Ana Sofia Tavares Figueiredo

Bachelor's degree in Biology

Relevance of Epigenetics in the Pathogenic Mechanism of Spinocerebellar Ataxia Type 37

Dissertation to obtain master's degree in

Molecular Genetics and Biomedicine

Supervisor:	Doctor Isabel Silveira, Principal Researcher at i3S (University of Porto)
Co-Supervisor:	Doctor Ana Rita Grosso, Assistant Researcher at FCT-NOVA, Group Leader at UCIBIO

Jury:

President:	Doctor José Paulo Sampaio, Assistant Professor at FCT-NOVA
Arguing:	Doctor José Bessa, Principal Researcher at i3S (University of Porto)

November, 2020

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Relevance of Epigenetics in the Pathogenic Mechanism of Spinocerebellar Ataxia Type 37
Ana Figueiredo

Agradecimentos

Esta dissertação foi desenvolvida no grupo Genetics of Cognitive Dysfunction no Instituto de Investigação e Inovação em Saúde da Universidade do Porto (i3S-U.Porto), integrada no Mestrado em Genética Molecular e Biomedicina da Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa (FCT-UNL). Gostaria de agradecer a ambas as entidades pela excelente oportunidade que me foi proporcionada.

No decorrer deste ano letivo, várias pessoas contribuíram para este trabalho, tanto a nível científico, como a nível pessoal com amizade e apoio moral. A essas pessoas, devo um enorme agradecimento.

Em primeiro lugar, gostaria de agradecer à minha orientadora, Doutora Isabel Silveira, por me ter recebido no seu laboratório com vontade em integrar-me na sua equipa. Agradeço por todo o apoio dado, pelo conhecimento transmitido e por todas as críticas construtivas que me fizeram crescer profissionalmente. Muito obrigada.

Agradeço também à Doutora Ana Rita Grosso, minha co-orientadora, pelo input científico dado nesta dissertação e pelos esclarecimentos prestados ao longo do ano. Acima de tudo, agradeço pela sua enorme simpatia e pela pronta disponibilidade em ajudar sempre que necessário. Muito obrigada.

À Joana Loureiro, a pessoa que mais me ajudou no laboratório. Devo-lhe um enorme agradecimento por todo o tempo e paciência que dispensou a responder às minhas dúvidas teóricas, práticas e algumas dúvidas existenciais. Em todas as minhas otimizações falhadas, tinha sempre um conselho assertivo a dar e uma palavra de consolo “Sofia, já passei por isso, vai funcionar”. Sem a ajuda da Joana e sem o trabalho previamente desenvolvido por ela, este trabalho não teria sido possível. Obrigada por todo o conhecimento transmitido e, principalmente, pelo companheirismo.

À Filipa, um enorme agradecimento pelo tempo despendido comigo, por todas as dúvidas esclarecidas, por toda a ajuda nos trabalhos práticos e pela demonstração de novas técnicas. Agradeço também pelo seu trabalho previamente desenvolvido, que ajudou ao desenvolvimento da minha dissertação. Muito obrigada por tudo.

Aos amigos do Serrim, que embora longe com um em cada canto do país e do mundo, estão sempre presentes (nas mil e trezentas video-chamadas por dia). Saudades vossas e saudades de Aveiro.

Aos amigos Rústicos de Odf, pelas conversas de fazer rir qualquer apático, pelos cafés ao fim de semana quando estavam mais do que três pessoas na terrinha, por estarem presentes em todos os momentos importantes e menos importantes e, principalmente, pela amizade.

À Cláudia, que me acompanha há mais de 10 anos. Obrigada pela paciência gigante em aturar o meu mau humor matinal, de fim da tarde e da noite (sempre, portanto). Obrigada pelos melhores conselhos, pelos cozinhados e pelas boleias em dias de chuva torrencial. Não preciso de dizer muito, porque tu já sabes tudo.

Rafael, obrigada por existires e por teres entrado na minha vida. Obrigada pela compreensão e pelo apoio ao longo destes anos, por estares sempre comigo nos melhores e nos piores momentos. Tens um parágrafo, mas merecias uma página inteira + 1 caixa de sushi. Um xi-coração para ti.

Por último, mas não menos importante: aos meus pais (paistrocinadores), por me terem trazido ao mundo e pelo apoio incondicional que sempre me deram. Obrigada por acreditarem em mim e por investirem na minha educação. Se hoje estou a fazer o que mais gosto, é graças a vocês – devo-vos tudo. A toda a restante família, um bem-haja.

Abstract

Spinocerebellar ataxia 37 (SCA37) is an autosomal-dominant neurodegenerative disease characterized by cerebellar atrophy, gait and limb incoordination, and dysarthria as the first symptom. SCA37 is caused by an $(ATTTC)_n$ insertion within a nonpathogenic $(ATTTT)_n$ located in a 5' UTR intron of *DAB1*. The age of onset in patients correlates with the number of ATTTCs and there is an increase in repeat insertion size during transmission to the next generation, with larger increases when the father is the transmitting parent. Haplotype analysis suggested that genetic factors flanking the mutant allele act as cis-elements influencing repeat instability. To identify cis-elements of genetic instability involved in the origin of mutant SCA37 chromosomes, we investigated single nucleotide polymorphisms (SNPs) and methylation status in the SCA37 repeat flanking region. SNPs were assessed by Sanger sequencing and DNA methylation by bisulfite sequencing of chromosomes containing nonpathogenic $(ATTTT)_{<50}$, $(ATTTT)_{>50}$ and interrupted, and mutant SCA37 alleles.

