	G	C	A	C	T	A	A	A	C	T	C	T	A	G	A	T	C	T	A	T	A	T	C
Coding Change		G-V		D-G	A-T	E-V		E-G	E-D	R-H	G-V		N-S	L-S				S-T		P-T		D-Y	L-S		G-R			P-L	F-S	A-S	Q-K	A-V	Q-R	A-V	S-P
No. of genomic sequences	1 (A)												1			5									3						2	2 (A)	18		

Table 5.2 Variants in mDux10 clones

Sequence variants are shown in orange or green, blue blocks are the same base as the genomic consensus sequence. The number of clones with each set of variants is shown. Coding changes are shown underneath. The number of genomic sequences amplified by other members of Jane Hewitt's group that contain each variant is indicated. (A) indicates a base change in the genomic sequences at the same position, but a different base.

Position	3238	3240	3245	3246	3247	3260	3263	3265	3293	3326	3359	3366	3386	3395	3400	3408	3411	3415	3420	3427	3428	3431	3432	3444	3456	3483	3485	3489	3490	3496	3497	3500	3501	3505	3507	3518
No. of clones	C	C	T	C	T	T	A	G	C	T	C	G	G	C	C	C	G	G	C	A	A	C	C	A	C	C	T	A	G	A	A	C	G	C	C	T
1	C	C	T	C	T	T	A	G	C	T	C	G	G	C	C	C	G	G	C	A	A	C	C	A	C	C	T	A	G	A	A	C	G	C	C	T
1	C	C	T	C	T	T	G	G	C	T	C	G	G	C	C	C	G	G	C	A	A	C	C	A	C	C	T	A	G	A	A	C	G	C	C	T
2	C	C	T	C	C	T	A	G	C	G	C	G	T	C	C	C	G	A	C	G	A	C	C	A	C	C	C	A	G	A	A	C	G	C	T	T
1	C	C	T	C	C	T	A	G	T	G	C	G	T	T	C	C	G	A	C	G	A	C	C	A	C	G	C	A	G	A	A	C	G	C	T	T
1	C	C	T	C	C	T	A	G	T	G	C	G	T	C	C	C	G	A	C	G	A	C	C	A	C	G	C	A	G	A	A	C	G	C	T	T
1	C	C	T	C	C	T	A	G	T	G	C	G	T	C	C	C	G	A	C	G	A	C	C	A	C	C	C	A	G	A	A	C	G	C	C	T
2	A	T	C	T	3bp del	A	G	A	C	G	G	T	T	T	A	T	A	A	G	G	2bp del	T	G	G	A	C	C	G	A	T	T	T	A	T	T	C
19	A	T	C	T	3bp del	A	G	A	C	G	G	T	T	T	A	T	A	A	G	G	2bp del	T	G	G	A	C	C	G	A	T	T	T	A	T	T	T
Coding Change	I-L		P-L		F-L	Q-L	G-D	K-E	A-V	R-M	R-P	D-E	M-R	L-S	T-P		W-STOP	T-A	L-F	I-V	Frameshift	S-L					I-T		G-S	I-F	N-I	T-I		P-S		F-S
No. of genomic sequences					19		1	3																					1							

Table 5.3 Variants in mD4Z42 clones

Sequence variants are shown in orange or green, blue blocks are the same base as the genomic consensus sequence. The number of clones with each set of variants is shown. Coding changes are shown underneath. The number of genomic sequences amplified by other members of Jane Hewitt's group that contain each variant is indicated.

Position	3588	3614	3626	3627	3668	3672	3678	3684	3691	3697	3698	3707	3712	3722	3740	3741	3748	3749	3765	3772	3773	3808	3809	3811	3824	3859	3863	3868	3895	3901	3954	3960	3971	3985	4027	4080	4093	4100	4104	4107	4120	4124
No. of clones	C	T	T	T	C	A	C	T	A	A	G	G	G	G	G	G	C	C	T	C	C	C	G	T	T	C	G	G	G	GAGA	A	C	C	C	G	G	T	C	C	A	T	C
4	C	T	T	T	C	A	C	T	A	A	G	G	G	G	G	G	C	C	T	C	C	C	G	T	T	C	G	G	G	GAGA	A	C	C	C	G	G	T	C	C	A	T	C
2	T	T	T	T	C	A	C	T	A	A	G	G	G	G	G	G	C	C	T	C	C	C	G	T	T	C	G	G	G	GAGA	A	C	C	C	G	G	T	C	C	A	T	C
2	C	T	T	T	C	A	C	T	A	A	G	G	G	G	G	G	C	C	T	C	C	C	G	C	T	C	G	G	G	GAGA	A	C	C	C	G	G	T	C	C	A	T	C
1	C	T	T	T	C	A	C	T	A	A	G	G	G	G	G	G	C	C	T	C	C	C	C	C	T	C	G	G	G	GAGA	A	C	C	C	G	G	T	C	C	A	T	C
6	C	T	T	C	C	T	T	T	A	A	G	G	G	G	T	G	C	C	T	C	C	C	C	T	T	C	G	G	DEL	4BP DEL	A	C	C	C	A	G	G	T	C	A	T	C
2	C	T	T	C	T	A	T	T	A	A	A	G	G	G	T	G	C	C	T	C	C	C	G	T	T	C	G	G	G	4BP DEL	C	DEL	G	T	G	G	T	T	C	A	T	C
14	C	G	C	C	T	A	C	A	C	C	G	DEL	A	A	G	T	T	T	DEL	T	G	T	G	T	G	G	C	A	G	4BP DEL	A	C	C	C	G	C	T	C	G	G	C	T
Coding Changes	W-G	E-D		P-S	Outside ORF																																						

Table 5.4 Variants in mD4Z42 clones

Sequence variants are shown in orange, blue blocks are the same base as the genomic consensus sequence. The number of clones with each set of variants is shown. Coding changes are shown underneath. The number of genomic sequences amplified by other members of Jane Hewitt's group that contain each variant is indicated.

5.2.3 BAC clone repeats

For each of the regions amplified by RT-PCR, a large number of sequences matched the BAC clone 142C15 (Figure 5.9a)), with a total of 43% of clones apparently originating from this region. As this smaller Dux cluster contains only two full repeats it is perhaps surprising that so many clones align here. However, as the mouse embryo RNA comes from a different mouse strain to the one amplified for the genomic sequences, it is possible the Swiss-Webster mice have a larger number of repeats in this Dux cluster. It is also possible that the large number of amplification products from this region is due to amplification or cloning bias

Surprisingly, the sequences which matched the BAC clone did not appear to come from the same repeats, nor were they all expressed from full length repeats (Figure 5.7) (Figure 5.8 Sequence alignment). Each sequence was manually aligned with the other repeats in order to confirm it was identical to only one of the regions, this was confirmed in all cases.

Six mATG sequences and 2 mD4Z42 sequences aligned with the third repeat, which contains a full length Dux ORF (Figure 5.7). As transcripts have been identified from the beginning and end of this repeat it is likely that this ORF is transcribed. Transcripts have also been identified from the first two partial repeats, even though these regions do not appear to contain a functional ORF (Figure 5.7). The organisation of this region may vary in Swiss Webster mice, resulting in functional ORFs that could account for this expression.

a

	mATG	mDux10	mD4Z42	mTGA
No. of clones sequenced	22	26	28	31
No. of clones matching BAC 142C15	6	7	19	14
% of clones matching BAC 142C15	27%	27%	68%	45%