We found a total of nine SNPs and confirmed that the repeat flanking region is highly polymorphic. SNP 7 and SNP 9 were present in all nonpathogenic large, interrupted and mutant SCA37 alleles studied, whereas only a small number of short nonpathogenic chromosomes carried these SNPs; they were associated with an increase in the $(ATTTT)_n$ tract, suggesting they might be a factor for repeat instability. These two SNPs are both in CpG dinucleotides, causing their elimination. The location of these SNPs shows an above average occupancy score for CTCF-binding in several cancer and embryonic human cells, which reinforces their involvement in repeat instability.

In conclusion, this work allowed the identification of genetic variants that could have modified the epigenetics of the *DAB1* $(ATTTT)_n$ flanking region and led to repeat instability. The occurrence of variants in important cis-regulatory elements might have created the ideal conditions for the mutational mechanism, originating the SCA37 $(ATTTC)_n$ insertion.

Keywords: Neurodegenerative diseases, repeat instability, Single Nucleotide Polymorphisms (SNPs), DNA methylation.

Resumo

A ataxia espinocerebelosa 37 (SCA37) é uma doença neurodegenerativa autossómica-dominante caracterizada por atrofia cerebelar, perda de coordenação dos membros e da marcha e disartria como primeiro sintoma. SCA37 é causada por uma inserção $(ATTTC)_n$ numa sequência $(ATTTT)_n$ normal, localizada na 5'UTR da região intrónica do *DAB1*. A idade de aparecimento da doença correlaciona-se com o número de ATTTCs, existindo um aumento do tamanho da inserção durante a transmissão à geração seguinte, sendo maior quando o pai é o transmissor. A análise dos haplótipos sugere que fatores genéticos presentes no alelo mutado atuam como elementos-cis, influenciando a instabilidade da repetição. Para identificar elementos-cis de instabilidade envolvidos na origem dos cromossomas mutados, foi investigado a presença de polimorfismos de nucleótido único (SNPs) e o padrão de metilação da região flanqueante à repetição. Os SNPs nos alelos não-patogénicos $(ATTTT)_{<50}$, $(ATTTT)_{>50}$ e interrompidos e mutados foram analisados por sequenciação de Sanger e a metilação do DNA por sequenciação-bissulfito.

Foram encontrados nove SNPs, confirmando-se que a região flanqueante da repetição é altamente polimórfica. Os SNPs 7 e 9 estão presentes em todos os alelos grandes, interrompidos e mutados, mas apenas num pequeno número de alelos não-patogénicos pequenos. Estes estão associados a um aumento no tamanho do $(ATTTT)_n$, sugerindo que podem ser um fator de instabilidade da repetição. SNP7 e SNP9 estão ambos em CpGs causando a sua eliminação, e a sua localização apresenta um nível de ocupação acima da média para a ligação do CTCF em células cancerígenas e embrionárias humanas, reforçando o seu envolvimento na instabilidade.

Concluindo, este trabalho permitiu a identificação de variantes genéticas que podem ter modificado o padrão epigenético da região da repetição no *DAB1*, levando à sua instabilidade. A ocorrência de variantes em elementos-cis regulatórios pode ter criado condições favoráveis aos mecanismos mutacionais que originaram a inserção $(ATTTC)_n$ responsável pela SCA37.

Palavras-chave: Doenças neurodegenerativas, instabilidade alélica, polimorfismos de alelo único (SNPs), metilação do DNA

List of contents

List of Figures

List of Tables

List of Abbreviations

4C	Chromosome conformation capture-on-chip
5-hmC	5-hydroxymethylcytosine
5-mC	5-methylcytosine
ASM	Allele-specific methylation
ATXN7	*Ataxin 7*
ATXN8	*Ataxin 8*
ATXN8OS	*Ataxin 8 oposite strand*
BEAN	Brain expressed associated with *Nedd4*
bp	Base pairs
BPES	Blepharophimosis syndrome
BS-seq	Bisulfite sequencing
BSS	Baratela-Scott syndrome
C9orf72	*Chromosome 9 open reading frame 72*
CGI	CpG Island
CI	Confidence interval
CpG	Cytosine-phosphate-guanine
CTCF	CCTC-binding factor
CTCF-BS	CTCF binding site
ddNTP	Dideoxynucleotide
DM1	Myotonic dystrophy 1
DM2	Myotonic dystrophy 2
DNA	Deoxyribonucleic acid
DNMT	*De novo* methyltransferases
dsDNA	Double-strand deoxyribonucleic acid
ESME®	Epigenetic Sequencing Methylation Analysis Software
FAME	Familial adult myoclonic epilepsy
FMR1	*Fragile X mental retardation 1*
FMRP	Fragile X mental retardation protein
FRDA	Friedreich's ataxia
FSHD	Fascioscapulohumeral muscular dystrophy
FTD/ALS	Frontotemporal lobar degeneration/Amyotrophic lateral sclerosis
FXS	Fragile X syndrome
FXTAS	Fragile X tremor associated syndrome
GCD	Genetics of Cognitive Dysfunction
HD	Huntington disease
hMeDIP-seq	Hydroxymethylated DNA immunoprecipitation sequencing
Hpf	Hours post-fertilization
IPSC	Induced pluripotent stem cells
iRNA	Interference RNA
LCM	Laser Capture Microdissection
LINE	Long interspersed nucleotide elements
MA	Mutant allele
MAF	Minor allele frequency
MBP	Methyl-CpG binding proteins
MCS	MeDIP-seq CpG score
MeDIP	Methylated DNA Immunoprecipitation
miRNA	Micro RNA
MJD	Machado Joseph disease
MRE-seq	Methylation sensitive restriction enzyme sequencing
mRNA	Messenger RNA

NA	Normal allele
NGS	Next generation sequencing
NIID	Neuronal intranuclear inclusion disease
OPMD	Oculopharyngeal muscular dystrophy
PCR	Polymerase chain reaction
PolyA	Polyalanine
PolyQ	Polyglutamine
PWM	Position weight matrix
RAN	Repeat-associated non-AUG
RBP	RNA binding protein
RNA	Ribonucleic acid
RNABP	RNA binding proteins
RU	Repeat units
SBMA	Spinal bulbar muscular atrophy
SCA	Spinocerebellar ataxia
SINE	Short interspersed nucleotide elements
SNP	Single nucleotide polymorphism
ssDNA	Single-strand deoxyribonucleic acid
STR	Short tandem repeats
TAB-seq	Tet-assisted bisulfite sequencing
TET	Ten eleven translocation
TFBS	Transcription factor binding sites
TSS	Transcription start site
UTR	Untranslated region
WG-Bis-seq	Whole-genome bisulfite sequencing
WG-oxBis-seq	Whole-genome oxidative bisulfite sequencing
XYLT1	*Xylosyltransferase 1*

1. INTRODUCTION

1.1 Repetitive elements in human genome

The study of the human genome has shown, contrarily to the expectations, that only about 1.1% are exonic coding sequences translated in proteins. Non-coding DNA, initially referred as "junk DNA", has been gaining scientific attention because it plays a crucial role in the control and regulation of gene expression [1]. The non-coding DNA contains repetitive sequences, belonging to two main categories [1]:

1) Interspersed repeats, they are classified into transposons or retrotransposons. While transposons can only be removed from one place and inserted into another, retrotransposons have the ability to copy themselves, increasing its number. Members of this family include long interspersed nucleotide elements (LINE) and short interspersed nucleotide elements (SINE) [2].

2) Tandem repeats, present in 10-15% of the mammalian genome, are divided into three main groups: microsatellites (1 to 9 nucleotides), minisatellites (10 to 50 nucleotides) and satellites (more than 50 nucleotides) [1].

Microsatellites, also called short tandem repeats (STR), are an important component of our genome, representing approximately 3%, a proportion similar to protein-coding sequences. Thus they might play an important role in the human genome [3]. These repetitive sequences can be a source of genetic variability, and they are present at ~15% of all human genes, in promoters, introns or coding regions [4]. However, the expansion of these repeats beyond a given threshold is usually associated with genetic instability and disease[5].

1.2 Discovery of repetitive DNA associated to disease

In the early 90's, it was discovered that STR expansions are the genetic cause of several neurological diseases. The first repeat expansion diseases identified were fragile X syndrome (FXS) and spinal bulbar muscular atrophy (SBMA) [6,7]. Nowadays, it is known that STR expansions are responsible for more than 40 neurological and neuromuscular diseases [8], including spinocerebellar ataxias (SCAs), Huntington disease (HD) and myotonic dystrophy type 1 and 2 (DM1 and DM2) [3].

1.3 Repeat expansion diseases

Repeat expansion diseases can be caused by several repetitive motifs, varying in the genomic location and repeat tract sequence and length, being positioned in: **i) coding regions (exons)**, as occurs in numerous diseases mediated by polyglutamine (e.g., HD [9] and SCA7 [10]) and polyalanine proteins (blepharophimosis syndrome (BPES) [11] and oculopharyngeal muscular dystrophy (OPMD) [12]); **ii) 5' untranslated regions (UTRs)** as observed in fragile X tremor associated syndrome (FXTAS) [13] and SCA12 [14] ; **iii) 3' UTRs**, as seen in DM1 [15]; **iv) introns**, as in DM2 [16] or SCA37 [17] and **v) promoters**, as occurs in Baratela-Scott syndrome (BSS) [18] (**Figure 1.1**).