b

Figure 5.9 mDux RT-PCR products which appear to originate from BAC 142C15

a) Summary of the mDux products that originate from the BAC clone 142C15 b) Schematic representation of the region of BAC clone 142C15 which contains mouse Dux repeats is shown. Repeats are indicated in orange. Repeats 3 and 5 are full length repeats, repeats 1, 2 and 4 are partial repeats. The positions of mATG, mDux10, mD4Z42 and mTGA RT-PCR products are indicated in purple.

```
1    AGGTCTATGCTAATGTTCACAGGACACTACCTGAGGGAAGGGGCAGGGAAGATTATTCCA
61   TGGGTTTAATGGACTGCTTATATTCTTACCAATGCTTGGTGCTGAACTGTGCAATGATTT
121  ATAAGCCAGTCCTTGCAAATTTACTACTCATAGAACGTTGGTTTTGTTTTGTTTTGATTG
181  TTTCCATTGTTTTCAATAAAGTTTATTACTTTCAACAAATTTTGTTTGTTTGTGTGTTGT
241  TTTTGAGTCTGGGGGTTAAATAGGCTCTTTTGATCTAACCGGGGAATGTGTATGAGCTTT
301  CCAATTAAAAGGGATAAAAGACATAGGAAGGTAATCAGGAGGGATGCAGAGAGGCCAGGT
361  AGGAAGGGGAGAGAACAAGGAAGGCTGGGCTCTGTCCAATGGAGAGATGGGCTTGGTTGC
421  TCTAGTAACAAGCCAGTGGAATGTCAATTGAAAGAAGAATGTGTGTAAGGCATAACCATG
481  TAAAGAAAGATCATGCTGGGAATACAGTGCAGCCATGCGGGAACATCTTATTCCAAGCTT
541  GACACACAGTTCATCCTTCCATCCACCCATGCATTGATCTGGAGTTCTTGCTACTCCTAT
601  TGGTTTTAATCCCCAACACCCACCCTTATTCCTCTTTTGATCTTAAGATCCCCAAGTCTC
661  CAAGGTCATAATGTTTACTTGGAGCTCCAAGTTGAACCTTGACAAACATGGAAAGAATGT
721  GTTTATCATTCTACAATGGCCTATAATTTTACAAAGAACATTTTGGCATGCCAGCAGGAT
781  AAAACAATATCTCACCTGCTGGTCATCAGCTCAGGGAATCCTCCTTGAGGTGTCTCCAGT
841  GTACAAGGCTAACTCCAGAGATCTAACAGTACAGGCCATACCCTTAGAAAACAGCACTAT
901  GCAAAAGATGATCAACACTTTTCTTCACACTGCTCAATAGAAAATGAACAATTTAATGAT
961  ATCTCAAGCTTCCAGGCTCCTTATCATACTGTCTGTGAAAGAAACAAATGGGAAGGTCCC
1021 TCTCCTTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCC
1081 TCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCCCTCCCCCTCTCCCTCCCCCTCTC
1141 CCTCCCCATCTCCCTCCCCATCTCCCTCCCCATCTCCCTCCCCATCTCCCTCCCCTTCTT
1201 TTCTCACACCCCCTCTTTCTCTTACTTGCTATTGCTCTCTTAATCTCTATCTCTCTCTCC
1261 CTCCCTCCATTCCTTCTTTGATCCCCCTCCCTTTTCTGTCCATGCCTCACTTTCACCATC
1321 TCTTCCACTCTCTCTCCACATCCTTTCTCCCTCCCTTCTGCACTCGCTCCCTCCCCTCCA
1381 CTTTCTCCCTTCCTACATAATTCCTTCTCTCTATCGATCCTTCCCTCCCTTCTGCACTCC
1441 CTCACCCAATCCTCTCCTCACTCTGTCCCTCAATCCCTTCCCACTCTCTTACTCTCTCTC
1501 TCTCTCTTCCAGTCTGTCTGTCTGTTTGTCTCTTCCTGTTTTACTTTGCCATGGAACCCC
1561 ATCCATAGTCCTGCCATTCACATTCCTCAGGATTGTGTGTGTGTGTGTGAGAGAGAGTGT
1621 GGTCTGAGGGTGTGCCTGTTCACAATTGACTCAGTGTGTTGTTGCTAAGGGCTTGGGGGC
1681 AGTTTGGGGTTGCATTTTTGATCCCAGTACAGGTGAAGTTCGGTCTTGTCCTCATAGCAC
1741 AGCCAGATTTGGAAAGCCAAGGAAGGGGGTCTGAAACACTCTAGAGATAGGAAGGAGAAG
1801 GGGATGTCTAGAACTCCTTCTTGATCAACTGCTGGATGACGTCCAAATAGAAAAGTATGC
1861 TCCAGCCCCTCTGGATTTGGATGGAGACCCTGGTGGCCGGGTGCATGAAGGTTCCCGGGA
1921 CAGCTTTTTGCCACAGGAAGAAACAGGAAGTACAGGCATGGATACTTCAAGCCCTAGAAA
1981 AACACTCTAGTTCTGCAAAGAGTCCCAGCATTCCCAAATCACAGCCCTGTGGAGCAGGAT
2041 AGCAGCCTTCCATATAACCCAGGGAAGGAGTCCACTCTCCCACTTCCCCTCAGGTAGTGT
2101 CTGTACTATTTAGTGAAGCCGACCTAGTGTCCCATCTGAATTCCCGGGGTCGACTGGAGC
2161 CAGTTACTTGGTGTGGCCCACAGCTCAAGATCAAGACTTGCTGGCGTCACAAGGGTCGGA
2221 TGGGGCCCCTACAGGTCTGGAAGGCAGAGAGCATGAAGGTGCCCAGGAGAGCTTGTTGCC
2281 AGAGGAAGAAGCAGGAAGTACGGGAAAGGATACTTTGAGCTCTAGCGATTCACCCTCCTT
2341 CTGAAGAGAGTCCCAGACTTCCCAAAAGGCACAGCCCTGTGGAGCGGGCCAAAAAGATAC
2401 CCGGAATCAAGCAGACAGCACAGGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH134                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH440                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH443                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH136                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH147                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH140                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH438                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH135                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH436                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH126                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH123                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH127                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH132                          GGCCCTCTGGGAATTCTCCCTGATCAACTGCAGGGCAA
AH130                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH118                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH439                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH144                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH117                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH122                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCGGGGCGA
AH143                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA
AH137                          GGCCCTCTGGAAATTCTCCCTGATCAACTGCAGGGCAA

2461 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH134 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH440 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH443 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH136 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH147 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
```

```
AH140 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH438 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH135 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH436 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH126 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH123 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH127 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH132 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH130 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH118 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH439 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH144 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH117 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH122 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH143 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG
AH137 AGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAGAAATCCTGGCAGCAG

 2521 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH134 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH440 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH443 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH136 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH147 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH140 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH438 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH135 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH436 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH126 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH123 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH127 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH132 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH130 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH118 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH439 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH144 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH117 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH122 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH143 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT
AH137 GGAGCATGAAGGTTCCCAGGACAGCTTACTGCGCCTGGATGAAGCAGCAAATTCGGGCAT

 2581 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH134 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH440 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH443 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH136 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH147 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH140 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH438 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH135 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGGGTTGCAGCCTCCCCAGGTGG
AH436 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH126 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH123 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH127 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH132 GGGTACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH130 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH118 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH439 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH144 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH117 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH122 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH143 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG
AH137 GGATACCTTGATCACTAGCATTTGACCAACCTTGTGCAGAGTTGCAGCCTCCCCAGGTGG

 2641 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH134 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH440 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH443 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH136 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH147 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH140 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH438 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH135 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH436 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH126 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH123 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH127 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH132 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
```

```
AH130 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH118  CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH439 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH144 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH117 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH122 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH143 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAGTGGGTTCATAGACTCTC
AH137 CACAGCCATCTGGACCAGGCCAAGCACAGGCCCCCACTCAGAATGGGTTCATAGACTCTC

 2701 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH440 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGAGCCTGCCC
AH443 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH136 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH147 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH140 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH438 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH135 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH436 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH126 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH123 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH127 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH132 TGGAGCTCTCCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH130 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH118 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH439 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH144 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCGTGCCC
AH117 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH122 TGGAGCTCTCCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH143 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC
AH137 TGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCC

 2761 CTGTGAATGTGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTC
AH134 CTGTGAATGTGGAGGAA
AH440 CTGTGAATGTGGAGGAA
AH443 CTGTGAATGTGGAGGAA
AH136 CTGTGAATGTGGAGGAA
AH147 CTGTGAATGTGGAGGAA
AH140 CTGTGAATGTGGAGGAA
AH438 CTGTGAATGTGGAGGAA
AH135 CTGTGAATGTGGAGGAA
AH436 CTGTGAATGTGGAGGAA
AH126 CTGTGAATGTGGAGGAA
AH123 CTGTGAATGTGGAGGAA
AH127 CTGTGAATGTGGAGGAA
AH132 CTGTGAATGTGGAGGAA
AH130 CTGTGAATGTGGAGGAA
AH118 CTGTGAATGTGGAGGAA
AH439 CTGTGAATGTGGAGGAA
AH144 CTGTGAATGTGGAGGAA
AH117 CTGTGAATGTGGAGGAA
AH122 CTGTGAACGTGGAGGAA
AH143 CTGTGAATGTGGAGGAA
AH137 CTGTGAATGTGGAGGAA

 2821 CAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGTACCCCTTGCTTCT
 2881 AGAAACCCGAGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGGGACCCAT
 2941 CCCAGCAAGGATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCT
 3001 CGCTAACTATAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCC
 3061 GGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACAGATGGAGAGGC
 3121 CAATAGTGACTCCTCAACAACAAGGACCCTAAAGGTAACCCCAAAAGAAGGGCCACACCA
 3181 AGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACTAGGCTTCACTA
 3241 CAGAGGACACACACTCCCTGAGGGCAATGGGGAGAGTGGACTCCTTCCCTGGGTTATATG
 3301 GACTGCTGTTATCCTTACAGATGCTTCGTGCAGAGCTGTGCAAGGTTTTACAGGCCAGTC
 3361 TTGTAATATCTACTACCCATAGGTCTTTTGTTTGTTTTCTTTTTCTTTTTCACTTTCTTT
 3421 TTCATTTTTTTTCTTTTTCTTTTTTTTAGGGGTGGGTTGTCTTTGCTGGGTTTGGTTTGG
 3481 TTTTGTGTAGTTTGTTTCCATTGCTTTCAATAAAGTTTATTGATTTTAACAAAATTTGTT
 3541 TGTGTGTTTGTGTGCTGTTTTGTGGGATGCGGGGTGGGTTGAATAGGCTGTTTTGTTCTA
 3601 CCCGGAGAACGTGCATGAGAATTCCATTTAAAGGGATGAAAGACATAGGAAGGTATTCAG
 3661 GAGGGATGCAGAGAAGCCAGGTAGAAAAGGGGAGAGAACAAGGAAGGCTGGGCTGTGTCC
 3721 AGTGGAGAGACGGGCTTGCTTGATCTAGTAACAAGCAGGTGGAATGTCAACTGAAAGAAG
 3781 AATGTGTGTAAGGCATAACCTTGTCAAGAAAGATCATCTTGGGAATACAGTGCAGCTAGG
 3841 TGTGAACATCTTATTCCAAGCTGAAACACAGCCATCAGTCTATTCTTCATCCATCCATCC
 3901 ATCCATCCATCTATCCATCCTTGCATCCATTCATCCATGCATGCCTAAATCAATCTGTAG
 3961 ATCTTTCAATCCTATTGGGATTAATCCCCAGCACCCTCCCTTATCCCTCTTTTGATCTTG
 4021 AGATTCCCAGGTGTTCAAGGTCATGCTGTTTATATGGAGCTCCAAGTTGATCTTTGACCC
 4081 ACATTGGAAGGAGACGGTATGTTTACCATTCTACAATGACCGACAATTCTACAGAGAGCC
 4141 TTATGGCATGCCAGCAGGACAAAACAATCTCTCATTTGCTGGCCATCACCTCAGGACAAC
```

```
 4201 TTATTTGAAGTGTCTCCAGTGTTCAAGGCTAACTCCAGAGATCTAAGAGCACAGAACATA
 4261 CTGCCAGCAAACACAGCACATGCAAGAAGACGATCAACTCTTTTCCTCACCCTGCTCCAT
 4321 AGAAAGTGCACAACCTACTGGTGTCTCCAGCTTCCAGGCTCCTTTTCATACAGTCTGTGA
 4381 AAGAAACACCTTGTGAGGTGTCTCCCTCTCTCTCTGTCTCTCTCTGTCTGTCTGTCTGTC
 4441 TCTATCTCTCTCTCTCTCCCTTCATTCCTCTTTTGCTCCCCCTCCCATTTCCCCACTTGC
 4501 CTCCATTTCACCATCTCTTCCACTCTCTGTCTCCATCCCCATACTTCTATCCTCCCATCT
 4561 TCACTCCCCCATCCACAATCTACCTCCCTACTTCCCTATCTCTCTCTATCCATTCTTCCC
 4621 TTCCTTCTGCACTCTGTCACTCTCTCCCTACCACCCTCCACCCTCTGTCCCTAAATCCCT
 4681 CCCCCCTTCTCTTTACATCTGTGTTTGCCTCTCTCTTTGTGTCTTCCTCTGCCCCTAACC
 4741 CCACCAATGGTCGTGACTTTATCCTCCCTTAGGATATTTGTGAGCATGATGTGTGTGTGT
 4801 GTGTATGTGTGTGTTTGTGCTTGTGTGTGCCCGCATGTGTGCATGTGTTTGTGTGTGCAT
 4861 GCGTGCGTGCATGCATGCAAATGTGTGTAAGTGTGTGTGTGGTCTGAGGGTGTGCCTGTT
 4921 CACAATTGTCTCTGTGTGTTGCTGCCAGGTGCCTAGGGGCTGTTTGGACTTTCATTTTGA
 4981 TCTCAGTACAGGTGATGTTCCCTCTTGTCCTCATAGCACAGTCAGACTTGGAAAGTCAAG
 5041 GAAGGGGGTCTGAAACACTCTAGAGATAGGATGGAGGTGGTGATGTCTTTGGATCTCAGA
 5101 CCATGATGTTGGGATGGTCAGTGGGAGGTCTTGTCTTTGTAAGCTGATTGAATCCGGATG
 5161 GGATGAACTGAGCATGGCTTCCATAGGGCTTGGGATTCGGGGAAGGCTGAGTCCAATTCC
 5221 CCAAGCTTTACTGAAGACTGCTCTCCTTCTCATAGGTGTCCTGGTACCTGTGCCACCTGC
 5281 CAGTCAATGAATGACATGGCTGATGGGAATGGCGAGTCCTCTGACTCTTGTGTGCTCCCT
 5341 GAGTGTGGGTCTAGACTGGCGACCCCGTGGCTTGCCAGGGATGAGGAGCCTTTGGGGAGA
 5401 TTTTGCTGAGTGTCAGAGGACGTTTGAGGTCCCCCTCCTGGTGCCCTGACGTCAGGTGCC
 5461 AGGCGTGGCAGGCATGGCAAGGATGCTGGAGGTCTGCCTCCTGGTGCCCTGATGTCAGGT
 5521 GCCAGGCGTGGCAGGTGTGGTGCTTCTTGGAAGCACCGCGAGGAAGGGGTAGGCATGTTC
 5581 TGTAGCCCAAGGCGCCTAACTGGTAGGTAGTGGGTGGGACTACCTGAGCAGAGGCAGAGA
 5641 TATTTAAGGGGTAGTGGTCACAGCCACTCTGCTGGCAGTTGCTGCAGCTTGTGTTTGTTC
 5701 TGAAGCTGTCCTGAGTCGATTCTCCCAAGGTGAGTACTTCTGGCAGGCCGTCATTGGCAC
 5761 CATGGCAGAAGCTGGCAGCCCTGTTGGTGGCTGTGGTGTGGCACAGGAATCCCGGCAGCG
 5821 CAGGAAGACGGTTTGGCAGGACTTGCAAGAGGAGGCCCTATCAGCTTTCAACTAGAAGAG
 5881 ATACCTGTACTTTTAGGTGCTGGCCAGGCAAATGGGGATCCCAGATTGCTGAATTTGGGT
 5941 GTGGTTTCTGAATTGCAGGAAATCGCACTGGAGGGGGAGGGGCATGCCTCAAAGAGGTTC
 6001 ACCTGAGGCTCCAACCAAATAGCCTCACCACAGCTCCAGGAAGAACTAGGCTCCAGGGTA
 6061 CAGGGTGGAGGCATGCGCTCATCCAGCAGAAGGCCTCACACTGGACTCACTTTGTCACAG
 6121 CGCAGGATCCTAGCACAAGCATTTGAGAGGAACCCACGACCAGGCTGTGCTACCAGGGAG
 6181 GAGCTGGCACTTGAGACAGGGTTGCCCGAGGACATGATCCACACATGGTTGAAAAACAAA

 6241 AGAGCTCGGCGCCACAGGAGGGGCAGGCCCACAGCTCAAGATCAAGACTTGCTGGCCTCA
AH432                                                 ACTTGCTGGCCTCA
AH435                                                 ACTTGCTGGCCTCA
AH408                                                 ACTTGCTGGCCTCA
AH411                                                 ACTTGCTGGCCTCA
AH421                                                 ACTTGCTGGCCTCA
AH415                                                 ACTTGCTGGCCTCA
AH418                                                 ACTTGCTGGCCTCA

 6301 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH432 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH435 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH408 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH411 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH421 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH415 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG
AH418 CAAGTGTCCGGTGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAG

 6361 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH432 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH435 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH408 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH411 AGCTCATTGCCACAGGAAGAGGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH421 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH415 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC
AH418 AGCTCATTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTAC

 6421 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGC
AH432 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGTAGGC
AH435 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGC
AH408 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGC
AH411 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGC
AH421 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGGCACAGCCCCGTGGAGCAGGC
AH415 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGC
AH418 TCGCCCTCTTTCTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGC

 6481 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCCCTCTGGAACTTCTCCTTGATGAA
AH432 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCC
AH435 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCC
AH408 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCC
AH411 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCC
AH421 CAACAAGAGGCCCCCACTCAAGCAGGCAACGCAGGCT
```

```
AH415 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCC
AH418 CAAAAAGAGGTCCCCACTCAAGCAGGCAACGTAGGCC

 6541 CTGCAGGACGAAGTACAGGTGAAAGAGCATGTGCCAGACCCTTTGGATTTGGGTAGTGAT
 6601 CCTGGCGCCAGGGAGCCTGAAGGTTCCCAGGACAGCTTACAGAGCCTGGATGAAGCAGCA
 6661 AATTCAGGCTGGCATACCTCGGTCCCGAGCATCTCGTCAACCTTGTGCAGAGAGTCCCAG
 6721 CCTTCCCAAGTGGCACAGCCCTCTGGACCAGGACAAGCACAGGCCCCCACTCAAAGTGGG

 6781 TTCATAGACCCTCTGGAGCTCTTTCTCGATGAACTGCTGACTGAAGTCCAACTTGAGGAG
AH99                                                      CTTGAGGAG
AH173                                                     CTTGAGGAG
AH190                                                     CTTGAGGAG
AH187                                                     CTTGAGGAG
AH110                                                     CTTGAGGAG
AH100                                                     CTTGAGGAG
AH189                                                     CTTGAGGAG
AH170                                                     CTTGAGGAG
AH194                                                     CTTGAGGAG
AH108                                                     CTTGAGGAG
AH188                                                     CTTGAGGAG
AH105                                                     CTTGAGGAG
AH106                                                     CTTGAGGAG
AH107                                                     CTTGAGGAG

 6841 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH99  CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACGCCTGAG
AH173 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH190 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH187 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH110 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH100 CAGGGGCCTGCCCCCGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCCGAG
AH189 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH170 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH194 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH108 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH188 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH105 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH106 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG
AH107 CAGGGGCCTGCCCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAG

 6901 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH99  CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH173 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH190 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH187 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH110 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH100 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH189 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH170 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH194 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH108 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH188 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH105 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH106 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT
AH107 CTGCCTCTCACCCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGT

 6961 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH99  ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH173 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH190 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH187 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH110 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH100 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH189 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH170 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH194 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH108 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH188 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH105 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH106 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA
AH107 ACCCCTTGCATCTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGA

7021 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH99  GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH173 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH190 GAAGGTACCCATTTCAGCAAGGATGTGCACGAAAACGCAACTCCAGTAACTTCCCGAAAT
AH187 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
```

```
AH110 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCTCGAAAT
AH100 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH189 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH170 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH194 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH108 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH188 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH105 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH106 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT
AH107 GAAGGTACCCATTTCAGCAAGGATGTGCACAAAAACGCAACTCCAGTGACTTCCCGAAAT

7081  GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH99  GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH173 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH190 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH187 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH110 GCAAGGTGTCTTGCTAGCTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH100 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH189 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH170 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH194 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH108 GCAAGGTGTCTTGCTAACTATAAGAATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH188 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH105 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH106 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT
AH107 GCAAGGTGTCTTGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTT

7141  CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH99  CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH173 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH190 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH187 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH110 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH100 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH189 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH170 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH194 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH108 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH188 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH105 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH106 CTGCAGCGCCCAGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACA
AH107 CTGTAGCGCCCAGGGATTA

 7201 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH99  GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH173 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH190 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAAATAACCCCAAAAGAA
AH187 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH110 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH100 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH189 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH170 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH194 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH108 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH188 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH105 GATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA
AH106 GATGGAGAGGCCAATAGTGACTCCCCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAA

 7261 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH99  GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH173 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH190 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH187 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH110 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH100 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH189 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH170 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH194 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH108 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH188 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH105 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT
AH106 GGGCCACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACT


 7321 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH99  AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH173 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGAAGAGTGGACTCCTTCCC
```

```
AH190 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH187 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH110 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH100 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH189 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH170 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH194 AGGCTTCACTACAGGGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH108 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH188 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH105 AGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC
AH106 AGGCTTCACTACAGGGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCC

 7381 TGGGTTGTATGGACTGCTGCTATTCTTACAGATGCTTCGTGCAGAGCTGTGCAAGGTTTT
AH99  TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH173 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH190 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH187 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH110 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH100 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH189 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH170 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH194 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH108 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH188 TGGATTGTATGGACTGCTGCTATTCTTACAG
AH105 TGGGTTGTATGGACTGCTGCTATTCTTACAG
AH106 TGGGTTGTATGGACTGCTGCTATCCTTACAG

 7441 ACAGGCCAGTCTTTTAATATCTACTACCCATAGGTCTTTTGTTTGTTTTCTTTTTCTTTT
 7501 TCACTTTCTTTTTCATTTTTTTTTCTTTATTTTTTTGGGTGGGTTGTCTTTGCTGGGTTT
 7561 GGTTTGGTTCTGTGTAGTTTGTTTCCATTGCTTTCAATAAAGTTTATTGATTTTGACAAA
 7621 ATTTGTTTGTGTGTTTGTGTACTGTGTTGTGGGATGAGGGGTGGGTTGAAGAGGCTGTTT
 7681 TGTTCTACCCGGAGAATGTGCATGAGAATTCCATTTAAAGGGATGAAAGACATAGGAAGG
 7741 TATTCAGGAGGGATGCAGAGAGGCCAGGTAGAAAAGGGGAGAGAACAAGGAAGGCTGGGC
 7801 TGTGTCCAGTGGAGAGACGGGCTTGCTTGATCTAGTAACAAGCAGGTGGAATGTCAACTG
 7861 AAAGAAGAATGTGTGTAAGGCATAACCTTGTCAAGAAAGATCATCTTGGGAATACAGTGC
 7921 AGCTAGGTGTGAACATCTTATTCCAAGCTTGAAACACAGCCATCAGTCTATTCTTCATCC
 7981 ATCCATCCATCTATCCATTCTTGCATCCATTCATCCATGCATGCCTAAATAATCTGGAGG
 8041 TCTTTCAATCCTTTTGGGATTAATCCCCAGCACCCTCCCTTATCCTTCTTTTGATCTTGA
 8101 GATTCCCAGGTGTTCAAGGTCATGCTGTTTACATGGAGCTCCAAGTTGATCCTTGACCCA
 8161 CATTGGAAGGACTCGGTATGTTTAATATTCTACAATGACCGACAATTCTACAGAGAGCCT
 8221 TATGGCATGCCAGCAGGACAAAACAATCTCTCATTTGCTGGCCATCACCTCAGGACAACT
 8281 TATTTGAAGTGTCTCCAGTGTTCAAGGCTAACTCCAGAGATCTAAGAGCACAGAACATAC
 8341 CGCCAGCAAACACAGCACATGCAAGAAGATGATCAACTCTTTTCTTCACCCTTCTCCATA
 8401 GAAAGTGCAAAACCTACTGGTGTCTCAAGCTTCCAGGCTCCTTTTCATACAGTCTGTGAA
 8461 AGAAACACCTTGTGAGGTGTCTCCCTCTCTCTCTGTCTCTGTCTGTCTGTCTCTCTGT
 8521 CTCTCTCTCTCTCCCTCCATTCCTCTTTTGCTCCCCCTCCCATTTCCCTCCTTGCCTCCA
 8581 TTTCACCATCTCTTCCACTCTCTGTCTCCATCCCCATCCTTCTACCCTCCCATATTCACT
 8641 CCCCCCATCCACTTTCTACCTCCCTACTTCCCTATCTCTCTCTATGGATTCTTCCCTTCC
 8701 TTCTGCACTCTGTCACTCTCTCCCTACCACCCTCCACCCTCTGTCCCTAAATCCCTTCCC
 8761 CCCTTCTCTCCACATCTGTGTTTGTCTCTCTCTTCGTGTCTTCCTCTGCCCCTAACCCCA
 8821 CCCATGGTCGTGACTTTATCTTCCCTTAGGATATTTGTGAGCATGATGTGTGTGTTTTG
 8881 TGTGTGTGTGTGTGTTTGTGCTTGTGTGTGCCCGCATGTGTGTACGTGTTTGTGTGTGCG
 8941 TGTGTGCGTGTGTGCGTGCATGCATGCAAATGTGTGTATGTGTGTGTGGTCTGAGGGT
 9001 GTGCCTGTTCACAATTGTCTCTGTGTGTTGCTGCCAGGTGCCTAGGGGCTGTTTGGACTT
 9061 TCATTTTGATCTCAGTACAGGTGATGTTCCCTCTTGTCCTCATAGCACAGTCAGACTTGG
 9121 AAAGTCAAGGAAGGGGGTCTGAAACACTCTAGAGATAGGATGGAGGTGGTGATGTCTTTG
 9181 GATCTCAGACCATGATGTTGGGATCGTCAGTGGGTGGTCTTGTCTTTGTAAGCTGATTGA
 9241 ATCCGGATGGGATGAACTGAGCATGGCTTCCATAGGGCTTGGGATTCCTGGAAGGCTGAG
 9301 TCCAATTCCCCAAGCTTTACTGAAGACTGCTCCCCTTCTCATAGGTGTCCTGGTACCTGT
 9361 GCCACCTGCCAGTCAATGAATGACATGGCTGATGGGAATGGCGAGTCCTCTGACTCTTGT
 9421 GTGCTCCCTGGGTGTGGGTCTAGACTGGCGACCCCGTGGCTTGCCAGGGATGAGGAGCCT
 9481 TTGGGGAGATTTTGCTGAGTGTCAGAGGATGCTTGAGGTCCGCCTCCTGGTGCCCTGATG
 9541 TCAGGTGCCAGGCGTGGCAGGCTTGGCGCTTCTTGGGGGCAATGCGAGAAAGGGGTAGGC
 9601 ATGTTCTGTAGCCCAAGGCGCCTAACTGGTAGGTAGTGGGCGGGACTACCTGAGCAGAGG

 9661 CAGAGGTATTTAAGGGGCAGTGGTCACAGCCACTCTGCTGGCAGTTGCTGCAGCTTGTGT
AH379                              GCCACTCTGCTGGCAGTTGCTGCAGCTTGTGT
AH377                              GCCACTCTGCTGGCAGTTGCTGCAGCTTGTGT
AH384                              GCCACTCTGCTGGCAGTTGCTGCAGCTTGTGT
AH378                              GCCACTCTGCTGGCAGTTGCTGCAGCTTGTGT
AH381                              GCCACTCTGCTGGCAGTTGCTGCAGCTTGTGT
AH383                                                          GTGT

 9721 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT
AH379 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT
AH377 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT
AH384 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT
```

```
AH378 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT
AH381 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT
AH383 TTGTTCTGAAGCTGTCCTGAGTCGAGTCTCCCAAGGTAAGGACTCCTGGGAGGCCGTCAT

 9781 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
AH379 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
AH377 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
AH384 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
AH378 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
AH381 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
AH383 TGGCACCATGGCAGAAGCTGGCAGCCCTGTTGGTGGCAGTGGTGTGGCACGGGAATCCCG
          M  A  E  A  G  S  P  V  G  G  S  G  V  A  R  E  S  R

 9841 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
AH379 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
AH377 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
AH384 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
AH378 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
AH381 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
AH383 GCGGCGCAGGAAGATGGTTTGGCAGGCCTGGCAAGAGCAGGCCCTGCTATCAACTTTTAA
    R  R  R  K  M  V  W  Q  A  W  Q  E  Q  A  L  L  S  T  F  K

 9901 GGAGAAGAGATACCTGAGCTTCAAGGAGAGGAAGGAGCTGGCCAAGCGAATGGGGGTCTC
AH379 GGAGAAGAGATACCTGAGCTT
AH377 GGAGAAGAGATACCTGAGCTT
AH384 GGAGAAGAGATACCTGAGCTT
AH378 GGAGAAGAGATACCTGAGCTT
AH381 GGAGAAGAGATACCTGAGCTT
AH383 GGAGAAGAGATACCTGAGCTT
    E  K  R  Y  L  S  F  K  E  R  K  E  L  A  K  R  M  G  V  S

 9961 AGATTGCCGCATCCGCGTGTGGTTTCAGAACCGCAGGAATCGCAGTGGAGAGGAGGGGCA
      D  C  R  I  R  V  W  F  Q  N  R  R  N  R  S  G  E  E  G  H

10021 TGCCTCAAAGAGGTCCATCAGAGGCTCCAGGCGGCTAGCCTCGCCACAGCTCCAGGAAGA
      A  S  K  R  S  I  R  G  S  R  R  L  A  S  P  Q  L  Q  E  E

10081 GCTTGGATCCAGGCCACAGGGTAGAGGCCTGCGCTCATCTGGCAGAAGGCCTCGCACTCG
      L  G  S  R  P  Q  G  R  G  L  R  S  S  G  R  R  P  R  T  R

10141 ACTCACCTCGCTACAGCGCAGGATCCTAGGGCAAGCCTTTGAGAGGAACCCACGACCAGG
      L  T  S  L  Q  R  R  I  L  G  Q  A  F  E  R  N  P  R  P  G

10201 CTTTGCTACCAGGGAGGAGCTGGCGCGTGACACAGGGTTGCCCGAGGACACGATCCACAT
      F  A  T  R  E  E  L  A  R  D  T  G  L  P  E  D  T  I  H  I

10261 ATGGTTTCAAAACCGAAGAGCTCGGCGGCGCCACAGGAGGGGCAGGCCCACAGCTCAAGA
      W  F  Q  N  R  R  A  R  R  R  H  R  R  G  R  P  T  A  Q  D

10321 TCAAGACTTGCTGGCGTCACAAGGGTCGGATGGGGCCCCTACAGGTCCGGAAGGCAGAGA
      Q  D  L  L  A  S  Q  G  S  D  G  A  P  T  G  P  E  G  R  E

10381 GCGTGAAGGTGCCCAGGAGAGCTTGTTGCCACAGGAAGAAGCAGGAAGTACGGGCATGGA
      R  E  G  A  Q  E  S  L  L  P  Q  E  E  A  G  S  T  G  M  D

10441 TACCTCGAGCCCTAGCGACTTGCCCTCCTTCTGCAGAGAGTCCCAGCCTTTCCAAGTGGC
      T  S  S  P  S  D  L  P  S  F  C  R  E  S  Q  P  F  Q  V  A

10501 ACAGCCCCGTGGAGCAGGCCAACAAGAGGCCCCCACTCAAGCAGGCAACGCAGGCTCTCT
      Q  P  R  G  A  G  Q  Q  E  A  P  T  Q  A  G  N  A  G  S  L

10561 GGAACCCCTCCTTGATCAGCTGCTGGATGAAGTCCAAGTAGAAGAGCCTGCTCCAGCCCC
      E  P  L  L  D  Q  L  L  D  E  V  Q  V  E  E  P  A  P  A  P

10621 TCTGAATTTGGATGGAGACCCTGGTGGCAGGGTGCATGAAGGTTCCCAGAAGAGCTTTTG
      L  N  L  D  G  D  P  G  G  R  V  H  E  G  S  Q  K  S  F  W

10681 GCCACAGGAAGAAGCAGGAAGTACAGGCATGGATACTTCAAGCCCCAGCGACTCAAACTC
      P  Q  E  E  A  G  S  T  G  M  D  T  S  S  P  S  D  S  N  S

10741 CTTCTGCAGAGAGTCCTTTCCTTCCCAAGTGGCACAGCCCTGTGGATCGGGCCAAGAAGA
      F  C  R  E  S  F  P  S  Q  V  A  Q  P  C  G  S  G  Q  E  D

10801 TGCCCGCACTCAAGCAGACAGCACAGGCCCTCTGGAACTCCTCCTCCTTGATCAACTGCT
      A  R  T  Q  A  D  S  T  G  P  L  E  L  L  L  L  D  Q  L  L

10861 GGACGAAGTCCAAAAGGAAGAGCATGTGCCAGTCCCACTGGATTGGGGTAGAAATCCTGG
```

```
       D  E  V  Q  K  E  E  H  V  P  V  P  L  D  W  G  R  N  P  G

10921 CAGCAGGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGCAGCAAATTC
       S  R  E  H  E  G  S  Q  D  S  L  L  P  L  E  E  A  A  N  S

   10981 GGGCATGGATACCTCGATCCCTAGCATCTGGCCAACCTTCTGCAGAGAATCCCAGCCTCC
       G  M  D  T  S  I  P  S  I  W  P  T  F  C  R  E  S  Q  P  P

11041 CCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCCCCCACTCAAGGTGGGAACAC
       Q  V  A  Q  P  S  G  P  G  Q  A  Q  A  P  T  Q  G  G  N  T

11101 GGACCCCCTGGAGCTCTTGCTCTATCAACTGTTGGATGAAGTCCAAGTAGAAGAGCATGC
       D  P  L  E  L  L  L  Y  Q  L  L  D  E  V  Q  V  E  E  H  A

11161 TCCAGCCCCTCTGAATTGGGATGTAGATCCTGGTGGCAGGGTGCATGAAGGTTCGTGGGA
       P  A  P  L  N  W  D  V  D  P  G  G  R  V  H  E  G  S  W  E

11221 GAGCTTTTGGCCACAGAGAGAAGCAGGAAGTACAGGCCTGGATACTTCAAGCCCCAGCGA
       S  F  W  P  Q  R  E  A  G  S  T  G  L  D  T  S  S  P  S  D

11281 CTCAAACTCCTTCTGCAGAGAGTCCCAGACTTCCCAAGTGGCACAGCCCTGTGGATCGGG
       S  N  S  F  C  R  E  S  Q  T  S  Q  V  A  Q  P  C  G  S  G

11341 CCAAGAAGATGCCCGCACTCAAGCAAACAGCACAGGCCCTCTGGAACTCCTCCTCCTTGA
AH442                                  GGCCCTCTGGAACTCCTCCTCCTTGA
AH441                                  GGCCCTCTGGAACTCCTCCTCCTTGA
       Q  E  D  A  R  T  Q  A  N  S  T  G  P  L  E  L  L  L  L  D

11401 TCAACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAG
AH442 TCAACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAG
AH441 TCAACTGCTGGACGAAGTCCAAAAGGAAGAGCATGTGCCAGCCCCACTGGATTGGGGTAG
       Q  L  L  D  E  V  Q  K  E  E  H  V  P  A  P  L  D  W  G  R

11461 AAATCCTGGCAGCAGGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGC
AH442 AAATCCTGGCAGCAGGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGC
AH441 AAATCCTGGCAGCAGGGAGCATGAAGGTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGC
       N  P  G  S  R  E  H  E  G  S  Q  D  S  L  L  P  L  E  E  A

11521 AGCAAATTCGGGCATGGATACCTCGATCCCTAGCATCTGGCCAACCTTCTGCAGAGAATC
AH442 AGCAAATTCGGGCATGGATACCTCGATCCCTAGCATCTGGCCAACCTTCTGCAGAGAATC
AH441 AGCAAATTCGGGCATGGATACCTTGATCCCTAGCATCTGGCCAACCTTCTGCAGAGAATC
       A  N  S  G  M  D  T  S  I  P  S  I  W  P  T  F  C  R  E  S

11581 CCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCAAAGCACAGGCCCCGACTCAAGG
AH442 CCAGCCTCCCCAAGTGGCACTGCCCTCTGGACCAGGCCAAGCACAGGCCCCGACTCAAGG
AH441 CCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAGGCCAAGCACAGGCCCCGACTCAAGG
       Q  P  P  Q  V  A  Q  P  S  G  P  G  Q  A  Q  A  P  T  Q  G

11641 TGGGAACACGGACCCTCTGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGA
AH442 TGGGAACACGGACCCTCTGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGA
AH441 TGGGAACACGGACCCTCTGGAGCTCTTCCTTGATCAACTGCTGACCGAAGTCCAACTTGA
AH177                                                      CTTGA
AH193                                                      CTTGA
       G  N  T  D  P  L  E  L  F  L  D  Q  L  L  T  E  V  Q  L  E

11701 GGAGCAGGGGCCTTCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATGGACACAACACC
AH442 GGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAA
AH441 GGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAA
AH177 GGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATGGACACAACACC
AH193 GGAGCAGGGGCCTGCCCCTGTGAATGTGGAGGAAACATGGGAGCAAATGGACACAACACC
       E  Q  G  P  A  P  V  N  V  E  E  T  W  E  Q  M  D  T  T  P

11761 TGATCTGCCTCTCACTCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGA
AH177 TGATCTGCCTCTCACTCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGA
AH193 TGATCTGCCTCTCACTCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGA
       D  L  P  L  T  P  E  E  Y  Q  T  L  L  D  M  L  .

11821 CAGTACCTCTTGCTTCTAGAAACCGGAAAGGCCAAAGTCCTGAAGAGACCCGATTTGGAA
AH177 CAGTACCTCTTGCTTCTAGAAACCCGAAAGGCCAAAGTCCTGAAGAGACCCGATTTGGAA
AH193 CAGTACCTCTTGCTTCTAGAAACCCGAAAGGCCAAAGTCCTGAAGAGACCCGATTTGGAA

11881 CTGGAGAAGTGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTCCAGTGACTTCC
AH177 CTGGAGAAGTGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTCCAGTGACTTCC
AH193 CTGGAGAAGTGACCCATCCCAGCAAGGATGTGCATCAAAAACCCAACTCCAGTGACTTCC
```

```
11941 CGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGATAATAATGAAG
AH177 CGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGATAATAATGAAG
AH193 CGAAATGCAAGGTGTCTCGCTAACTATAAGGATTGATTGCAGGTGGGGATAATAATGAAG

12001 TGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACA
AH177 TGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACA
AH193 TGCCTTCTCCAGGGCCCGGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACA

12061 GGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACGCCA
AH177 GGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACGCGGAGC:TAAAGATAACGCCA
AH193 GGGACAGATGGAGAGGCCAATAGTGACTCCTCAACAACACGGAGC:TAAAGATAACGCCA

12121 AAAGAAGGGCTACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACT
AH177 AAAGAAGGGCTACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACT
AH193 AAAGAAGGGCTACACCAAGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACT

12181 AGGACTAGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATGGGGAGAGTGGACTC
AH177 AGGACTAGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATGGGGAGAGTGGACTC
AH193 AGGACTAGGCTTCACTACAGAGGACACACACTCCCTGAGGGCAATGGGGAGAGTGGACTC

12241 CTTCCTTGGCTTATATGGACTGCTGTTATCCTTACAGATGCTTCGTGCAGAGCTGTGCAA
AH177 CTTCCTTGGCTTATATGGACTGCTGTTATCCTTACAG
AH193 CTTCCTTGGCTTATATGGACTGCTGTTATCCTTACAG

12301 GGTTTTACAGGCCAGTCTTTTAATATCTACTACCCATAGGTCTTTTGTTTGTTTTCTTTT
12361 TCTTTTTCACTTTCTTTTTCATTTTTTTTCTTTTTCTTCTTTTAAGGGGTGGGTTGGCTT
12421 TGTTGGGTTTGGTTTGGTTTTGTGTAGTTTGTTTCCATTGCTTTCAATAAAGTTTATTGA
12481 TTTTAACAAAATTTGTTTGTGTGTTTGTGTGCTGTTTTGTGTGATGAGGGGTGGGTTGAA
12541 TAGGCTGTTTTGTTCTACCCGGAGAACGTGCATGAGAATTCCATTTAAAGGGATTAAAGA
12601 CATAGGAAGGTATTCAGGAGGGATGCAGAGAGGCCAGGTAGAAAAGGGGAGAGAACAAGG
12661 AAGGCTGGGCTATGTCCAGTGGAGAGACGGGCTTGCTTGATCTAGTAACAAGCAGGTGGA
12721 ATGTCAACTGAAAGAAGAATGTGTGTAAGGCATAACCTTGTCAAGAAAGATCATCTTGGG
12781 AATACAGTGCAGCTAGGTGTGAACATCTTATTCCAAGCTTGAAACACAGCCATCAGTCTA
12841 TTCTTCATCCATCCATCCATCCATCCATCCATGCCTAAATCAATCTGTAGATCTTT
12901 CAATCCTATTGGGATTAATCCCCAGCACCCTCCCTAATCCCTCTTTTGATCTTGAGATTC
12961 CCAGGTGTTCAAGGTCATACTGTTTATATGGAGCTCCAAGTTGATCCTTGACCCACATTG
13021 GAAGGAGACGGTATGTTTACCATTCTACAATGACCGACAATTCTTTTTTTTTTTTTTTTT
13081 TTTTTCGTTTTCAGCTTTTTATTTTTTTATGACCGACAATTCTACAGAGAGCCTTAAAGC
13141 ATGCCAGCAGGACAAAACAATCTCTCATTTGCTGGCCATCACTTCAGGACAACTTATTTG
13201 AAGTGTCTCCAGTGTTCAAGGCTAACTCCAGAGATCTAAGAGCACAGAACATACCGCCAG
13261 CAAACACAGCACATGCAAGAAGACGATCAACTCTTTTCTTCAACCTGCTCCATAGAAAGT
13321 GCACAACCTACTGGTGTCTCAAGCTTCCAGGCTCCTTTTTATACAGTCTGTGAAAGAAAC
13381 ACCTTGTGAGGTGTCTCCCTCTCTCTCTGTCTCTGTCTGTCTGTCTGTCTCTGTCTGTCT
13441 GTCTGTCTCTGTCTGTCTCTCTCTCCCTCCATTCCTCTTTTGCTCCCCCTCCATTTCCCT
13501 CCCTGCCTCCATTTCACCATCTCTTCCCCTCTCTGTCTCCATCCCCATTCTTCTACCCTC
13561 CCATATTCACTCCCCCCCATCCACTTTCTACCTCCCTACTTCCCTATCTCTCTGTATCGA
13621 TTCTTCCCTTCCTTCTGCACTCTGTCACTCTCTCCTTACCACCCTCCACCCTCTGTCCCT
13681 AAATCCCTTCCCCCCTTCTCTCCACATCTGTGTTTGTCTCTCTCTTCGTGTCTTCCTCTG
13741 CCCCTAACCCCAACCATGGTCGTGACTTTATCTTCACTGGAGTTGGGTTTTTGATGCACA
13801 TCTGGCCAACCTTCTGCAGAGAATACCAGCCTCCCCAAGTGGCACAGCCCTCTGGACCAG
13861 GCCAAGCACAGGCCCCCACTCAAGGTGGGAACACGGACCCCCTGGAGCTCTTGCTCTATC
13921 AACTGTTGGATGAAGTCCAAGTAGAAGAGCATGCTCCAGCCCGTCAGAATTGGGATGTAG
13981 ATCCTGGTGGCAGGGTGCATGAAGGTTCGTGGGAGAGCTTTTGGCCACAGGAAGAAGCAG
14041 GAAGTACAGGCCTGGATACTTCAAGCCCCAGCGACTCAAACTCCTTCTGCAGAGAGTCCC
14101 AGACTTCCCAAGTGGCACAGCCCTGTGGATCGGGCCAAGAAGATGCCCGCACTCAAGCAG

14161 ACAGCACAGGCCCTCTGGAACTCCTCCTCCTTGATCAACTGCTGGACGAAGTCCAAAAGG
AH145         GGCCCTCTGGAACTCCTCCTCCTTGCTCTNCTGCTGGACGAAGTCCAAAAGG

14221 AAGAGCATGTGCCAGTCCCACTGGATTGGGGTAGAAATCCTGGCAGCAGGGAGCATGAAG
AH145 AAGAGCATGTGCCAGTCCCACTGGATTGGGGTAGAAATCCTGGCAGCAGGGAGCATGAAG

14281 GTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGCAGCAAATTCGGGCATGGATACCTCGA
AH145 GTTCCCAGGACAGCTTACTGCCCCTGGAGGAAGCAGCAAATTCGGGCATGGATACCTCGA

14341 TCCCTAGCATCTGGCCAACCTTCTGCAGAGAATCCCAGCCTCCCCAAGTGGCACAGCCCT
AH145 TCCCTAGCATCTGGCCAACCTTCTGCAGAGAATCCCAGCCTCCCCAAGTGGCACAGCCCT


14401 CTGGACCAGGCCAAGCACAGGCCCCCACTCAAGGTGGGAACACGGACCCTCTGGAGCTCT
AH145 CTGGACCAGGCCAAGCACAGGCCCCCACTAAAGGTGGGAACACGGACCATCTGGAGCTCT

14461 TCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
AH145 TCCTTGATCAACTGCTGACCGAAGTCCAACTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
AH174                              CTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
AH98                               CTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
AH172                              CTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
```

```
AH112                                  CTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
AH175                                  CTTGAGGAGCAGGGGCCTGCCCCTGTGAATG
AH103                                  CTTGAGGAGCAGGGGCCTGCCCCTGTGAATG

14521 TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGAAT
AH145 TGGAGGAA
AH174 TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGAAT
AH98  TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGGAT
AH172 TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGAAT
AH112 TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGAAT
AH175 TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGAAT
AH103 TGGAGGAAACATGGGAGCAAATGGACACAACACCTGATCTGCCTCTCACTCCAGAAGAAT

14581 ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACAGTACCTCTTGCTTCTAGAAACCCG
AH174 ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACTGTACCTCTTGCTTCTAGAAACCCG
AH98  ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACTGTACCTCTTGCTTCTAGAAACCCG
AH172 ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACTGTACCTCTTGCTTCTAGAAACCCG
AH112 ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACTGTACCTCTTGCTTCTAGAAACCCG
AH175 ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACTGTACCTCTTGCTTCTAGAAACCCG
AH103 ATCAGACTCTTCTAGATATGCTCTGACTCCCCGACTGTACCTCTTGCTTCTAGAAACCCG

14641 AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG
AH174 AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG
AH98  AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG
AH172 AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG
AH112 AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG
AH175 AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG
AH103 AGAGGCCAAAGTCCTGAAGAGACCCGATTTGGAACTGGAGAAGTGACCCATCCCAGCAAG

14701 GATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCTCGCTAACTA
AH174 GATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGCGTCTCCCTAACTA
AH98  GATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCTCCCTAACTA
AH172 GATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCTCCCTAACTA
AH112 GATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCTCCCTAACTA
AH175 GATGTGCATCAAGAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCTCCCTAACTA
AH103 GATGTGCATCAAAAACCCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCTCCCTAACTA

14761 TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG
AH174 TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG
AH98  TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG
AH172 TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG
AH112 TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG
AH175 TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG
AH103 TAAGGATTGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTCCAGGGCCCGGGGATTAG

14821 GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC
AH174 GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC
AH98  GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC
AH172 GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC
AH112 GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC
AH175 GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC
AH103 GAAATCAGCCCTGAAAGTCGAGACTCTGCTACAGGGACAGATGGAGAGGCCAATAGTGAC

14881 TCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC
AH174 TCCTCAACAACAAGGAGTCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC
AH98  TCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC
AH172 TCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC
AH112 TCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC
AH175 TCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC
AH103 TCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCAAGTGACTGGC

14941 TCCAGTGGACCCCAGGAAATCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC
AH174 TCCAGTGGACCCCAGGAAGTCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC
AH98  TCCAGTGGACCCCAGGAAATCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC
AH172 TCCAGTGGACCCCAGGAAATCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC
AH112 TCCAGTGGACCCCAGGAAATCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC
AH175 TCCAGTGGACCCCAGGAAATCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC
AH103 TCCAGTGGACCCCAGGAAATCACACAGGACACTAGGACTAGGCTTCACTACAGAGGACAC

15001 ACACTCCCTGAGGGCAATGGGGAGAGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
AH174 ACACTCCCTGAGGGCAATGGGGAGAGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
AH98  ACACTCCCTGAGGGCAATGGGGAGAGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
AH172 ACACTCCCTGAGGGCAATGGGGAGTGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
AH112 ACACTCCCTGAGGGCAATGGGGAGAGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
AH175 ACACTCCCTGAGGGCAATGGGGAGAGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
AH103 ACACTCCCTGAGGGCAATGGGGAGAGTGGACGCCTTCCTTGGCTTATATGGACTGCTGTT
```

```
15061 ATCCTTACAGATGCTTCGTGCAGAGCTGTGCAAGGTTTTACAGGCCAGTCTTTTAATATC
AH174 ATCCTTACAG
AH98  ATCCTTACAG
AH172 ATCCTTACAG
AH112 ATCCTTACAG
AH175 ATCCTTACAG
AH103 ATCCTTACAG

15121 TACTACCCATAGGTCTTTTGTTTGTTTTCTTTTTCTTTTTCACTTTCTTTTTCATTTTTT
```