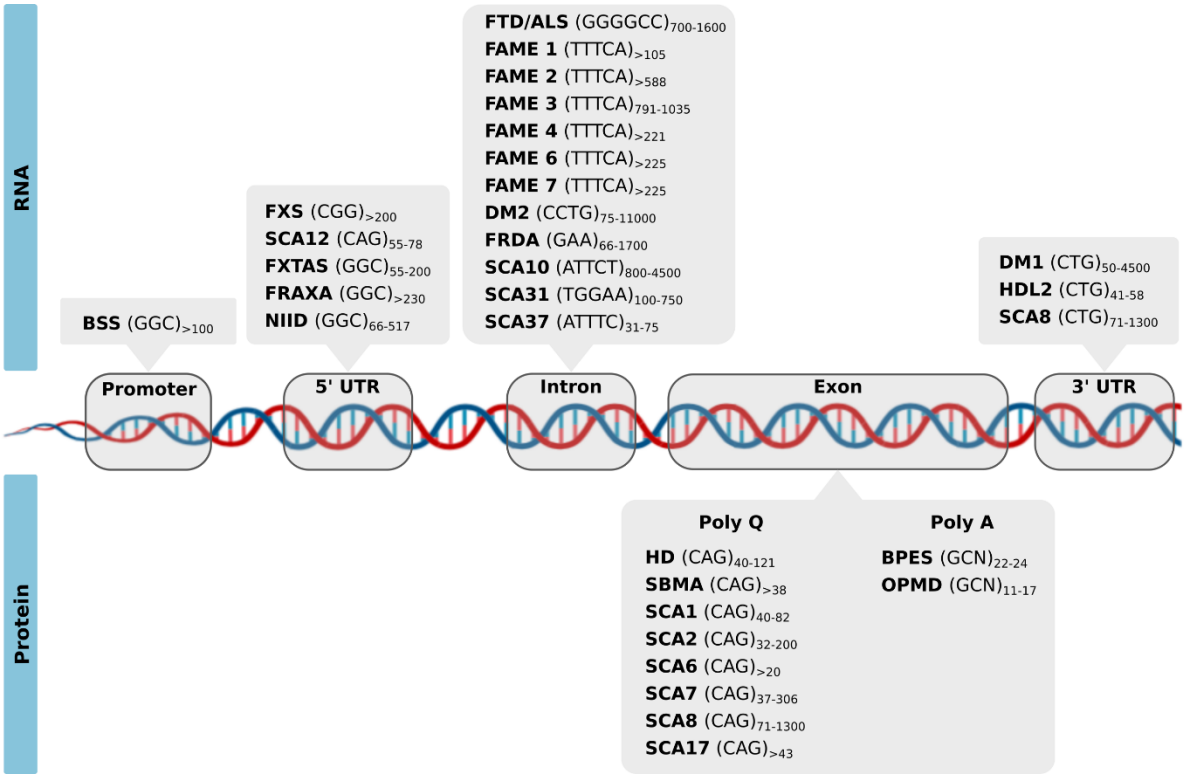

Figure 1.1 – Disease associated STR expansions can be located at coding and noncoding regions. In noncoding regions, STR expansions can be located at promoters, 3' and 5' untranslated regions (UTR) or introns. In coding regions, the expansion of STRs results in the translation of proteins with polyglutamine (PolyQ) or polyalanine (PolyA) tracts. *Schematic representation designed based on Loureiro et al 2016 and Sznajder & Swanson 2019.*

There is a relationship between the number of repeats and its location in the genome. Repeat tracts located in noncoding regions have a higher propensity for expansion than the coding sequences [5]. Unstable trinucleotide repeats are the most common STR causing disease [19] but there are also tetranucleotides (CCTG), pentanucleotides (ATTCT, TGGAA, ATTTC and AAGGG),

hexanucleotides (GGCCTG, CCCTCT, and GGGGCC), and one dodecanucleotide (CCCCGCCCCGCG) [20].

Until recently, the known repeat diseases were caused by only one simple STR expansion. Recently, it was identified a new type of mutation in repetitive regions of the genome. First, Sato et al 2009 [21] identified a $(TGGAA)_n$ pathological insertion located within a non-pathologic $(TAAAA)_n$ positioned in a noncoding region of *brain expressed associated with Nedd4 (BEAN)* gene, as the molecular cause of SCA31. In 2017, Seixas and collegues, led by Dr. Isabel Silveira, identified a pathological $(ATTTC)_n$ insertion in the noncoding region of *DAB1* gene, located in chromosomal region 1p32.2, being the molecular cause of SCA37 [17]. After this, 6 other repeat insertions have been reported as the genetic cause of six familial adult myoclonic epilepsies (FAME 1, 2, 3, 4, 6 and 7), in several genes with different functions, like *SAMD12* [22], *TNRC6A* and *RAPGEF* [23], *MARCH6* [24], *STARD7* [25] and *YEATS2* [26]. It is important to note that often the expansion of the same STR leads to the same phenotypic consequences regardless of the gene in which the expansion is located. This is the case of, for example, CAG repeat expansions that, though it can be located in different genes, the phenotypic result is SCA. These facts indicate that the pathology is not only related to the gene where the expansion is located, but also to the STR.

The enormous diversity of genes containing these repeat insertions highlights the importance of investigating dynamic mutations as genetic causes of neurological and/or neurodegenerative diseases.

1.4 Mechanisms of pathogenicity in repeat diseases

The pathogenic mechanisms associated with repeat expansions are complex. Depending on the genomic location of the repeat – if located in protein coding regions or noncoding regions (5' UTR, 3' UTR or intronic sequences) – three principal mechanisms can occur: **(a)** polyglutamine (PolyQ) or polyalanine (PolyA) gain-of-function; **(b)** gene loss-of-function and **(c)** RNA gain-of-function (Figure 1.2).

a) In coding regions, the expansion of CAG or GCN repeats results in the translation of proteins with PolyQ or PolyA tracts. This will interfere with protein normal function, altering its conformation and leading to intracellular inclusions, either in cytoplasm **(Figure 1.2, point 3)** or nucleus **(Figure 1.2, point 4)** [19,27]. Protein gain-of-function is observed in HD, SCA1, SCA2, SCA3, SCA6, SCA7 and SCA17 [28].

In noncoding regions, there are two main mechanisms:

b) **Gene loss-of-function**. It occurs when the repeat expansion leads to a decrease or a completely absence of gene expression, originating the reduction or absence of protein product **(Figure 1.2, point 6)**. This mechanism is present in, at least, two diseases: i) FXS,

where a CGG expansion above 200 repeats in the 5'UTR of *Fragile X Mental Retardation 1 (FMR1)* gene influences the methylation state of a proximal CpG island located at the promoter region, altering the conformation of active chromatin that results in complete silence of *FMR1* gene [29,30] and ii) Friedreich's ataxia (FRDA), the most common inherited ataxia, where a GAA expansion located in intron 1 of *frataxin* (*FXN)* gene leads to *FXN* silencing, what could be due to heterochromatin formation that blocks *FXN* transcription initiation and elongation [31].

c) **RNA gain-of-function.** Transcription of the repeat produces toxic RNAs that can trigger several parallel pathogenic mechanisms, including: i) RNA foci formation **(Figure 1.2, point 10)**; ii) Repeat-mediated RNA binding protein (RBP) sequestration and mis-splicing **(Figure 1.2, points 10 and 11)**; iii) Nucleocytoplasmic transport dysregulation **(Figure 1.2, point 13)**; iv) Transcription abortion by R-loop formation **(Figure 1.2, point 8)** and v) Repeat associated non-AUG (RAN) translation **(Figure 1.2, points 15 and 16)**, leading to the disruption of important cellular mechanisms, originating cellular stress. These RNA gain-of-function mechanisms have been seen in FTD/ALS and all of them contribute to the pathogenicity of the disease [32].

The pathogenic mechanisms underlying repeat expansion diseases are complex and have been comprehensively reviewed in Loureiro et al 2016 [33] and Sznajder 2019 [34].

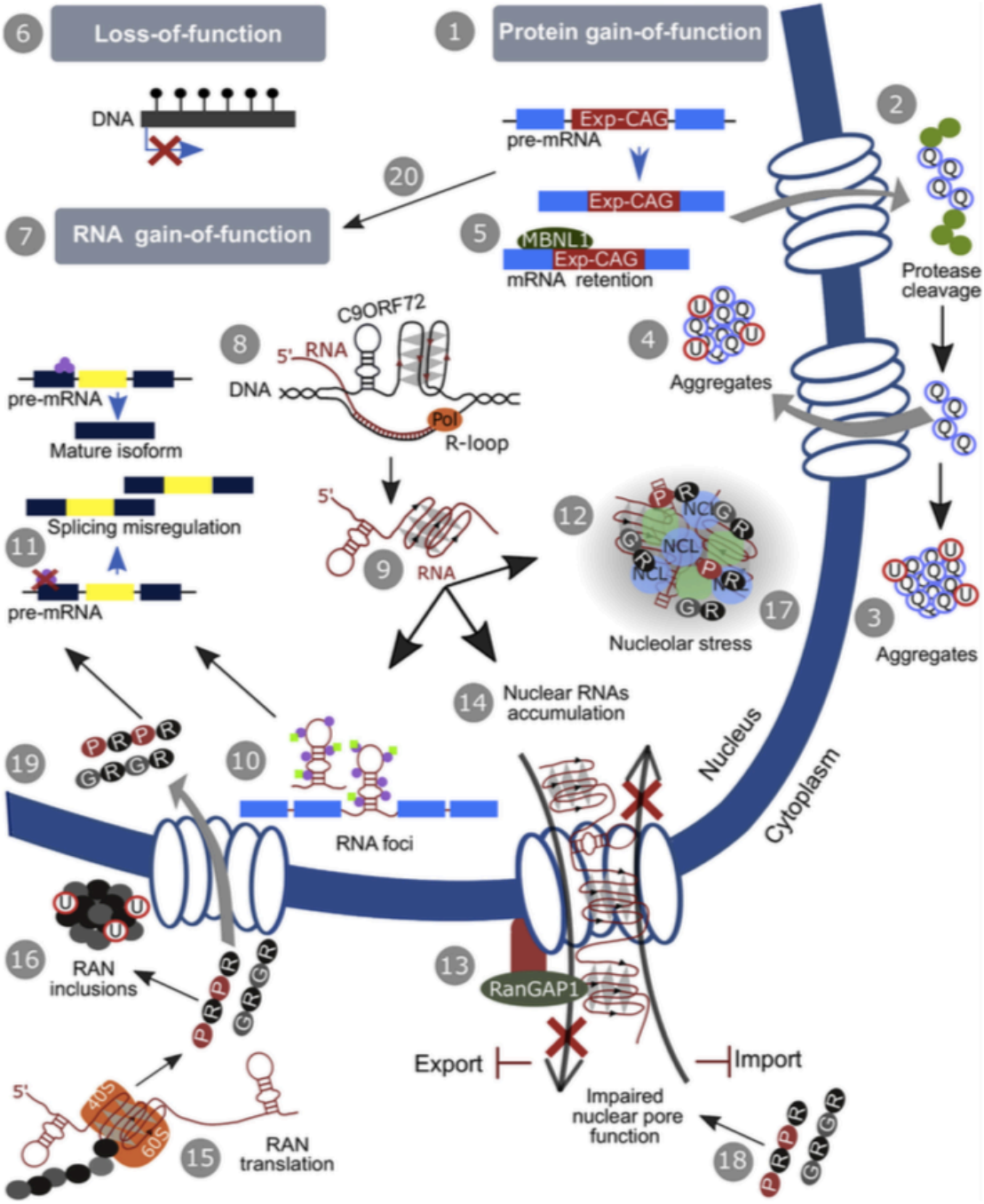

Figure 1.2 - Mechanisms of pathogenicity associated to repeat expansion diseases. There are three main pathogenic mechanisms. **In coding regions**, the production of homopolymeric stretches of glutamine (Q) or alanine (A) results in **protein gain-of-function (1)**. These PolyQ stretches are cleaved (2) and aggregates to cause cellular toxicity either in cytoplasm (3) or in the nucleus (4). The CAG expanded mRNA can also interact with MBNL1, that retains the mRNA in the nucleus, decreasing translation (5). **In noncoding regions,** there are two possible mechanisms: **gene loss-of-function (6)**, when the repeat expansion leads to a reduction or a completely absence of gene expression; and **RNA gain-of-function (7),** in which the production of toxic RNA molecules can trigger several parallel pathogenic mechanisms as: the formation of G-quadruplex structures and DNA:RNA loops (8 and 9); RNA foci formation (10); RNABP sequestration and splicing mis-regulation (11); nucleolar stress mediated by the recruitment of nucleolin (12); nucleocytoplasmic transport dysregulation (13) and the consequent accumulation of mRNAs in the nucleus (14); RAN-translation of repeat expansions (15) produces RAN proteins that are capable to form RAN inclusions in the cytoplasm (16) or in the nucleus (17), interfere in the nucleocytoplasmic transport (18) and disrupt mRNA splicing (19). *Adapted from Loureiro et al 2016.*