Figure 5.10 Alignment of RT-PCR clones matching BAC 142C15

The sequences of the RT-PCR were aligned to the BAC clone 142C15 using Sequencher. The predicted amino acid sequence of the ORF is shown underneath. Nucleotide changes from the consensus are highlighted in red.

5.2.4 Nested RT-PCR on mouse testis RNA

During the work described in this chapter, amplification of full length transcripts from the human array was achieved from human testis RNA using a nested RT-PCR reaction (Snider *et al.*, 2010). Therefore, nested primers were designed to amplify a 2143bp product from the mouse array, to determine whether full length Dux transcripts are expressed in mouse testis.

Mouse testis RNA was treated with DnaseI as described previously and RT-PCR was carried out using Superscript III and BioTaq polymerase. The nested RT-PCR reaction was also performed with mouse midterm embryo RNA and RNA extracted from ES cells that were kindly provided by Dr Paul Scotting. RNA extraction was carried out as previously described (section 2.4.1). Nested primers are shown in Figure 5.11a.

Amplification of a fragment of around 2000bp was seen from testis RNA (Figure 5.11b). If this product represented Dux it would be the largest transcript identified in mouse this far. The RT-PCR product was cloned into pSMART and sequenced. Sequencing from one end of the clone failed and so sequence data from only half the length of these PCR products could be analysed (approximately 1000bp).

AH648 and AH642 were aligned with the BAC clone 142C15 and had 33 and 26 variants respectively so were unlikely to originate from here. They were then aligned with the genomic consensus sequence and had 14 and 5 variants respectively (Table 5.6). The number of genomic sequences in which each variant is found is indicated in Table 5.6. Of the 14 variants identified in AH648, 12 of them were also present in a single genomic clone which is highly variant compared with the other genomic sequences.

The three sequences, AH635, AH631 and AH634 were also aligned with the genomic consensus. The first 556bp of these sequences aligned with the consensus sequence between positions 3584 and 4137, with 31 base changes and a 4p deletion. The other half of the sequence aligned with the genomic consensus between positions 1628 and 2141, with 73 changes from the consensus. The two halves of these sequences (section 1 and section 2) were then aligned with the BAC clone. The first 556 base pairs (section 1) aligned with repeat number 2 of the BAC clone, AH634 was identical to the BAC sequence, AH635 had one variant

position and AH631 had 2 variant positions. The final 497bp (section 2) aligned with repeat 2; however it aligns 615bp upstream of the first half of the sequence (Figure 5.12). AH631 had 1 variant position, AH634 had 3 variant positions and AH635 had 5 variant positions. This data suggests that these three RT-PCR products originate from a recombinant repeat sequence which has not yet been identified in genomic sequencing of the region. The sequence was used in a BLAST search of the mouse genome and aligned with 99% identity with the mouse chromosome 10 contig (NT 039500.7) across the whole length of the sequence, confirming the presence of a recombinant repeat sequence in the mouse genome.

a

b

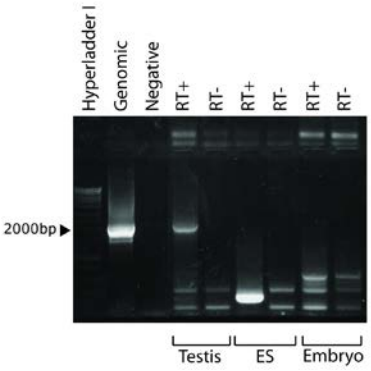


Figure 5.11 Amplification of full length mouse Dux transcripts by nested RT-PCR

a) Schematic representation of the mouse Dux array with nested RT-PCR primers indicated. b) RT-PCR of mouse RNA from testis and midterm embryo (Ambion) and from RNA extracted from mouse ES cells (kindly provided by Dr Paul Scotting). Amplification of the correct 'full length' product was only seen from Testis RNA.

Position	2825	2866	2869	3050	3095	3240	3245	3247	3263	3270	3277	3474	3490	3568	3587	3678	3696	3735	4153
Clone	T	A	T	G	C	C	T	T	A	C	G	A	G	C	C	C	C	G	C
AH648	C	T	G	G	T	A	C	T	G	A	DEL	T	A	T	T	C	T	ND	NA
AH642	ND	ND	ND	C	C	C	T	C	A	C	G	A	G	C	C	T	C	A	A
Coding Changes	V-A	T-S	F-V	R-T	A-V		L-P	F-L	D-G		Frameshift		G-S	P-S	T-I	Outside ORF			
No. of genomic sequences	1	1	1		1	1		10	1	1	1		1	1	1	1			

Table 5.5 Variants in the mDux sequences AH648 and AH642

Sequence variants are shown in orange, blue blocks are the same base as the genomic consensus sequence. Coding changes are shown underneath. The number of genomic sequences amplified by other members of Jane Hewitt's group that contain each variant is indicated.

Figure 5.12 Schematic representation of the alignment of AH631, AH634 and AH635 with BAC clone 142C15

The region of BAC clone 142C15 which contains mouse Dux repeats is shown. Repeats are indicated in orange. Repeats 3 and 5 are full length repeats, repeats 1, 2, 4and 6 are partial repeats. The positions of sections 1 and 2 of the clones AH631, AH634 and AH635 are indicated in purple.

```
106621            TAGTGGTCACAGCCACTCTGCTGGCAGTTGCTGCAGCTTGTGTTTGTTCTGAAGCTGTCC

106681            TGAGTCGATTCTCCCAAGGTGAGTACTTCTGGCAGGCCGTCATTGGCACCATGGCAGAAG
AH634 Section 2                                           TCATTGGCACCATGGCAGAAG
AH631 Section 2                                           TCATTGGCACCATGGCAGAAG

106741            CTGGCAGCCCTGTTGGTGGCTGTGGTGTGGCACAGGAATCCCGGCAGCGCAGGAAGACGG
AH634 Section 2   CTGGCAGCCCTGTTGGTGGCTGTGGTGTGGCACAGAAATCCCGGCAGCGCAGGAAGACGG
AH635 Section 2                                                   AGAAGACGG
AH631 Section 2   CTGGCAGCCCTGTTGGTGGCTGTGGTGTGGCACAGGAATCCCGGCAGCGCAGGAAGACGG

106801            TTTGGCAGGACTTGCAAGAGGAGGCCCTATCAGCTTTCAACTAGAAGAGATACCTGTACT
AH634 Section 2   TTTGGCAGGACTTGCAAGAAGAGGCCCTATCAGCTTTCAACTAGAAGAGATACCTGTACT
AH635 Section 2   TTTGGCAGGACTTGCAAGAAGAGGCCCTATCAGCTTTCAACTAGAAGAGATACCTGTACT
AH631 Section 2   TTTGGCAGGACTTGCAAGAGGAGGCCCTATCAGCTTTCAACTAGAAGAGATACCTGTACT

106861            TTTAGGTGCTGGCCAGGCAAATGGGGATCCCAGATTGCTGAATTTGGGTGTGGTTTCTGA
AH634 Section 2   TTTAGGTGCTGGCCAGGCAAATGGGGATCCCAGATTGCTGAATTTGGGTGTGGTTTCTGA
AH635 Section 2   TTTAGGTGCTGGCCAGGCAAAAGGGGATCCCAGATTGCTGAATTTGGGTGTGGTTTCTGA
AH631 Section 2   TTTAGGTGCTGGCCAGGCAAATGGGGATCCCAGATTGCTGAATTTGGGTGTGGTTTCTGA

106921            ATTGCAGGAAATCGCACTGGAGGGGGAGGGGCATGCCTCAAAGAGGTTCACCTGAGGCTC
AH634 Section 2   ATTGCAGGAAATCGCACTGGAGGGGGAGGGGCATGCCTCAAAGAGGTTCACCTGAGGCTC
AH635 Section 2   ATTGCAGGAAATCGCACGGGAGGGGGAGGGGCATGCCTCAAAGAGGTTCACCTGAGGCTC
AH631 Section 2   ATTGCAGGAAATCGCACTGGAGGGGGAGGGGCATGCCTCAAAGAGGTTCACCTGAGGCTC

106981            CAACCAAATAGCCTCACCACAGCTCCAGGAAGAACTAGGCTCCAGGGTACAGGGTGGAGG
AH634 Section 2   CAACCAAATAGCCTCACCACAGCTCCAGGAAGAACTAGGCTCCAGGGTACAGGGTGGAGG
AH635 Section 2   CAACCAAATAGCCTCACCACAGCTCCAGGAAGAACTAGGCTCCAGGGTACAGGGTGGAGG
AH631 Section 2   CAACCAAATAGCCTCACCACAGCTCCAGGAAGAACTAGGCTCCAGGGTACAGGGTGGAGG

107041            CATGCGCTCATCCAGCAGAAGGCCTCACACTGGACTCACTTTGTCACAGCGCAGGATCCT
AH634 Section 2   CATGCGCTCATCCAGCAGAAGGCCTCACACTGGACTCACTTTGTCACAGCGCAGGATCCT
AH635 Section 2   CATGCGCTCATCCAGCAGAAGGCCTCACACTGGACTCACTTTGTCACAGCGCAGGATCCT
AH631 Section 2   CATGCGCTCATCCAGCAGAAGGCCTCACACTGGACTCGCTTTGTCACAGCGCAGGATCCT

107101            AGCACAAGCATTTGAGAGGAACCCACGACCAGGCTGTGCTACCAGGGAGGAGCTGGCACT
AH634 Section 2   AGCACAAGCATTTGAGAGGAACCCACGACCAGGCTGTGCTACCAGGGAGGAGCTGGCACT
AH635 Section 2   AGCACAAGCATTTGAGAGGAACCCACGACCAGGCTGTGCTACCAGGGAGGAGCTGGCACT
AH631 Section 2   AGCACAAGCATTTGAGAGGAACCCACGACCAGGCTGTGCTACCAGGGAGGAGCTGGCACT

107161            TGAGACAGGGTTGCCCGAGGACATGATCCACACATGGTTGAAAAACAAAAGAGCTCGGCG
AH634 Section 2   TGAGACAGGGTTGCCCGAGGACATGATCCACACATGGTTGAAAAGCAAAAGAGCTCGG
AH635 Section 2   TGAGACAGGGTTGCCCGAGGACATGATCCACACATGGTTGAAAAGCAAAAGAGCTCG
AH631 Section 2   TGAGACAGGGTTGCCCGAGGACATGATCCACACATGGTTGAAAAACAAAAGAGCTC

107221            CCACAGGAGGGGCAGGCCCACAGCTCAAGATCAAGACTTGCTGGCCTCACAAGTGTCCGG
107281            TGGGGCCCCTGCAGGTCCGGTAGGCAGGGGCCATGAAGTTGCACAGGAGAGCTCATTGCC
107341            ACAGGAAGAAGCAGGAAGTACGGGCATGGATACTACAAGCACCAGCTACTCGCCCTCTTT
107401            CTGCAGAGAGTCCCAACTTTCCCAAGTGTCACAGCCCCGTGGAGCAGGCCAAAAAGAGGT
107461            CCCCACTCAAGCAGGCAACGTAGGCCCTCTGGAACTTCTCCTTGATGAACTGCAGGACGA
107521            AGTACAGGTGAAAGAGCATGTGCCAGACCCTTTGGATTTGGGTAGTGATCCTGGCGCCAG
107581            GGAGCCTGAAGGTTCCCAGGACAGCTTACAGAGCCTGGATGAAGCAGCAAATTCAGGCTG
107641            GCATACCTCGGTCCCGAGCATCTCGTCAACCTTGTGCAGAGAGTCCCAGCCTTCCCAAGT
107701            GGCACAGCCCTCTGGACCAGGACAAGCACAGGCCCCCACTCAAAGTGGGTTCATAGACCC
107761            TCTGGAGCTCTTTCTCGATGAACTGCTGACTGAAGTCCAACTTGAGGAGCAGGGGCCTGC
107821            CCCTGTGAATGTGGAGGAAACAGGGGAGCAAATGGACACAACACCTGAGCTGCCTCTCAC
AH634 Section 1                GGAGGAAACAGGGGAGCAAATGGACACAACACCTGAGCTGCCTCTCAC
AH631 Section 1                GGAGGAAACAGGGGAGCAAATGGACACAACACCTGAGCTGCCTCTCAC
AH635 Section 1                GGAGGAAACAGGGGAGCAAATGGACACAACACCTGAGCTGCCTCTCAC

107881            CCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGTACCCCTTGCAT
AH634 Section 1   CCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGTACCCCTTGCAT
AH631 Section 1   CCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGTACCCCTTGCAT
AH635 Section 1   CCCAGAAGAATATCAGACTCTTCTAGATATGCTCTGACTCCCTGACAGTACCCCTTGCAT

107941            CTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGAGAAGGTACCCA
AH634 Section 1   CTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGAGAAGGTACCCA
AH631 Section 1   CTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGAGAAGGTACCCA
AH635 Section 1   CTAGACACCCGCGAGGCCAAATCCTAAAGAGACCCAATTTGGAACTGGAGAAGGTACCCA


108001            TTTCAGCAAGGATGTGCACAAAAATGCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCT
AH634 Section 1   TTTCAGCAAGGATGTGCACAAAAATGCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCT
AH631 Section 1   TTTCAGCAAGGATGTGCACAAAAATGCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCT
AH635 Section 1   TTTCAGCAAGGATGTGCACAAAAATGCAACTCCAGTGACTTCCCGAAATGCAAGGTGTCT
```

```
     108061        TGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTGCAGCGCCC
AH634 Section 1    TGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTGCAGCGCCC
AH631 Section 1    TGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTGCAGCGCCC
AH635 Section 1    TGCTAACTATAAGGATGGATTGCAGGTGGGGATAATAATGAAGTGCCTTCTGCAGCGCCC

108121             AGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACAGATGGAGAGGC
AH634 Section 1    AGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACAGATGGAGAGGC
AH631 Section 1    AGGGATTATGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACAGATGGAGAGGC
AH635 Section 1    AGGGATTAGGAAATCAGCCCTGAAAGTGAGAGACTCTGCTACAGGGACAGATGGAGAGGC

108181             CAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCA
AH634 Section 1    CAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCA
AH631 Section 1    CAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCA
AH635 Section 1    CAATAGTGACTCCTCAACAACAAGGAGCCTAAAGATAACCCCAAAAGAAGGGCCACACCA

108241             AGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACTAGGCTTCACTA
AH634 Section 1    AGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACTAGGCTTCACTA
AH631 Section 1    AGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACTAGGCTTCACTA
AH635 Section 1    AGTGACTGGCTCCAGTGGACCCCAGGAAATCACACGGGACACTAGGACTAGGCTTCACTA

108301             CAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCCTGGGTTGTATG
AH634 Section 1    CAGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCCTGGGTTGTATG
AH631 Section 1    CCGAGGACACACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCCTGGGTTGTATG
AH635 Section 1    CAGAGGACGCACACTCCCTGAGGGCAATCGGGAGAGTGGACTCCTTCCCTGGGTTGTATG

108361             GACTGCTGCTATTCTTACAGATGCTTCGTGCAGAGCTGTGCAAGGTTTTACAGGCCAGTC
AH634 Section 1    GACTGCTGCTATTCTTACAG
AH631 Section 1    GACTGCTGCTATTCTTACAG
AH635 Section 1    GACTGCTGCTATTCTTACAG
```

Figure 5.13 Alignment of AH631, AH634 and AH635 with BAC clone 142C15

The sequences of the RT-PCR products were aligned to the BAC clone 142C15 using Sequencher. The predicted amino acid sequence of the ORF is shown underneath. Nucleotide changes from the consensus are highlighted in red.

5.2.5 Does the mouse array have the same DNA methylation pattern as the human repeats?

In humans, individuals unaffected with FSHD have high levels of methylation at the D4Z4 array (Tsien *et al.*, 2001). The array is hypomethylated on the disease allele of FSHD1 patients, and on both 4q and 10q alleles in FSHD2 patients. The methylation status of the mouse array was investigated to see whether it has the same methylation.

Liver and testis tissues were harvested from C57/BL6 mice by Jane Hewitt, and DNA was extracted as described in section 2.6.1. The DNA was then digested with MspI or HpaII restriction enzymes. These enzymes are isoschizomers which cleave the sequence CCGG and have different sensitivities to methylation. HpaII is sensitive to methylation of either C in the sequence. MspI is able to cut methylated CpG dinucleotides, however, it may be also be sensitive to CpNpG methylation.