Another molecular mechanism that can contribute for disease pathogenicity is bidirectional transcription. It has been identified in several repeat diseases like SCA7 [35], SCA8 [36], HD [37], HDL2 [38], FXTAS [39], DM1 [40] and frontotemporal degeneration/amyotrophic lateral sclerosis (FTD/ALS) [41], and it happens when the transcription occurs in both sense and antisense orientation. Normally, sense strand is protein coding and is present at a higher copy number, comparing with the antisense strand that is noncoding. However, the antisense transcript can regulate gene expression at different levels [42–44], including: (a) interference with the sense transcript, causing the collision of RNA polymerases that are transcribing both strands in opposite directions; (b) recruitment of chromatin remodeling complexes that are responsible for histone modifications and DNA methylation, interfering then with gene expression; (c) sense and antisense transcripts duplex formation, that could be confused with miRNAs, inducing the iRNA pathway, leading to gene silencing. In SCA8, bidirectional transcription results in the production of a polyQ protein from *ATXN8* gene and production of a toxic noncoding RNA from the antisense transcriptional unit (*ATXN8OS*). Thus, SCA8 is caused by both, RNA and protein gain-of-function [36], being an excellent demonstration of the combinatory effect of different pathogenic mechanisms, abrogating the idea that each disease is associated to a unique pathogenic mechanism.

1.5 Spinocerebellar ataxia (SCA)

Spinocerebellar ataxia (SCA) is an autosomal dominant inherited rare neurodegenerative disease [45] with an estimated prevalence of 1.6 to 5.6 out of 100 000 inhabitants [46,47]. SCA is characterized by a progressive cerebellar ataxia in which the cerebellum begins to degenerate, mainly due to Purkinje cell loss, and also brainstem and spinal cord degeneration [48,49]. Phenotypically, it results in an unsteady gait, clumsiness and dysarthria [45]. Currently, there is no treatment that can stop or delay the disease progression. SCA is considered a very heterogeneous neurological disease, once the causing mutations can occur in a variety of genes and genetic context, differing consequently in the phenotype. Currently, there are more than 40 different types of SCA identified [50], divided in 3 major genetic categories: **i)** SCA caused by tri-, tetra-, penta- or hexanucleotide repeat expansions in coding regions, that leads to the production of proteins with polyglutamine tracts (e.g., SCA1, SCA2, SCA3, SCA6, SCA7, SCA8 and SCA17) ; **ii)** the ones caused by tri-, tetra-, penta- or hexanucleotide repeat insertions in noncoding regions, that are normally associated to RNA toxicity (e.g., SCA8, SCA10, SCA12, SCA31, SCA36 and SCA37); **iii)** SCAs caused by missense, insertion, deletion or duplication mutations (e.g., SCA5, SCA11, SCA13, SCA14, SCA27) [51,52].

As referred before, Seixas and collegues identified a pathogenic $(ATTTC)_n$ repeat insertion located between a normal $(ATTTT)_n$ repeat expansion in *DAB1* gene as the genetic cause of SCA37[17]. We will, hereinafter, focus on SCA37.

1.5.1 Spinocerebellar ataxia type 37 (SCA37)

SCA37 was identified during a population-based survey in three different Portuguese families [46] and was posteriorly mapped and described by Seixas et al 2017 [17] (Figure 1.3). Affected individuals show a pure cerebellar ataxia with late onset, present gait and limb incoordination, dysarthria as first symptom and dysmetria in the lower and upper extremities [17].