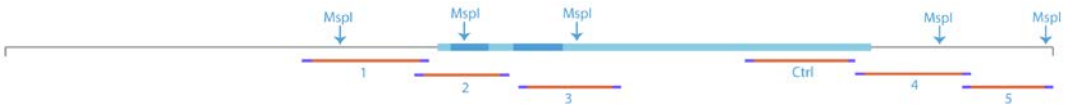
There are 5 MspI/HpaII restriction sites in the mouse Dux array and primers were designed to flank these sites (Figure 5.14a). If the site is unmethylated, both enzymes will be able to cleave the DNA and no product will be amplified from DNA digested with either enzyme. However, if there is methylation at the CpG dinucleotide, the HpaII enzyme's activity will be blocked and a product will be amplified from DNA digested with this enzyme. The MspI enzyme would be unaffected by this methylation and so no product would be amplified from DNA digested with MspI (Figure 5.14b). As a control for DNA quality, primers that amplify a product between the MspI/HpaII sites were also used, these reactions should be positive when using DNA digested with either enzyme.

In order to be sure that the DNA had been fully digested, primers were designed that flank a MspI/HpaII site in the pgk-1 gene. This gene is known to be methylated only on the inactive X chromosomes in mice (Singer-Sam *et al.*, 1990) and so a product should only be amplified in DNA from female mice. Because this site is methylated in females and unmethylated in males it could be used as both a positive and negative control. In addition, an MspI/HpaII site in the DHFR gene has been shown to be unmethylated at every stage in development (Kafri *et al.*, 1992), this acts as a good digestion control as no product should be amplified in any digested

samples. Finally, amplification of the Igfr2 gene acts as a positive control to show that the enzyme is unable to cut methylated DNA (primers M114 and M115, Table 2.4). This imprinted gene is known to be methylated only in the maternal oocytes of mice. The amplicon contains 3 MspI sites, these will be methylated on the maternal chromosome and so the product is expected to be produced only with the HpaII digests. (Figure 5.15a)

When the PCRs were performed on the sites within the Dux array, products were amplified from DNA digested with both enzymes, although the intensity of the bands in the MspI digested DNA were lower (Figure 5.15b). Although the controls suggested that the DNA was digested, the control genes are not in the same region as the Dux array, so it's possible that some of the repeats remain intact. In addition, it has been suggested that the MspI enzyme may be sensitive to methylation of the second C in the recognition sequence, if this is the case then methylation at this position would prevent cleavage of the site and a product would be seen.

a



b

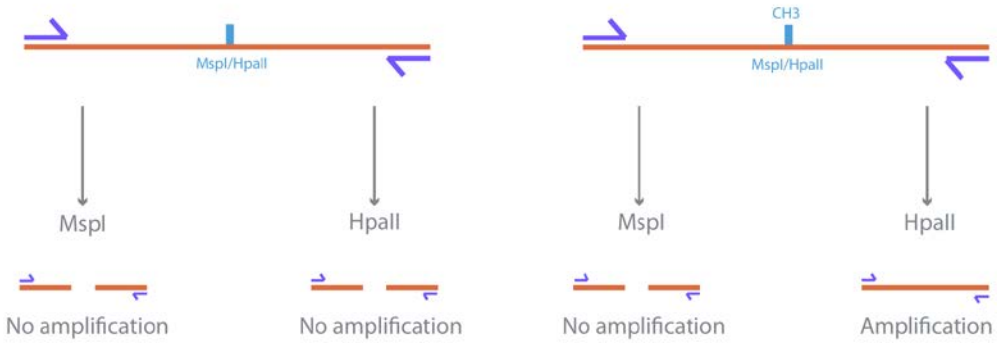


Figure 5.14 Location of MspI/HpaII sites in the mouse Dux repeat

a) Schematic representation of a single Dux repeat unit (5' – 3' orientation) with the 5 MspI/HpaII indicated. Primers flanking the sites are shown underneath in purple with their amplification product in orange. The control primer pair does not have an MspI/HpaII site within the product so should be amplified in DNA digested with either enzyme. b) Diagram to illustrate the expected results, dependent on methylation state and enzyme used for digestion. Purple arrows indicate primers. The restriction enzyme site is indicated. The CH3 highlights the methylated site.

a

	Msp Ctrl	Pgk1	DHFR	Igf2R
MspI	✓	X	X	X
HpaII	✓	✓(F)	X	✓

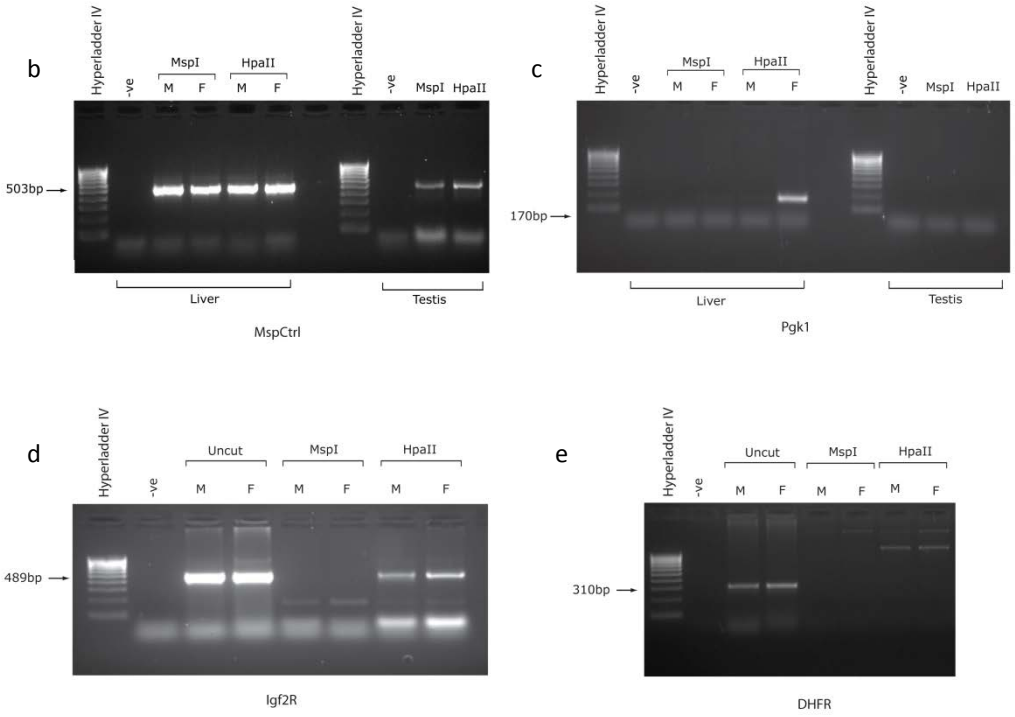


Figure 5.15 Confirmation of complete digestion of mouse DNA

a) Table to show expected results from control PCRs if digestion is complete b) PCR using mspCtrl primers which are within the array but do not contain a MspI/HpaII restriction site. c) PCR using Pgk1 primers. d) PCR using Igf2R primers. e) PCR using DHFR primers.

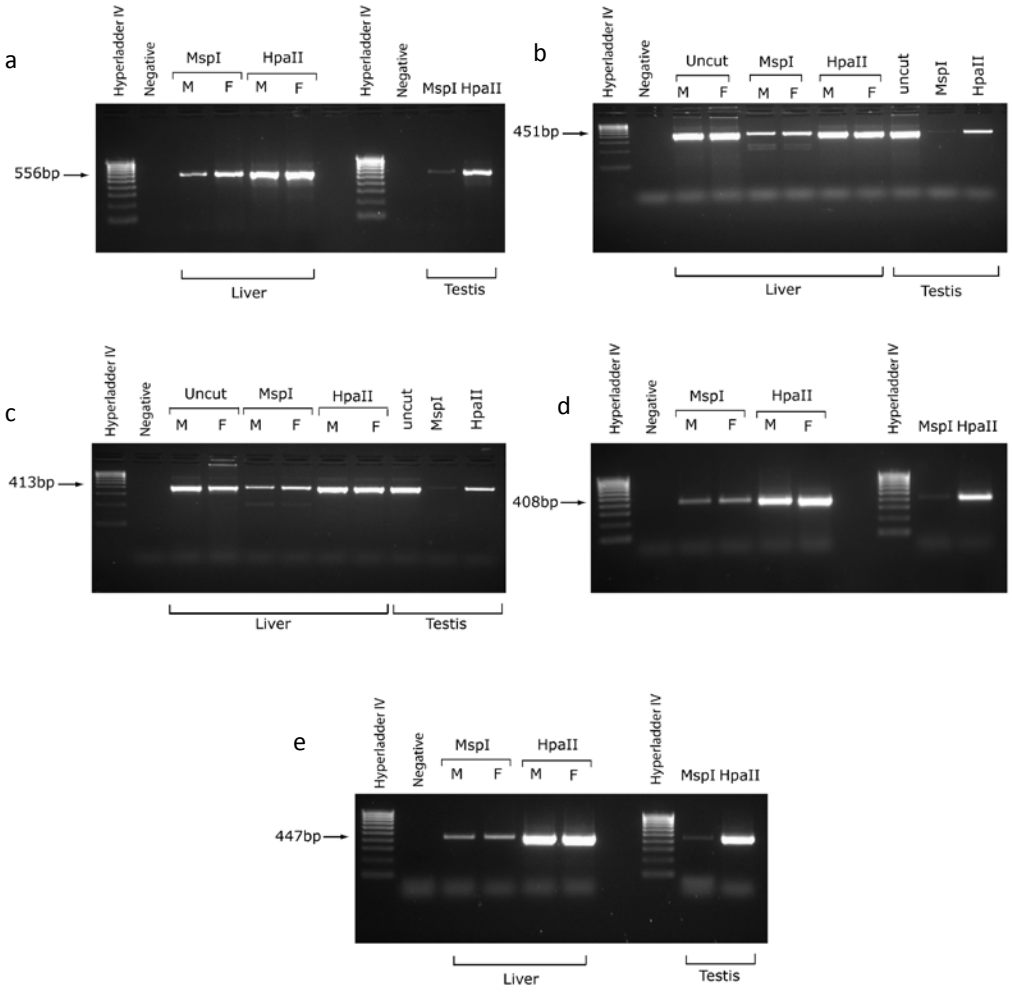


Figure 5.16 Methylation sensitive PCRs on MspI/HpaII digested DNA

Methylation sensitive PCRs on MspI or HpaII digested DNA. Primer sequences are shown in Table 2.4 a) Primer pair Msp1 b) Primer pair Msp2 c) Primer pair Msp3 d) Primer pair Msp4 e) Primer pair Msp5

5.3 Discussion

The main aim of the work in this chapter was to investigate expression from the mouse Dux array. Although no complete full length ORFs have been identified, smaller fragments from along the array have been amplified and sequenced and their variants compared with genomic sequence data from Jane Hewitt's laboratory.

Of the 107 RT-PCR products that were successfully cloned and sequenced, 46 aligned with the BAC clone 142C15. Two full length repeat units, and a further 4 partial repeats were identified in the sequence of the BAC. Only the first full repeat (repeat 3) appears to have a functional ORF. Thus, it was surprising to find that not all of the RT-PCR products that appear to originate from this BAC region aligned to repeat 3. In fact, only 17% of sequences aligned to this functional ORF. As discussed previously, the mouse strain from which the BAC clone originates is a C57/Bl6 mouse, while the RNA used for RT-PCR in this study is from Swiss-Webster mice. The outbred CD1 mouse was shown to have a more complex pattern of repeats (Clapp *et al.*, 2007), so it is possible that the outbred Swiss Webster mice also contain variations in repeat structure. If this is the case, the sequences which align to partial repeats without ORFs in this clone, may well come from repeats which remain functional in the Swiss Webster mice. Alternatively, these smaller fragments may represent regulatory RNAs such as those recently suggested for the human array (Snider *et al.*, 2009).

The majority of the remaining sequences, 50 out of 61, aligned with the genomic consensus sequence that had been compiled from genomic sequences amplified by others in Jane Hewitt's laboratory. Within these sequences, 14 variant positions were identified and 10 of these resulted in coding changes within the ORF. None of the variants occurred within the 86bp of homeodomain sequence in the mATG RT-PCR product of the conserved C-Terminal domain region in the mTGA product. 6 of the variant positions were also identified in the genomic sequences. It is surprising that so many of the variants resulted in coding changes, 71% of variants in these mouse RT-PCR products result in coding changes, while in the human RT-PCR products only 42% of variants result in a coding change (Table 4.1). None of the changes occurred in the conserved homeodomain and C-terminal domain sequences, so it's

possible that the sequence outside of these regions is less constrained. The levels of sequence variation between the sequences that match this consensus and the expressed sequences amplified from the hEC transcripts (table 4.2) are very similar, based on this sequence data a variant occurs, on average, every 63 base pairs in the mouse sequences and every 58 base pairs in the human sequences. However, the sequence of the repeats in the 142C15 region varies considerably compared to genomic sequences from the other Dux clusters, the repeats in this region may be less constrained, however, as a large proportion of transcripts appear to originate from here the region is likely to be transcribed. It is possible this 142C15 region has a different function to the other Dux clusters.

Finally, 11 sequences did not appear to originate from either the genomic sequences or the 142C15 BAC clone. It is likely that these sequences come from repeats which have not yet been fully sequenced.

No full length clones were able to be sequenced from the nested RT-PCR, however transcripts of around 1000bp were sequenced. Two of the 5 sequences aligned with the genomic consensus. One of these sequences, AH648, had 14 variant positions compared with the consensus, however a single genomic sequence has 12 of these changes so it is likely there are other variant repeats in mouse Dux region.

The other three sequences were found to be rearranged, as the two halves of the sequence aligned with the 142C15 sequence separately, with a 615bp gap in the middle. This is unlikely to be as a result of splicing of the transcript as a BLAST search of the full RT-PCR product aligns with the chromosome 10 contig with 99% identity. No rearranged sequences have yet been identified from the genomic mDux sequences.

The second aim of the work in this section was to investigate the methylation status of the mDux repeats. Five MspI/HpaII sites were identified in each repeat unit and amplification of PCR products surrounding these sites was investigated. The results of this analysis could not be linked with individual repeat units so this work would give a view of methylation state across the whole of the array.

There were initial problem with confirming complete digestion of the mouse DNA, however 4 control PCRs were included to overcome these. Once the control PCRs confirmed

complete digestion the MspI PCRs were performed on the DNA. For each of the 5 sites, amplification was seen from DNA digested with both MspI and HpaII enzymes, however there appeared to be less amplification from the MspI digested DNA.

As the control PCRs used to confirm complete digestion are located in different parts of the genome to the mouse Dux repeats, it's possible that although these regions have been fully digested the mouse array region has not. If the repeat units are compacted into heterchromatin, it's possible that the enzymes could not easily access their restriction sites. Although it has been confirmed that the HpaII enzyme is sensitive to methylation on either C of the recognition sequence, it has also been suggested that the MspI enzyme is sensitive to methylation of the external c of the sequence. If this external C is sometimes methylated, and the MspI enzyme is sensitive to this methylation, this could explain the amplification of products from the MspI digested DNA.

Chapter 6. General Discussion

This chapter gives an overview of advances in FSHD research that developed during the work described in this thesis. When the work described here began, the data showing expression from D4Z4 was new and most research was still focussed on a position effect mechanism on neighbouring genes. Over the last few years there have been a number of significant publications confirming expression from the D4Z4 array and providing an explanation for the haplotype specific pathogenicity.

In 2009, Snider *et al.* reported amplification of smaller fragments at the 3' and 5' ends of the ORF, in most cases their transcripts matched the final repeat of the array, although variant bases in some of the transcripts were suggestive of transcription from internal D4Z4 repeats. The group were also able to confirm the presence of the two splice forms reported by Dixit *et al*. (2007). In 2010, this group were able to report expression of two transcripts (DUX4-fl and DUX4-s) from the final repeat of the array, as they used a primer in the pLAM region their approach would not have amplified transcripts from internal repeats. The DUX4-fl transcript was amplified from 5/10 FSHD myoblast cell cultures, as well as testis and germ cell tumour RNA, while the DUX4-s transcript was present in all skeletal muscle controls and some of the FSHD samples. It is interesting that the full length transcripts were not identified in all of the FSHD samples. This would suggest that either the full length transcript is not required for the disease, or that it is expressed at a low abundance so that it is not always possible to detect. The requirement for a nested RT-PCR protocol in order to detect amplification is also consistent with a low abundance of *DUX4* mRNA.

This explanation for the difficulty in amplification of DUX4 mRNA was confirmed by Snider *et al.* (2010) reporting that only a small percentage of nuclei show *DUX4* expression. Using RT-PCR on limiting dilutions of myoblast cells along with immunostaining of cells with monoclonal antibodies to DUX4, the group were able to show that only 0.1% of FSHD muscle nuclei express DUX4. This observation could also help to explain why the RT-PCR results vary between reactions; the success of amplification will be dependent on the number of DUX4

expressing cells in the sample used, which is likely to vary. Snider and colleagues reported that muscle cells which show expression of the DUX4-fl transcript contain nuclear foci, which are characteristic of DUX4-induced apoptosis. It is possible that the culturing conditions of the myoblast cells affect the number of DUX4 positive cells in a sample; for example, apoptotic cells may no longer attach to the culture flasks and be lost on passaging or collection.

Expression from the testis and germ cell tumour RNA was detected at levels 100 fold higher than in FSHD muscle cells. The data described in this thesis support the finding as the hEC cell line used was isolated from a teratocarcinoma, a tumour derived from germ cells. In contrast to the myoblast data, identification of transcripts from this cell line did not require a nested RT-PCR and results were easily repeatable across reactions.

In 2010, Lemmers *et al.* reported that FSHD patients carry specific variants in the chromosome 4 region, distal to the last repeat of the array, which create a poly adenylation signal for transcripts derived from D4Z4. 4qB chromsomes lack this region, including the poly A site, while chromosome 10q carries the sequence ATCAAA which is not known to be a poly A signal. On FSHD permissive alleles, a polymorphism creates the signal ATTAAA, a commonly used poly A signal in humans. Lemmers and colleagues transfected C2C12 cells with constructs where the poly A signal of permissive chromosomes was replaced with a signal from a non-permissive chromosome and vice versa. They were able to show that constructs with the permissive poly A signal produced stable DUX4 constructs, while no transcript was detected from those with a non-permissive signal. In the mouse array, the most commonly used poly A signal, AATAAA, is present within the individual repeat unit, perhaps explaining why expression from the mouse array is more robust.

Lemmers *et al.* studied FSHD patients with unusual hybrid structures that contain both Bln^{+} and Bln^{-} repeat units. They identified a patient whose FSHD allele resides on chromosome 10 rather than chromosome 4. This array starts with two and a half Bln^{+} repeats and ends with one and a half Bln^{-} repeat units. The permissive distal end of the chromosome 4 had been transferred to chromosome 10 and importantly, none of the FSHD candidate genes that reside proximal to the D4Z4 array on chromosome 4 were transferred (Lemmers *et*

al., 2010a). This data provides strong evidence that the genes proximal to the array do not have a key role in FSHD pathogenesis.