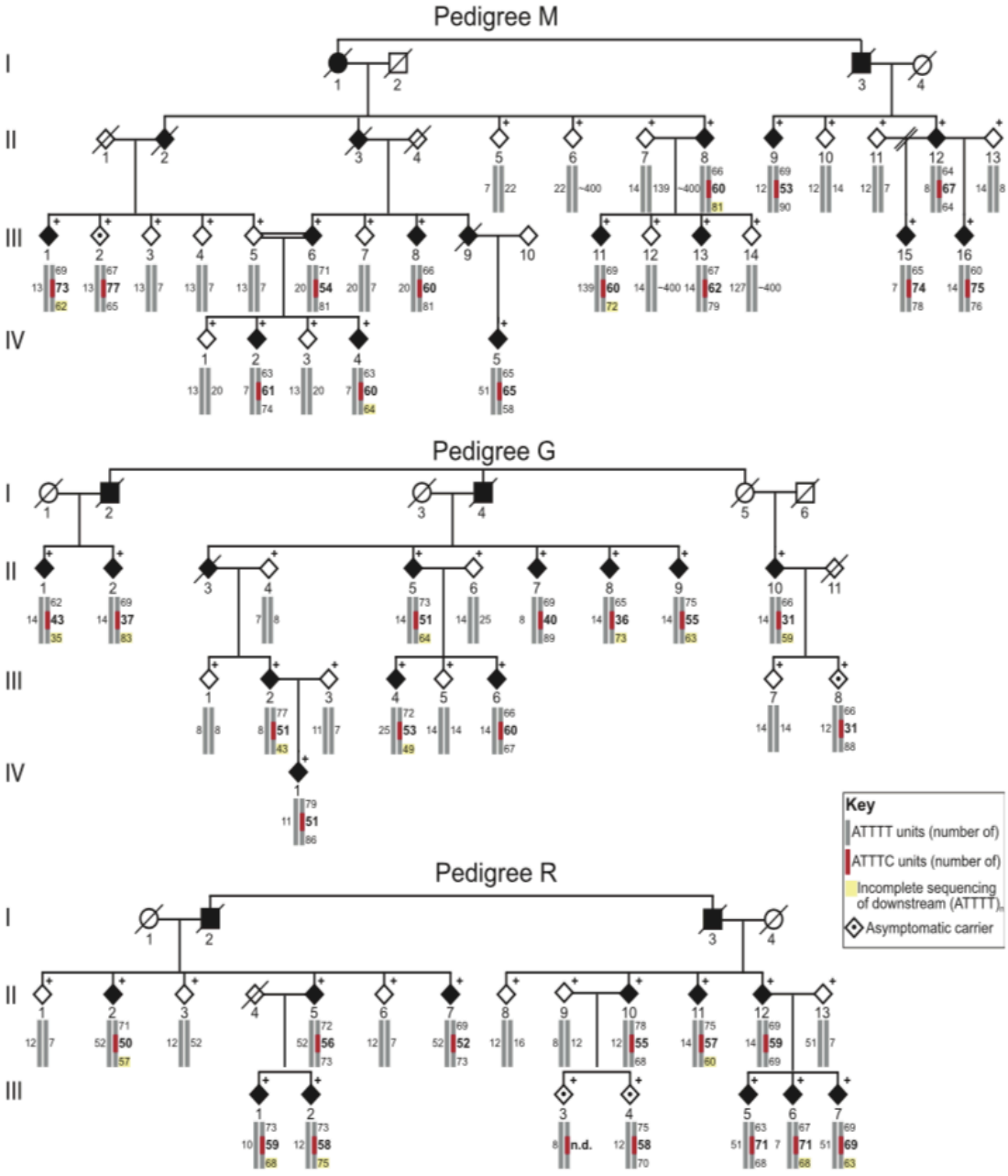

Figure 1.3 - Pedigree structures of the three Portuguese SCA 37 affected families. Symbols in pedigrees were modified for privacy protection. Abbreviations: n.d. not determined; +, individual for whom DNA is available. *Adapted from Seixas et al 2017.*

1.5.1.1 Molecular and genetic context of SCA37

SCA37 is caused by an $(ATTTC)_n$ repeat insertion within in a normal nonpathogenic $(ATTTT)_n$ expansion located in 5'UTR intronic region of *DAB1* gene (DAB1, reelin adaptor protein), in the chromosomal region 1p32.2 [17]. DAB1 protein is a signal transducer functioning downstream of reelin – a glycoprotein important in neurodevelopment. It belongs to a signaling pathway that controls the neuron positioning during neurodevelopment and adult neurogenesis [53], and it is known that *Dab1* mice mutants shows the *scrambler* and *yotari* phenotypes [54], confirming the extreme importance of this gene to a correct neurodevelopment.

The normal nonpathogenic alleles have the (ATTTT) repeat expansion varying between 7 and 400 units. The pathogenic alleles have the (ATTTC) insertion repeated for 31 to 75 times. Thus, generally, the affected individuals have the following configuration **$(ATTTT)_{60\text{-}79}$ $(ATTTC)_{31\text{-}75}$ $(ATTTT)_{50\text{-}90}$** [17] (Figure 1.4 B). Both ATTTT and ATTTC repeats are not interrupted [55].

The pathologic repeat insertion is located in the middle of an Alu Jb element [17] (Figure 1.4 A). This is interesting because there are several other diseases in which the repeat insertion is located either in middle or in the poly-A regions of Alu elements [56], as FRDA [57], SCA10 [58], SCA31 [21] and FAME [22,59]. Due to similarities between the different loci containing repeat insertions, the mutational events responsible for the formation of these expansions could occur frequently in the genome. Loureiro et al., 2019 [55] reported that, in SCA37 locus, the pathogenic repeat insertions resulted from a nucleotide substitution followed by an increase in repeat size. Additionally, the fact that the repeat region is highly polymorphic, mutable and unstable and is located in the middle of an Alu element (that are highly dispersed in the genome [1]) could suggest an implication of AT-rich repeats of Alu elements in other neurologic and/or neurodegenerative diseases.