Together, these data support a model for FSHD whereby expression from the D4Z4 array in differentiated tissues is repressed due to increased H3K9me3, possibly facilitated by the repetitive nature of the array. On contraction of this array, the chromatin in the region is relaxed and transcription of a full length transcript occurs from the most distal repeat; only if that allele contains a polymorphism creating a functional poly A signal is the transcript stable. Snider *et al.* (2010) also suggested that the DUX4-s transcript is produced as a mechanism to control transcripts that escape repression, but that this mechanism is disrupted on contraction of the array. Expression of the DUX4-fl transcript is thought to be toxic to cells, and might therefore induce the muscle damage seen in FSHD patients.

While expression of *DUX4* in differentiated muscle cells appears to be the cause of this disease, there is an increasing amount of data that supports a normal role of the DUX4 gene during development. Firstly, the DUX gene has been conserved for >100mya, strongly suggesting a protein coding function for the gene (Clapp *et al.*, 2007). Secondly, expression of *DUX4* transcripts has been identified in the germ-line lineage by Snider *et al.* (2010) and by the work described in this thesis. Importantly, this expression is not restricted to FSHD permissive alleles, Snider *et al.* reported expression from both 4q and 10q alleles from testis RNA, while the work described here amplified transcripts from a hEC cell line containing only 4qB163 and 10qA166 alleles.

Analysis of the RT-PCR products amplified during this work has provided information on the variation seen between repeats of the D4Z4 array. Neither the sequence of, or expression from the internal repeat units has previously been examined. The transcripts contained a number of variant sites indicating that expression occurs from multiple units of the repeat array.

In order to investigate the normal function of the DUX4 protein, binding sites and partners for the protein will need to be investigated which could be achieved by chromatin immunoprecipitation assays. Although it has been reported that the DUX4 protein binds to

the promoter region of *PITX1* causing its upregulation, work described in this thesis was unable to confirm that overexpression of DUX4 induces *PITX1* expression.

An animal model for FSHD would be useful to investigate the function of the *DUX* genes. To date, DUX4 has been overexpressed in both *Xenopus* and Zebrafish (Snider *et al.*, 2009; Wuebbles *et al.*, 2010; Wallace *et al.*, 2011) embryos, resulting in developmental malformations and heart defects. However, as these species' genomes do not contain any homologous *DUX* genes they are unlikely to provide a useful model for FSHD (Leidenroth and Hewitt, 2010). *DUX4* has also been overexpressed in adult mouse by injection into muscle (Wallace *et al.*, 2011), which produced muscle degeneration consistent with FSHD. However, there is currently no mouse model stably expressing DUX4 at the levels seen in FSHD patients. If the mouse *Dux* array was shown to be functionally equivalent to the D4Z4 array in humans, a mouse model of FSHD could potentially be developed by deletion of part of the mouse *Dux* array. In addition, if the *Dux* and *DUX4* genes do carry out equivalent functions, further studies of the mouse gene may be helpful in elucidating the normal role of *DUX4* during development.

Work described in this thesis has contributed to the understanding of the mouse *Dux* array. Firstly, analysis of the methylation status of the repeats indicates that the mouse array is methylated, as seen in humans unaffected with FSHD. Secondly, amplification and sequencing of transcripts from the mouse array has given an insight into sequence variation between repeats, and the genomic origins of mouse *Dux* transcripts.

Analysis of transcripts amplified from mouse RNA has shown that a large proportion of the RT-PCR products originated from a small cluster of Dux repeats, some of which do not appear to contain functional ORFs. Transcripts originating from these partial repeats may be non-functional and it is currently unclear what role these may have. There have been reports of small mRNA transcripts in humans and it has been suggested that these may have a regulatory role, controlling transcription from the array. Alternatively, these fragments may originate from full repeats that have not yet been identified in the mouse genome. During this work, three longer transcripts were identified which showed a rearrangement compared with genomic transcripts, supporting the suggestion that there are variant genomic repeats within

the mouse genome which have not yet been identified in genomic sequencing studies. It is possible that these variant repeats are not present in the inbred mouse strain used for the genomic sequencing, if there is significant variation at this region between inbred and outbred strains then comparison of the organisation of and expression from the arrays in multiple strains may provide helpful information about the function of the arrays in the mice.

Chapter 7. References

Agresti, A., Rainaldi, G., Lobbiani, A., Magnani, I., Dilernia, R., Meneveri, R., Siccardi, A.G., and Ginelli, E. (1987). Chromosomal location by in situ hybridization of the human SAU3A family of DNA repeats. *Human Genetics* **75**, 326-332.

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

Almstrup, K., Hoei-Hansen, C.E., Wirkner, U., Blake, J., Schwager, C., Ansorge, W., Nielsen, J.E., Skakkebaek, N.E., Meyts, E.R.D., and Leffers, H. (2004). Embryonic stem cell-like features of testicular carcinoma in situ revealed by genome-wide gene expression profiling. *Cancer Research* **64**, 4736-4743.

Altherr, M.R., Bengtsson, U., Markovich, R.P., and Winokur, S.T. (1995). Efforts toward understanding the molecular-basis of facioscapulohumeral muscular-dystrophy. *Muscle & Nerve* **Suppl 2**, S32-S38.

Alvarado, D.M., McCall, K., Aferol, H., Silva, M.J., Garbow, J.R., Spees, W.M., Patel, T., Siegel, M., Dobbs, M.B., and Gurnett, C.A. (2011). Pitx1 haploinsufficiency causes clubfoot in humans and a clubfoot-like phenotype in mice. *Human Molecular Genetics* **20**, 3943-3952.

Andrews, P.W., Matin, M.M., Bahrami, A.R., Damjanov, I., Gokhale, P., and Draper, J.S. (2005). Embryonic stem (es) cells and embryonal carcinoma (ec) cells: Opposite sides of the same coin. *Biochemical Society Transactions* **33**, 1526-1530.

Ansseau, E., Laoudj-Chenivesse, D., Marcowycz, A., Tassin, A., Vanderplanck, C., Sauvage, S., Barro, M., Mahieu, I., Leroy, A., Leclercq, I., Mainfroid, V., Figlewicz, D., Mouly, V., Butler-Browne, G., Belayew, A., and Coppee, F. (2009). Dux4c is up-regulated in FSHD. It induces the myf5 protein and human myoblast proliferation. *PLoS ONE* **4**.

Bakker, E., Wijmenga, C., Vossen, R., Padberg, G.W., Hewitt, J., Vanderwielen, M., Rasmussen, K., and Frants, R.R. (1995). The FSHD-linked locus D4F104S1 (p13E-11) on 4q35 has a homolog on 10qter. *Muscle & Nerve* **Suppl 2**, S39-S44.

Barro, M., Carnac, G., Flavier, S., Mercier, J., Vassetzky, Y., and Laoudj-Chenivesse, D. (2010). Myoblasts from affected and non affected FSHD muscles exhibit morphological differentiation defects. *Journal of Cellular and Molecular Medicine* **14**, 275-289.

Bastress, K.L., Stajich, J.M., Speer, M.C., and Gilbert, J.R. (2005). The genes encoding for D4Z4 binding proteins HMGB2, YY1, NCL, and MYOD1 are excluded as candidate genes for FSHD1B. *Neuromuscular Disorders* **15**, 316-320.

Beckers, M.-C., Gabriels, J., van der Maarel, S., De Vriese, A., Frants, R.R., Collen, D., and Belayew, A. (2001). Active genes in junk DNA? Characterization of DUX genes embedded within 3.3 kb repeated elements. *Gene* **264**, 51-57.

Bodega, B., Ramirez, G.D.C., Grasser, F., Cheli, S., Brunelli, S., Mora, M., Meneveri, R., Marozzi, A., Mueller, S., Battaglioli, E., and Ginelli, E. (2009). Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation. *BMC Biology* **7**, 41.

Booth, H.A.F., and Holland, P.W.H. (2007). Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene* **387**, 7-14.

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

Boulikas, T. (1993). Nature of DNA-sequences at the attachment regions of genes to the nuclear matrix. *Journal of Cellular Biochemistry* **52**, 14-22.

Boulikas, T. (1995). Chromatin domains and prediction of MAR sequences. *International Review of Cytology* **162A**, 279-388.

Brouwer, O.F., Padberg, G.W., Bakker, E., Wijmenga, C., and Frants, R.R. (1995). Early-onset facioscapulohumeral muscular-dystrophy. *Muscle & Nerve* **Suppl 2**, S67-S72.

Buzhov, B.T., Lemmers, R., Tournev, I., Dikova, C., Kremensky, I., Petrova, J., Frants, R.R., and van der Maarel, S.M. (2005). Genetic confirmation of facioscapulohumeral muscular dystrophy in a case with complex D4Z4 rearrangments. *Human Genetics* **116**, 262-266.

Cacurri, S., Piazzo, N., Deidda, G., Vigneti, E., Galluzzi, G., Colantoni, L., Merico, B., Ricci, E., and Felicetti, L. (1998). Sequence homology between 4qter and 10qter loci facilitates the instability of subtelomeric KpnI repeat units implicated in facioscapulohumeral muscular dystrophy. *American Journal of Human Genetics* **63**, 181-190.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.-F., and Fraser, P. (2002). Long-range chromatin regulatory interactions *in vivo*. *Nature Genetics* **32**, 623-626.

Celegato, B., Capitanio, D., Pescatori, M., Romualdi, C., Pacchioni, B., Cagnin, S., Vigano, A., Colantoni, L., Begum, S., Ricci, E., Wait, R., Lanfranchi, G., and Gelfi, C. (2006). Parallel protein and transcript profiles of FSHD patient muscles correlate to the D4Z4 arrangement and reveal a common impairment of slow to fast fibre differentiation and a general deregulation of MyoD-dependent genes. *Proteomics* **6**, 5303-5321.

Chen, Y.-W., Zhao, P., Borup, R., and Hoffman, E.P. (2000). Expression profiling in the muscular dystrophies: Identification of novel aspects of molecular pathophysiology. *Journal of Cell Biology* **151**, 1321-1336.

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

Clark, L.N., Koehler, U., Ward, D.C., Wienberg, J., and Hewitt, J.E. (1996). Analysis of the organisation and localisation of the FSHD-associated tandem array in primates: Implications for the origin and evolution of the 3.3 kb repeat family. *Chromosoma* **105**, 180-189.

Cockell, M., and Gasser, S.M. (1999). Nuclear compartments and gene regulation. *Current Opinion in Genetics & Development* **9**, 199-205.

Coulon, V., L'Honore, A., Ouimette, J.F., Dumontier, E., van den Munckhof, P., and Drouin, J. (2007). A muscle-specific promoter directs Pitx3 gene expression in skeletal muscle cells. *Journal of Biological Chemistry* **282**, 33192-33200.

Dˆrner, A., and Schultheiss, H.-P. (2007). Adenine nucleotide translocase in the focus of cardiovascular diseases. *Trends in Cardiovascular Medicine* **17**, 284-290.

Davie, J.R. (1995). The nuclear matrix and the regulation of chromatin organization and function. *International Review of Cytology, Vol 162a* **162A**, 191-250.

de Greef, J.C., Lemmers, R., Camano, P., Day, J.W., Sacconi, S., Dunand, M., van Engelen, B.G.M., Kiuru-Enari, S., Padberg, G.W., Rosa, A.L., Desnuelle, C., Spuler, S., Tarnopolsky, M., Venance, S.L., Frants, R.R., van der Maarel, S.M., and Tawil, R. (2010). Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology* **75**, 1548-1554.

de Greef, J.C., Lemmers, R.J.L.F., van Engelen, B.G.M., Sacconi, S., Venance, S.L., Frants, R.R., Tawil, R., and van der Maarel, S.M. (2009). Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Human Mutation* **30**, 1449-1459.

Deak, K.L., Lemmers, R., Stajich, J.M., Klooster, R., Tawil, R., Frants, R.R., Speer, M.C., van der Maarel, S.M., and Gilbert, J.R. (2007). Genotype-phenotype study in an FSHD family with a proximal deletion encompassing p13E-11 and D4Z4. *Neurology* **68**, 578-582.

Deidda, G., Cacurri, S., Piazzo, N., and Felicetti, L. (1996). Direct detection of 4q35 rearrangements implicated in facioscapulohumeral muscular dystrophy (FSHD). *Journal of Medical Genetics* **33**, 361-365.

DeLaurier, A., Schweitzer, R., and Logan, M. (2006). Pitx1 determines the morphology of muscle, tendon, and bones of the hindlimb. *Developmental Biology* **299**, 22-34.

Dickson, M.C. (1998). Characterisation of the subtelomeric region of human chromosome 4q, PhD Thesis, (University of Manchester)

Ding, H., Beckers, M.C., Plaisance, S., Marynen, P., Collen, D., and Belayew, A. (1998). Characterization of a double homeodomain protein (DUX1) encoded by a cDNA homologous to 3.3 kb dispersed repeated elements. *Human Molecular Genetics* **7**, 1681-1694.

Dixit, M., Ansseau, E., Tassin, A., Winokur, S., Shi, R., Qian, H., Sauvage, S., Mattotti, C., van Acker, A.M., Leo, O., Figiewicz, D., Barro, M., Laoudj-Chenivesse, D., Belayew, A., Coppee, F., and Chen, Y.W. (2007). DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. *Proceedings of the National Academy of Sciences, USA* **104**, 18157-18162.

Doerner, A., Pauschinger, M., Badorff, A., Noutsias, M., Giessen, S., Schulze, K., Bilger, J., Rauch, U., and Schultheiss, H.P. (1997). Tissue-specific transcription pattern of the adenine nucleotide translocase isoforms in humans. *Febs Letters* **414**, 258-262.

Fischbeck, K.H., and Garbern, J.Y. (1992). Facioscapulohumeral muscular dystrophy defect identified. *Nature Genetics* **2**, 3-4.

Forlani, G., Giarda, E., Ala, U., Di Cunto, F., Salani, M., Tupler, R., Kilstrup-Nielsen, C., and Landsberger, N. (2010). The MeCP2/YY1 interaction regulates ANT1 expression at 4q35: Novel hints for Rett syndrome pathogenesis. *Human Molecular Genetics* **19**, 3114-3123.

Funakoshi, M., Goto, K., and Arahata, K. (1998). Epilepsy and mental retardation in a subset of early onset 4q35-facioscapulohumeral muscular dystrophy. *Neurology* **50**, 1791-1794.

Gabellini, D., D'Antona, G., Moggio, M., Prelle, A., Zecca, C., Adami, R., Angeletti, B., Ciscato, P., Pellegrino, M.A., Bottinelli, R., Green, M.R., and Tupler, R. (2006). Facioscapulohumeral muscular dystrophy in mice overexpressing FRG1. *Nature* **439**, 973-977.

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

Gabriels, J., Beckers, M.C., Ding, H., De Vriese, A., Plaisance, S., van der Maarel, S.M., Padberg, G.W., Frants, R.R., Hewitt, J.E., Collen, D., and Belayew, A. (1999). Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. *Gene* **236**, 25-32.

Garrick, D., Fiering, S., Martin, D.I.K., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. *Nature Genetics* **18**, 56-59.

GibsonBrown, J.J., Agulnik, S.I., Chapman, D.L., Alexiou, M., Garvey, N., Silver, L.M., and Papaioannou, V.E. (1996). Evidence of a role for T-box genes in the evolution of limb morphogenesis and the specification of forelimb/hindlimb identity. *Mechanisms of Development* **56**, 93-101.

Gilbert, J.R., Stajich, J.M., Speer, M.C., Vance, J.M., Stewart, C.S., Yamaoka, L.H., Samson, F., Fardeau, M., Potter, T.G., Roses, A.D., and et al. (1992). Linkage studies in facioscapulohumeral muscular dystrophy (FSHD). *American Journal of Human Genetics* **51**, 424-427.

Gilbert, J.R., Stajich, J.M., Wall, S., Carter, S.C., Qiu, H., Vance, J.M., Stewart, C.S., Speer, M.C., Pufky, J., Yamaoka, L.H., Rozear, M., Samson, F., Fardeau, M., Roses, A.D., and Pericakvance, M.A. (1993). Evidence for heterogeneity in facioscapulohumeral muscular-dystrophy (FSHD). *American Journal of Human Genetics* **53**, 401-408.

Gottschling, D.E., Aparicio, O.M., Billington, B.L., and Zakian, V.A. (1990). Position effect at *saccharomyces cerevisiae* telomeres - reversible repression of pol-II transcription. *Cell* **63**, 751-762.

Grewal, P.K., van Geel, M., Frants, R.R., de Jong, P., and Hewitt, J.E. (1999). Recent amplification of the human FRG1 gene during primate evolution. *Gene* **227**, 79-88.

Griggs, R.C., Tawil, R., Storvick, D., Mendell, J.R., and Altherr, M.R. (1993). Genetics of facioscapulohumeral muscular-dystrophy - new mutations in sporadic cases. *Neurology* **43**, 2369-2372.

Gurnett, C.A., Alaee, F., Kruse, L.M., Desruisseau, D.M., Hecht, J.T., Wise, C.A., Bowcock, A.M., and Dobbs, M.B. (2008). Asymmetric lower-limb malformations in individuals with homeobox PITX1 gene mutation. *American Journal of Human Genetics* **83**, 616-622.

Hanakahi, L.A., Dempsey, L.A., Li, M.J., and Maizels, N. (1997). Nucleolin is one component of the B cell-specific transcription factor and switch region binding protein, lr1. *Proceedings of the National Academy of Sciences, USA* **94**, 3605-3610.

Hanel, M.L., Wuebbles, R.D., and Jones, P.L. (2009). Muscular dystrophy candidate gene FRG1 is critical for muscle development. *Developmental Dynamics* **238**, 1502-1512.

Haslbeck, K.M., Bierhaus, A., Erwin, S., Kirchner, A., Nawroth, P., Schlotzer, U., Neundorfer, B., and Heuss, D. (2004). Receptor for advanced glycation endproduct (RAGE)-mediated nuclear factor-kappa B activation in vasculitic neuropathy. *Muscle & Nerve* **29**, 853-860.

Hewitt, J.E., Lyle, R., Clark, L.N., Valleley, E.M., Wright, T.J., Wijmenga, C., van Deutekom, J.C.T., Francis, F., Sharpe, P.T., Hofker, M., Frants, R.R., and Williamson, R. (1994). Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystropothhy. *Human Molecular Genetics* **3**, 1287-1295.

Hirano, M., and DiMauro, S. (2001). ANT1, Twinkle, POLG, and TP: New genes open our eyes to ophthalmoplegia. *Neurology* **57**, 2163-2165.

Iqbal, Z., Roper, H., Pericak-Vance, M.A., Hung, W.Y., Delong, R., Cummings, W.J.K., and Siddique, T. (1992). Genetic heterogeneity in facioscapulohumeral disease. *American Journal of Human Genetics* **51**, A191.

Jardine, P., Jones, M., Tyfield, L., Upadhyaya, M., and Lunt, P. (1993). De-novo DNA rearrangement in atypical facioscapulohumeral muscular dystrophy. *Clinical Genetics* **44**, 167-167.

Jiang, G.C., Yang, F., van Overveld, P.G.M., Vedanarayanan, V., van der Maarel, S., and Ehrlich, M. (2003). Testing the position-effect variegation hypothesis for facioscapulohumeral muscular dystrophy by analysis of histone modification and gene expression in subtelomeric 4q. *Human Molecular Genetics* **12**, 2909-2921.

Jordens, E.Z., Palmieri, L., Huizing, M., Van Den Heuvel, L.P., Sengers, R.C.A., Dˆrner Wim Ruitenbeek, A., Trijbels, F.J., Valsson, J., Sigfusson, G., Palmieri, F., and Smeitink, J.A.M. (2002). Adenine nucleotide translocator 1 deficiency associated with Sengers syndrome. *Annals of Neurology* **52**, 95-99.

Jurica, M.S., Licklider, L.J., Gygi, S.P., Grigorieff, N., and Moore, M.J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA* **8**, 426-439.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. *Genes & Development* **6**, 705-714.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K.C., Hallinan, J., Mattick, J., Hume, D.A., Lipovich, L., Batalov, S., Engstrom, P.G., Mizuno, Y., Faghihi, M.A., Sandelin, A., Chalk, A.M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., Wahlestedt, C., Core, R.G.N.P., Genome Sci, G., and Consortium, F. (2005). Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566.

Kawamura-Saito, M., Yamazaki, Y., Kaneko, K., Kawaguchi, N., Kanda, H., Mukai, H., Gotoh, T., Motoi, T., Fukayama, M., Aburatani, H., Takizawa, T., and Nakamura, T. (2006). Fusion between CIC and DUX4 up-regulates PEA3 family genes in ewing-like sarcomas with t(4;19)(q35;q13) translocation. *Human Molecular Genetics* **15**, 2125-2137.

Kleinjan, D.J., and van Heyningen, V. (1998). Position effect in human genetic disease. *Human Molecular Genetics* **7**, 1611-1618.

Klinge, L., Eagle, M., Haggerty, I.D., Roberts, C.E., Straub, V., and Bushby, K.M. (2006). Severe phenotype in infantile facioscapulohumeral muscular dystrophy. *Neuromuscular Disorders* **16**, 553-558.

Kohler, J., Rupilius, B., Otto, M., Bathke, K., and Koch, M.C. (1996). Germline mosaicism in 4q35 facioscapulohumeral muscular dystrophy (FSHD1A) occurring predominantly in oogenesis. *Human Genetics* **98**, 485-490.

Kondo, T., Bobek, M.P., Kuick, R., Lamb, B., Zhu, X.X., Narayan, A., Bourc'his, D., Viegas-Pequignot, E., Ehrlich, M., and Hanash, S.M. (2000). Whole-genome methylation scan in ICF syndrome: Hypomethylation of non-satellite DNA repeats D4Z4 and NBL2. *Human Molecular Genetics* **9**, 597-604.

Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Le Beau, M.M., Fisher, A.G., and Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science* **296**, 158-162.

Kowaljow, V., Marcowycz, A., Ansseau, E., Conde, C.B., Sauvage, S., Mattotti, C., Arias, C., Corona, E.D., Nufiez, N.G., Leo, O., Wattiez, R., Iglewicz, D.F., Laoudj-Chenivesse, D., BelayeW, A., Coppe, F., and Rosa, A.L. (2007). The DUX4 gene at the FSHDIA locus encodes a pro-apoptotic protein. *Neuromuscular Disorders* **17**, 611-623.

Lanctot, C., Lamolet, B., and Drouin, J. (1997). The bicoid-related homeoprotein ptx1 defines the most anterior domain of the embryo and differentiates posterior from anterior lateral mesoderm. *Development* **124**, 2807-2817.

Lanctot, C., Moreau, A., Chamberland, M., Tremblay, M.L., and Drouin, J. (1999). Hindlimb patterning and mandible development require the *ptx1* gene. *Development* **126**, 1805-1810.

Laoudj-Chenivesse, D., Carnac, G., Bisbal, C., Hugon, G., Bouillot, S., Desnuelle, C., Vassetzky, Y., and Fernandez, A. (2005). Increased levels of adenine nucleotide translocator 1 protein and response to oxidative stress are early events in facioscapulohumeral muscular dystrophy muscle. *Journal of Molecular Medicine* **83**, 216-224.

Lapidot, M., and Pilpel, Y. (2006). Genome-wide natural antisense transcription: Coupling its regulation to its different regulatory mechanisms. *Embo Reports* **7**, 1216-1222.

Leidenroth, A., and Hewitt, J.E. (2010). A family history of DUX4: Phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evolutionary Biology* **10**.

Lemmers, R., de Kievit, P., Sandkuijl, L., Padberg, G.W., van Ommen, G.J.B., Frants, R.R., and van der Maarel, S.M. (2002). Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nature Genetics* **32**, 235-236.

Lemmers, R., Osborn, M., Haaf, T., Rogers, M., Frants, R.R., Padberg, G.W., Cooper, D.N., van der Maarel, S.M., and Upadhyaya, M. (2003). D4F104S1 deletion in facioscapulohumeral muscular dystrophy - phenotype, size, and detection. *Neurology* **61**, 178-183.

Lemmers, R., van der Maarel, S.M., van Deutekom, J.C.T., van der Wielen, M.J.R., Deidda, G., Dauwerse, H.G., Hewitt, J., Hofker, M., Bakker, E., Padberg, G.W., and Frants, R.R. (1998). Inter- and intrachromosomal sub-telomeric rearrangements on 4q35: Implications for facioscapulohumeral muscular dystrophy (FSHD) aetiology and diagnosis. *Human Molecular Genetics* **7**, 1207-1214.

Lemmers, R., van der Vliet, P.J., Klooster, R., Sacconi, S., Camano, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W., Miller, D.G., Tapscott, S.J., Tawil, R., Frants, R.R., and van der Maarel, S.M. (2010a). A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650-1653.

Lemmers, R., van der Vliet, P.J., van der Gaag, K.J., Zuniga, S., Frants, R.R., de Knijff, P., and van der Maarel, S.M. (2010b). Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete interchromosomal sequence transfers in human evolution. *American Journal of Human Genetics* **86**, 364-377.

Lemmers, R., van Overveld, P.G.M., Sandkuijl, L.A., Vrieling, H., Padberg, G.W., Frants, R.R., and van der Maarel, S.M. (2004a). Mechanism and timing of mitotic rearrangements in the subtelomeric D4Z4 repeat involved in facioscapulohumeral muscular dystrophy. *American Journal of Human Genetics* **75**, 44-53.

Lemmers, R., Wohlgemuth, M., Frants, R.R., Padberg, G.W., Morava, E., and van der Maarel, S.M. (2004b). Contractions of D4Z4 on 4qB subtelomeres do not cause facioscapulohumeral muscular dystrophy. *American Journal of Human Genetics* **75**, 1124-1130.

Lemmers, R., Wohlgemuth, M., van der Gaag, K.J., van der Vliet, P.J., van Teijlingen, C.M.M., de Knijff, P., Padberg, G.W., Frants, R.R., and van der Maarel, S.M. (2007). Specific sequence

variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *American Journal of Human Genetics* **81**, 884-894.

Lemmers, R.J.L.F., Peggy De Kievit, Michel van Geel, Michiel J. R. van der Wielen, Egbert Bakker, George W. Padberg, Frants, R.R., and Maarel, S.M.V.D. (2001). Complete allele information in the diagnosis of facioscapulohumeral muscular dystrophy by triple DNA analysis. *Annals of Neurology* **50**, 816-819.

Levis, R., Hazelrigg, T., and Rubin, G.M. (1985). Effects of genomic position on the expression of transduced copies of the *white* gene of *Drosophila*. *Science* **229**, 558-561.

Lewis, E.B. (1950). The phenomenon of position effect. *Advances in Genetics Incorporating Molecular Genetic Medicine* **3**, 73-115.

Li, K., Warner, C.K., Hodge, J.A., Minoshima, S., Kudoh, J., Fukuyama, R., Maekawa, M., Shimizu, Y., Shimizu, N., and Wallace, D.C. (1989). A human-muscle adenine-nucleotide translocator gene has 4 exons, is located on chromosome-4, and is differentially expressed. *Journal of Biological Chemistry* **264**, 13998-14004.

Logan, M., and Tabin, C.J. (1999). Role of Pitx1 upstream of Tbx4 in specification of hindlimb identity. *Science* **283**, 1736-1739.

Lunt, P.W., Compston, D.A.S., and Harper, P.S. (1989). Estimation of age-dependent penetrance in facioscapulohumeral muscular-dystrophy by minimizing ascertainment bias. *Journal of Medical Genetics* **26**, 755-760.

Lunt, P.W., and Harper, P.S. (1991). Genetic-counseling in facioscapulohumeral muscular-dystrophy. *Journal of Medical Genetics* **28**, 655-664.

Lunt, P.W., Jardine, P.E., Koch, M., Maynard, J., Osborn, M., Williams, M., Harper, P.S., and Upadhyaya, M. (1995a). Phenotypic genotypic correlation will assist genetic-counseling in 4q35-facioscapulohumeral muscular-dystrophy. *Muscle & Nerve* **Suppl 2**, S103-S109.

Lunt, P.W., Jardine, P.E., Koch, M.C., Maynard, J., Osborn, M., Williams, M., Harper, P.S., and Upadhyaya, M. (1995b). Correlation between fragment size at D4F104S1 and age at onset or at wheelchair use, with a possible generational-effect, accounts for much phenotypic variation in 4q35-facioscapulohumeral muscular dystrophy (FSHD). *Human Molecular Genetics* **4**, 951-958.

Lyle, R., Wright, T.J., Clark, L.N., and Hewitt, J.E. (1995). FSHD-associated repeat, D4Z4, is a member of a dispersed family of homeobox-containing repeats, subsets of which are clustered on the short arms of the acrocentric chromosomes. *Genomics* **28**, 389-397.

Macaione, V., Aguennouz, M., Rodolico, C., Mazzeo, A., Patti, A., Cannistraci, E., Colantone, L., Di Giorgio, R.M., De Luca, G., and Vita, G. (2007). RAGE-NF-kappa B pathway activation in response to oxidative stress in facioscapulohumeral muscular dystrophy. *Acta Neurologica Scandinavica* **115**, 115-121.

Manz, E., Schnieders, F., Brechlin, A.M., and Schmidtke, J. (1993). TSPY-related sequences represent A microheterogeneous gene family organized as constitutive elements in DYZ5 tandem repeat units on the human y-chromosome. *Genomics* **17**, 726-731.

Marshall, W.F., Fung, J.C., and Sedat, J.W. (1997). Deconstructing the nucleus: Global architecture from local interactions. *Current Opinion in Genetics & Development* **7**, 259-263.

Masny, P.S., Bengtsson, U., Chung, S.A., Martin, J.H., van Engelen, B., van der Maarel, S.M., and Winokur, S.T. (2004). Localization of 4q35.2 to the nuclear periphery: Is FSHD a nuclear envelope disease? *Human Molecular Genetics* **13**, 1857-1871.

Masny, P.S., Chan, O.Y.A., de Greef, J.C., Bengtsson, U., Ehrlich, M., Tawil, R., Lock, L.F., Hewitt, J.E., Stocksdale, J., Martin, J.H., van der Maarel, S.M., and Winokur, S.T. (2010). Analysis of allele-specific RNA transcription in FSHD by RNA-DNA fish in single myonuclei. *European Journal of Human Genetics* **18**, 448-456.

Mathews, K.D., Mills, K.A., Bosch, E.P., Ionasescu, V.V., Wiles, K.R., Buetow, K.H., and Murray, J.C. (1992). Linkage localization of facioscapulohumeral muscular dystrophy (FSHD) in 4q35. *American Journal of Human Genetics* **51**, 428-431.

Meneveri, R., Agresti, A., Dellavalle, G., Talarico, D., Siccardi, A.G., and Ginelli, E. (1985). Identification of A human clustered G+C-rich DNA family of repeats (SAU3A family). *Journal of Molecular Biology* **186**, 483-489.

Meneveri, R., Agresti, A., Marozzi, A., Saccone, S., Rocchi, M., Archidiacono, N., Corneo, G., Dellavalle, G., and Ginelli, E. (1993). Molecular-organization and chromosomal location of human gc-rich heterochromatic blocks. *Gene* **123**, 227-234.

Mills, K.A., Buetow, K.H., Xu, Y., Ritty, T.M., Mathews, K.D., Bodrug, S.E., Wijmenga, C., Balazs, I., and Murray, J.C. (1992). Genetic and physical mapping on chromosome-4 narrows the localization of the gene for facioscapulohumeral muscular-dystrophy (FSHD). *American Journal of Human Genetics* **51**, 432-439.

Miura, K., Kumagai, T., Matsumoto, A., Iriyama, E., Watanabe, K., Goto, K., and Arahata, K. (1998). Two cases of chromosome 4q35-linked early onset facioscapulohumeral muscular dystrophy with mental retardation and epilepsy. *Neuropediatrics* **29**, 239-241.

Monici, M.C., Aguennouz, M., Mazzeo, A., Messina, C., and Vita, G. (2003). Activation of nuclear factor-kappa B in inflammatory myopathies and duchenne muscular dystrophy. *Neurology* **60**, 993-997.

Morosetti, R., Mirabella, M., Gliubizzi, C., Broccolini, A., Sancricca, C., Pescatori, M., Gidaro, T., Tasca, G., Frusciante, R., Tonali, P.A., Cossu, G., and Ricci, E. (2007). Isolation and characterization of mesoangioblasts from facioscapulohumeral muscular dystrophy muscle biopsies. *Stem Cells* **25**, 3173-3182.

Moylan, J.S., and Reid, M.B. (2007). Oxidative stress, chronic disease, and muscle wasting. *Muscle & Nerve* **35**, 411-429.

Nassif, N., Penney, J., Pal, S., Engels, W.R., and Gloor, G.B. (1994). Efficiant copying of nonhomologous sequences from ectopic sites via p-element-induced gap repair. *Molecular and Cellular Biology* **14**, 1613-1625.

Nguyen, C.T., Gonzales, F.A., and Jones, P.A. (2001). Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: Correlation of accessibility, methylation, MeCP2 binding and acetylation. *Nucleic Acids Research* **29**, 4598-4606.

Osborne, R.J., Welle, S., Venance, S.L., Thornton, C.A., and Tawil, R. (2007). Expression profile of FSHD supports a link between retinal vasculopathy and muscular dystrophy. *Neurology* **68**, 569-577.

Ostlund, C., Garcia-Carrasquillo, R.M., Belayew, A., and Worman, H.J. (2005). Intracellular trafficking and dynamics of double homeodomain proteins. *Biochemistry* **44**, 2378-2384.

Ottaviani, A., Rival-Gervier, S., Boussouar, A., Foerster, A.M., Rondier, D., Sacconi, S., Desnuelle, C., Gilson, E., and Magdinier, F. (2009). The D4Z4 macrosatellite repeat acts as a CTCF and A-type lamins-dependent insulator in facio-scapulo-humeral dystrophy. *Plos Genetics* **5**.

Ottaviani, A., Rival-Gervier, S., Forster, A., Gilson, E., and Magdinier, F. (2006). The D4Z4 subtelomeric element behaves as a CTCF-dependent insulator and anchors telomeres to the nuclear periphery. *Neuromuscular Disorders* **16**, 713-713.

Padberg, G.W. (2004a). Facioscapulohumeral muscular dystrophy: A clinician's experience. In Facioscapulohumeral muscular dystrophy clinical medicine and molecular cell biology, M. Upadhyaya, and D.N. Cooper, eds. (Garland Science/BIOS Scientific Publishers), pp. 41-54.

Padberg, G.W. (2004b). Facioscapulohumeral muscular dystrophy: A clinician's experience. In Facioscapulohumeral muscular dystrophy, D.N.C. M.Upadhyaya, ed. (Garland Science/BIOS Scientific Publishers), pp. 41-54.

Padberg, G.W., Brouwer, O.F., Dekeizer, R.J.W., Dijkman, G., Wijmenga, C., Grote, J.J., and Frants, R.R. (1995). On the significance of retinal vascular-disease and hearing-loss in facioscapulohumeral muscular-dystrophy. *Muscle & Nerve* **Suppl 2**, S73-S80.

Petrov, A., Allinne, J., Pirozhkova, I., Laoudj, D., Lipinski, M., and Vassetzky, Y.S. (2008). A nuclear matrix attachment site in the 4q35 locus has an enhancer-blocking activity in vivo: Implications for the facio-scapulo-humeral dystrophy. *Genome Research* **18**, 39-45.

Petrov, A., Pirozhkova, I., Carnac, G., Laoudj, D., Lipinski, M., and Vassetzky, Y.S. (2006). Chromatin loop domain organization within the 4q35 locus in facioscapulohumeral dystrophy patients versus normal human myoblasts. *Proceedings of the National Academy of Sciences, USA* **103**, 6982-6987.

Pirozhkova, I., Petrov, A., Dmitriev, P., Laoudj, D., Lipinski, M., and Vassetzky, Y. (2008). A functional role for 4qA/B in the structural rearrangement of the 4q35 region and in the regulation of FRG1 and ANT1 in facioscapulohumeral dystrophy. *PLoS ONE* **3**, e3389.

Rappsilber, J., Ryder, U., Lamond, A., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Research* **12**, 1231-1245.

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

Rossi, M., Ricci, E., Colantoni, L., Galluzzi, G., Frusciante, R., Tonali, P.A., and Felicetti, L. (2007). The facioscapulohumeral muscular dystrophy region on 4qter and the homologous locus on 10qter evolved independently under different evolutionary pressure. *BMC Medical Genetics* **8**.

Saito, A., Higuchi, I., Nakagawa, M., Saito, M., Uchida, Y., Inose, M., Kasai, T., Niiyama, T., Fukunaga, H., Arimura, K., and Osame, M. (2000). An overexpression of fibroblast growth factor (FGF) and FGF receptor 4 in a severe clinical phenotype of facioscapulohumeral muscular dystrophy. *Muscle & Nerve* **23**, 490-497.

Saitoh, Y., Miyamoto, N., Okada, T., Gondo, Y., Showguchi-Miyata, J., Hadano, S., and Ikeda, J.E. (2000). The RS447 human megasatellite tandem repetitive sequence encodes a novel deubiquitinating enzyme with a functional promoter. *Genomics* **67**, 291-300.

Sarfarazi, M., Upadhyaya, M., Padberg, G., Pericak-Vance, M., Siddique, T., Lucotte, G., and Lunt, P. (1989). An exclusion map for facioscapulohumeral (Landouzy-Dejerine) disease. *Journal of Medical Genetics* **26**, 481-484.

Sarfarazi, M., Wijmenga, C., Upadhyaya, M., Weiffenbach, B., Hyser, C., Mathews, K., Murray, J., Gilbert, J., Pericakvance, M., Lunt, P., Frants, R.R., Jacobsen, S., Harper, P.S., and Padberg,

G.W. (1992). Regional mapping of facioscapulohumeral muscular-dystrophy gene on 4q35 - combined analysis of an international consortium. *American Journal of Human Genetics* **51**, 396-403.

Schwartz, C.M., Spivak, C.E., Baker, S.C., McDaniel, T.K., Loring, J.F., Nguyen, C., Chrest, F.J., Wersto, R., Arenas, E., Zeng, X.M., Freed, W.J., and Rao, M.S. (2005). Ntera2: A model system to study dopaminergic differentiation of human embryonic stem cells. *Stem Cells and Development* **14**, 517-534.

Scofield, D.G., Hong, X., and Lynch, M. (2007). Position of the final intron in full-length transcripts: Determined by NMD? *Molecular Biology and Evolution* **24**, 896-899.

Shapiro, M.D., Bell, M.A., and Kingsley, D.M. (2006). Parallel genetic origins of pelvic reduction in vertebrates. *Proceedings of the National Academy of Sciences, USA* **103**, 13753-13758.

Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jonsson, B., Schluter, D., and Kingsley, D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717-723.

Singer-Sam, J., Grant, M., LeBon, J.M., Okuyama, K., Chapman, V., Monk, M., and Riggs, A.D. (1990). Use of a HpaII-polymerase chain reaction assay to study DNA methylation in the Pgk-1 CpG island of mouse embryos at the time of X-chromosome inactivation. *Mol Cell Biol* **10**, 4987-4989.

Snider, L., Asawachaicharn, A., Tyler, A.E., Geng, L.N., Petek, L.M., Maves, L., Miller, D.G., Lemmers, R., Winokur, S.T., Tawil, R., van der Maarel, S.M., Filippova, G.N., and Tapscott, S.J. (2009). RNA transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: New candidates for the pathophysiology of facioscapulohumeral dystrophy. *Human Molecular Genetics* **18**, 2414-2430.

Snider, L., Geng, L.N., Lemmers, R., Kyba, M., Ware, C.B., Nelson, A.M., Tawil, R., Filippova, G.N., van der Maarel, S.M., Tapscott, S.J., and Miller, D.G. (2010). Facioscapulohumeral dystrophy: Incomplete suppression of a retrotransposed gene. *Plos Genetics* **6**.

Spaenkuch, B., and Strebhardt, K. (2005). RNA interference-based gene silencing in mice: The development of a novel therapeutical strategy. *Current Pharmaceutical Design* **11**, 3405-3419.

Stepien, G., Torroni, A., Chung, A.B., Hodge, J.A., and Wallace, D.C. (1992). Differential expression of adenine-nucleotide translocator isoforms in mammalian-tissues and during muscle-cell differentiation. *Journal of Biological Chemistry* **267**, 14592-14597.

Stratton, M.R., Reeves, B.R., and Cooper, C.S. (1989). Misidentified cell. *Nature* **337**, 311-312.

Sugimoto, A. (2004). High-throughput RNAi in *Caenorhabditis elegans*: Genome-wide screens and functional genomics. *Differentiation* **72**, 81-91.

Tam, R., Smith, K.P., and Lawrence, J.B. (2004). The 4q subtelomere harboring the FSHD locus is specifically anchored with peripheral heterochromatin unlike most human telomeres. *Journal of Cell Biology* **167**, 269-279.

Tawil, R., Forrester, J., Griggs, R.C., Mendell, J., Kissel, J., McDermott, M., King, W., Weiffenbach, B., Figlewicz, D., Cos, L., Langsam, A., Pandya, S., Martens, B., Brower, C., Herr, B., Downing, K., and Gorell, W.C. (1996). Evidence for anticipation and association of deletion size with severity in facioscapulohumeral muscular dystrophy. *Annals of Neurology* **39**, 744-748.

Thomas, J.O., and Travers, A.A. (2001). HMG1 and 2, and related 'architectural' DNA-binding proteins. *Trends in Biochemical Sciences* **26**, 167-174.

Thomas, M.J., and Seto, E. (1999). Unlocking the mechanisms of transcription factor YY1: Are chromatin modifying enzymes the key? *Gene* **236**, 197-208.

Thomas, N.S.T., Wiseman, K., Spurlock, G., MacDonald, M., Ustek, D., and Upadhyaya, M. (2007). A large patient study confirming that facioscapulohumeral muscular dystrophy (FSHD) disease expression is almost exclusively associated with an FSHD locus located on a 4qA-defined 4qter subtelomere. *Journal of Medical Genetics* **44**, 215-218.

Tonini, M.M.O., Lemmers, R., Pavanello, R.C.M., Cerqueira, A.M.P., Frants, R.R., van der Maarel, S.M., and Zatz, M. (2006). Equal proportions of affected cells in muscle and blood of a mosaic carrier of facioscapulohumeral muscular dystrophy. *Human Genetics* **119**, 23-28.

Tsien, F., Sun, B.D., Hopkins, N.E., Vedanarayanan, V., Figlewicz, D., Winokur, S., and Ehrlich, M. (2001). Methylation of the FSHD syndrome-linked subtelomeric repeat in normal and FSHD cell cultures and tissues. *Molecular Genetics and Metabolism* **74**, 322-331.

Tsumagari, K., Chang, S.-C., Lacey, M., Baribault, C., Chittur, S.V., Sowden, J., Tawil, R., Crawford, G.E., and Ehrlich, M. (2011). Gene expression during normal and FSHD myogenesis. *BMC Medical Genomics* **4**.

Tupler, R., Berardinelli, A., Barbierato, L., Frants, R., Hewitt, J.E., Lanzi, G., Maraschio, P., and Tiepolo, L. (1996). Monosomy of distal 4q does not cause facioscapulohumeral muscular dystrophy. *Journal of Medical Genetics* **33**, 366-370.

Tupler, R., Perini, G., Pellegrino, M.A., and Green, M.R. (1999). Profound misregulation of muscle-specific gene expression in facioscapulohumeral muscular dystrophy. *Proceedings of the National Academy of Sciences, USA* **96**, 12650-12654.

Upadhyaya, M., Jardine, P., Maynard, J., Farnham, J., Sarfarazi, M., Wijmenga, C., Hewitt, J.E., Frants, R., Harper, P.S., and Lunt, P.W. (1993). Molecular analysis of British facioscapulohumeral dystrophy families for 4q DNA rearrangements. *Human Molecular Genetics* **2**, 981-987.

Upadhyaya, M., Lunt, P., Sarfarazi, M., Broadhead, W., Farnham, J., and Harper, P.S. (1992). The mapping of chromosome 4q markers in relation to facioscapulohumeral muscular dystrophy (FSHD). *American Journal of Human Genetics* **51**, 404-410.

van der Maarel, S.M., Deidda, G., Lemmers, R., van Overveld, P.G.M., van der Wielen, M., Hewitt, J.E., Sandkuijl, L., Bakker, B., van Ommen, G.J.B., Padberg, G.W., and Frants, R.R. (2000). *De novo* facioscapulohumeral muscular dystrophy: Frequent somatic mosaicism, sex-dependent phenotype, and the role of mitotic transchromosomal repeat interaction between chromosomes 4 and 10. *American Journal of Human Genetics* **66**, 26-35.

van Deutekom, J.C.T., Bakker, E., Lemmers, R., vanderWielen, M.J.R., Bik, E., Hofker, M.H., Padberg, G.W., and Frants, R.R. (1996a). Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: Implications for genetic counselling and etiology of FSHD1. *Human Molecular Genetics* **5**, 1997-2003.

van Deutekom, J.C.T., Lemmers, R., Grewal, P.K., vanGeel, M., Romberg, S., Dauwerse, H.G., Wright, T.J., Padberg, G.W., Hofker, M.H., Hewitt, J.E., and Frants, R.R. (1996b). Identification of the first gene (FRG1) from the FSHD region on human chromosome 4q35. *Human Molecular Genetics* **5**, 581-590.

Van Deutekom, J.C.T., Wijmenga, C., Vantienhoven, E.A.E., Gruter, A.M., Frants, R.R., Hewitt, J.E., Padberg, G.W., Vanommen, G.J.B., and Hofker, M.H. (1993a). FSHD associated DNA rearrangements are due to deletions of integral copies of A 3.2 kb tandemly repeated unit. *Human Molecular Genetics* **2**, 2037-2042.

van Deutekom, J.C.T., Wljmenga, C., Tlenhoven, E.A.E.V., Gruter, A.-M., Hewitt, J.E., Padberg, G.W., Ommen, G.-J.B.v., Hofker, M.H., and Fronts, R.R. (1993b). FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Human Molecular Genetics* **2**, 2037-2042.

van Driel, R., Wansink, D.G., vanSteensel, B., Grande, M.A., Schul, W., and deJong, L. (1995). Nuclear domains and the nuclear matrix. *International Review of Cytology* **162A**, 151-189.

van Geel, M., Dickson, M.C., Beck, A.F., Bolland, D.J., Frants, R.R., van der Maarel, S.M., de Jong, P.J., and Hewitt, J.E. (2002). Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. *Genomics* **79**, 210-217.

van Geel, M., Heather, L.J., Lyle, R., Hewitt, J.E., Frants, R.R., and de Jong, P.J. (1999). The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. *Genomics* **61**, 55-65.

van Koningsbruggen, S., Straasheijm, K.R., Sterrenburg, E., de Graaf, N., Dauwerse, H.G., Frants, R.R., and van der Maarel, S.M. (2007). FRG1P-mediated aggregation of proteins involved in pre-mRNA processing. *Chromosoma* **116**, 53-64.

van Overveld, P.G.M., Enthoven, L., Ricci, E., Rossi, M., Felicetti, L., Jeanpierre, M., Winokur, S.T., Frants, R.R., Padberg, G.W., and van der Maarel, S.M. (2005). Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. *Annals of Neurology* **58**, 569-576.

van Overveld, P.G.M., Lemmers, R., Deidda, G., Sandkuijl, L., Padberg, G.W., Frants, R.R., and van der Maarel, S.M. (2000). Interchromosomal repeat array interactions between chromosomes 4 and 10: A model for subtelomeric plasticity. *Human Molecular Genetics* **9**, 2879-2884.

van Overveld, P.G.M., Lemmers, R., Sandkuijl, L.A., Enthoven, L., Winokur, S.T., Bakels, F., Padberg, G.W., van Ommen, G.J.B., Frants, R.R., and van der Maarel, S.M. (2003). Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nature Genetics* **35**, 315-317.

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

Weiffenbach, B., Bagley, R., Falls, K., Hyser, C., Storvick, D., Jacobsen, S.J., Schultz, P., Mendell, J., Willems van Dijk, K., Milner, E.C., and et al. (1992). Linkage analyses of five chromosome 4 markers localizes the facioscapulohumeral muscular dystrophy (FSHD) gene to distal 4q35. *American Journal of Human Genetics* **51**, 416-423.

Weiffenbach, B., Dubois, J., Storvick, D., Tawil, R., Jacobsen, S.J., Gilbert, J., Wijmenga, C., Mendell, J.R., Winokur, S., Altherr, M.R., Schultz, P., Olandt, S., Frants, R.R., Pericakvance, M., and Griggs, R.C. (1993). Mapping the facioscapulohumeral muscular dystrophy gene is complicated by chromosome-4q35 recombination events. *Nature Genetics* **4**, 165-169.

Wijmenga, C., Brouwer, O.F., Padberg, G.W., and Frants, R.R. (1992a). Transmission of de-novo mutation associated with facioscapulohumeral muscular-dystrophy. *Lancet* **340**, 985-986.

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

Wijmenga, C., Hewitt, J.E., Sandkuijl, L.A., Clark, L.N., Wright, T.J., Dauwerse, H.G., Gruter, A.-M., Hofker, M.H., Moerer, P., Williamson, R., van Ommen, G.-J.B., Padberg, G.W., and Frants, R.R. (1992b). Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nature Genetics* **2**, 26-30.

Wijmenga, C., Padberg, G.W., Moerer, P., Wiegant, J., Liem, L., Brouwer, O.F., Milner, E.C.B., Weber, J.L., Vanommen, G.B., Sandkuyl, L.A., and Frants, R.R. (1991). Mapping of facioscapulohumeral muscular dystrophy gene to chromosome 4q35-qter by multipoint linkage analysis and *in situ* hybridization. *Genomics* **9**, 570-575.

Wijmenga, C., Sandkuijl, L.A., Moerer, P., Vanderboorn, N., Bodrug, S.E., Ray, P.N., Brouwer, O.F., Murray, J.C., Vanommen, G.J.B., Padberg, G.W., and Frants, R.R. (1992c). Genetic-linkage map of facioscapulohumeral muscular-dystrophy and 5 polymorphic loci on chromosome 4q35-qter. *American Journal of Human Genetics* **51**, 411-415.

Wijmenga, C., Winokur, S.T., Padberg, G.W., Skraastad, M.I., Altherr, M.R., Wasmuth, J.J., Murray, J.C., Hofker, M.H., and Frants, R.R. (1993). The human skeletal muscle adenine nucleotide translocator gene maps to chromosome 4q35 in the region of the facioscapulohumeral muscular dystrophy locus. *Human Genetics* **92**, 198-203.

Winokur, S.T., Barrett, K., Martin, J.H., Forrester, J.R., Simon, M., Tawil, R., Chung, S.A., Masny, P.S., and Figlewicz, D.A. (2003a). Facioscapulohumeral muscular dystrophy (FSHD) myoblasts demonstrate increased susceptibility to oxidative stress. *Neuromuscular Disorders* **13**, 322-333.

Winokur, S.T., Bengtsson, U., Feddersen, J., Mathews, K.D., Weiffenbach, B., Bailey, H., Markovich, R.P., Murray, J.C., Wasmuth, J.J., and Altherr, M.R. (1994). The DNA rearrangement associated with facioscapulohumeral muscular dystrophy involves a heterochromatin-associated repetitive element: Implications for a role of chromatin structure in the pathogenesis of the disease. *Chromosome Research* **2**, 225-234.

Winokur, S.T., Bengtsson, U., Vargas, J.C., Wasmuth, J.J., and Altherr, M.R. (1996). The evolutionary distribution and structural organization of the homeobox-containing repeat D4Z4 indicates a functional role for the ancestral copy in the FSHD region. *Human Molecular Genetics* **5**, 1567-1575.

Winokur, S.T., Chen, Y.W., Masny, P.S., Martin, J.H., Ehmsen, J.T., Tapscott, S.J., van der Maarel, S.M., Hayashi, Y., and Flanigan, K.M. (2003b). Expression profiling of FSHD muscle supports a defect in specific stages of myogenic differentiation. *Human Molecular Genetics* **12**, 2895-2907.

Winokur, S.T., Schutte, B., Weiffenbach, B., Washington, S.S., McElligott, D., Chakravarti, A., Wasmuth, J.H., and Altherr, M.R. (1993). A radiation hybrid map of 15-loci on the distal long arm of chromosome-4, the region containing the gene responsible for facioscapulohumeral muscular-dystrophy (FSHD). *American Journal of Human Genetics* **53**, 874-880.

Wuebbles, R.D., Hanel, M.L., and Jones, P.L. (2009). FSHD region gene 1 (FRG1) is crucial for angiogenesis linking FRG1 to facioscapulohumeral muscular dystrophy-associated vasculopathy. *Disease Models & Mechanisms* **2**, 267-274.

Wuebbles, R.D., Long, S.W., Hanel, M.L., and Jones, P.L. (2010). Testing the effects of FSHD candidate gene expression in vertebrate muscle development. *Int J Clin Exp Pathol* **3**, 386-400.

Xing, Y., Johnson, C.V., Moen, P.T., Jr., McNeil, J.A., and Lawrence, J. (1995). Nonrandom gene organization: Structural arrangements of specific pre-mRNA transcription and splicing with sc-35 domains. *J Cell Biol* **131**, 1635-1647.

Xu, Z., Wei, W., Gagneur, J., Clauder-Muenster, S., Smolik, M., Huber, W., and Steinmetz, L.M. (2011). Antisense expression increases gene expression variability and locus interdependency. *Molecular Systems Biology* **7**.

Yamaoka, L., Speer, M.C., Stajich, J., Lewis, K., Clancy, R., Qiu, H., Kumar, A., Vance, J., Stewart, C., Rozear, M., Roses, A.D., Pericalvance, M.A., and Gilbert, J.R. (1995). Exclusion mapping of chromosomal regions which cross-hybridize to FSH1A associated markers in FSH1B. *American Journal of Human Genetics* **57**, 1905-1905.

Yip, D.J., and Picketts, D.J. (2003). Increasing D4Z4 repeat copy number compromises C2C12 myoblast differentiation. *Febs Letters* **537**, 133-138.

Zatz, M., Marie, S.K., Passosbueno, M.R., Vainzof, M., Campiotto, S., Cerqueira, A., Wijmenga, C., Padberg, G., and Frants, R. (1995). High proportion of new mutations and possible anticipation in Brazilian facioscapulohumeral muscular-dystrophy families. *American Journal of Human Genetics* **56**, 99-105.

Zeng, W.H., de Greef, J.C., Chen, Y.Y., Chien, R., Kong, X.D., Gregson, H.C., Winokur, S.T., Pyle, A., Robertson, K.D., Schmiesing, J.A., Kimonis, V.E., Balog, J., Frants, R.R., Ball, A.R., Lock, L.F., Donovan, P.J., van der Maarel, S.M., and Yokomori, K. (2009). Specific loss of histone H3 lysine 9 trimethylation and HP1 gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). *Plos Genetics* **5**.

Zhang, X.Y., Loflin, P.T., Gehrke, C.W., Andrews, P.A., and Ehrlich, M. (1987). Hypermethylation of human DNA-sequences in embryonal carcinoma-cells and somatic tissues but not in sperm. *Nucleic Acids Research* **15**, 9429-9449.

Zhou, D.M., He, Q.C.S., Wang, C.Y., Zhang, J., and Wong-Staal, F. (2006). RNA interference and potential applications. *Current Topics in Medicinal Chemistry* **6**, 901-911.