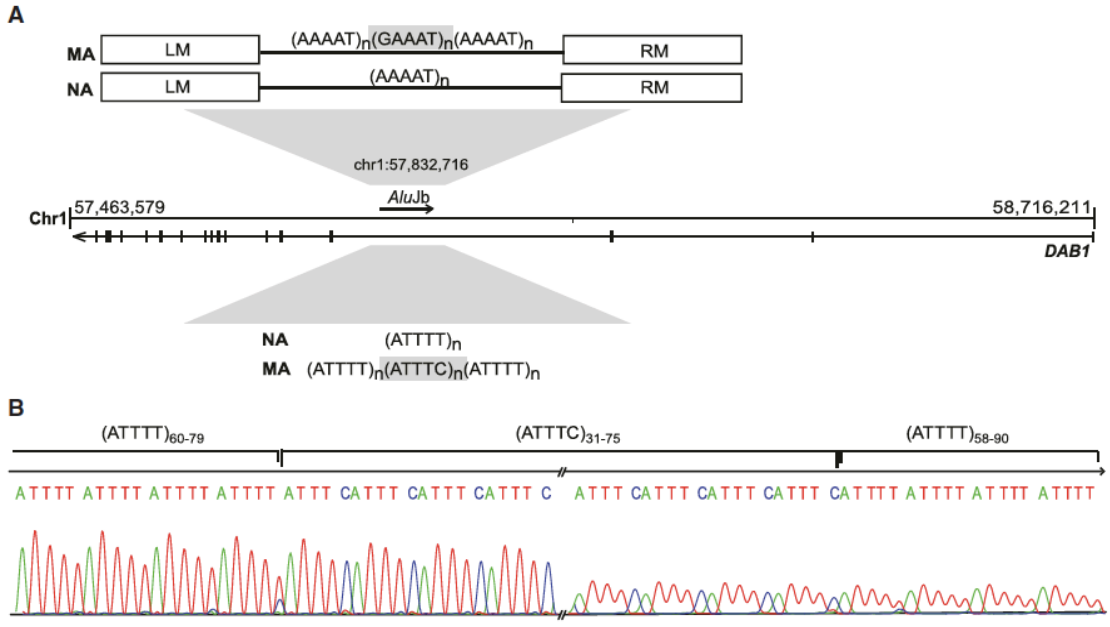

Figure 1.4 - Genetic and molecular context of the repeat. (A) Schematic representation of the pathogenic $(ATTTC)_n$ insertion at the mutant allele (MA) and the normal $(ATTTT)_n$ insertion at the normal allele (NA), represented either in the AluJb-oriented strand and *DAB1*-oriented strand. (B) Sequencing analysis of *DAB1* gene showing the $(ATTTC)_n$ repeat insertion within a normal $(ATTTT)_n$ repeat. *Adapted from Seixas et al., 2017.*

1.5.1.2 Mechanism of pathogenicity

In 2017, Seixas et al reported that the transfection of plasmids containing the pathogenic insertion $(ATTTC)_{58}$ flanked with ATTTTs and AluJb monomers in HEK293T cells, results in the formation of RNA aggregates, contrarily to cells transfected with normal ATTTT insertion, where there is no RNA foci formation. This RNA foci have the potential to cause cellular toxicity by triggering the RNA gain-of-function pathways mentioned in section 1.4. In other experiment, *in vivo* studies in zebrafish embryos demonstrated that the injection of a one- to two-cell-stage zebrafish with RNA containing the repeat insertion $(AUUUG)_n$ flanked by $(AUUUU)_n$, results in a higher mortality rate compared with the embryos injected with the normal alleles $(AUUUU)_n$, suggesting that the RNA repeat insertion has a deleterious effect *in vivo* (Figure 1.5).

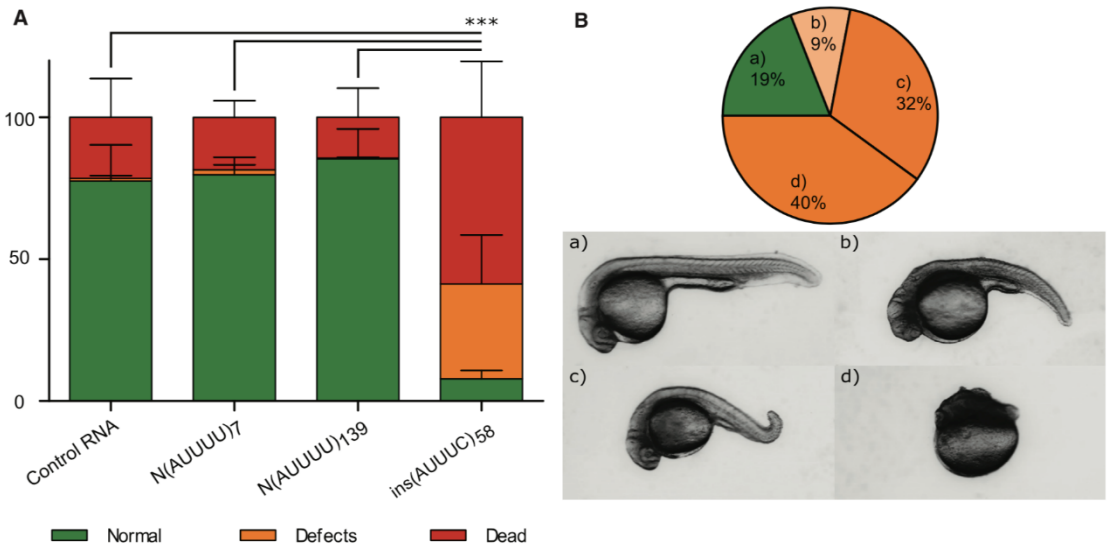

Figure 1.5 - SCA37 RNA gain-of-function. (A) Percentage of embryos that developed normal (wild-type phenotype), with developmental defects (defects) or died (dead) 24 hpf after the injection with a control RNA, the normal $(AUUUU)_n$ RNA and the pathogenic RNA $(AUUUC)_{58}$. (B) Different phenotypic classes observed upon injection with the pathogenic RNA $(AUUUC)_{58}$, being a) wild-type; b) severe defects in the tail and head; c) mild defects in the tail and c) severe defects in the anterior-posterior axis. *Adapted from Seixas et al., 2017*.

1.5.1.3 Genetic instability

Although no evidence of anticipation in the age of onset is reported for SCA37 [60], there is an inverse correlation between the repeat insertion size and the disease age of onset [17,61] (Figure 1.6), being this more evident in males than in females [61].

Additionally, it is noted an intergenerational instability in which the number of ATTTC repeats increase throughout transmissions. This instability is higher when the father is the transmitting parent (all the paternal transmissions results in an increase of repeat size) than in maternal transmissions (in which the repeat length increases only in 67% of the maternal

transmissions) [17]. So, the $(ATTTC)_n$ repeat insertion appears to be highly unstable, mainly when the father is the transmitting parent (Figure 1.7). Until the date, there are no evidence of ATTTC repeat contractions. All the described SCA37 families showed 100% penetrance, although this penetrance is age dependent [17] .

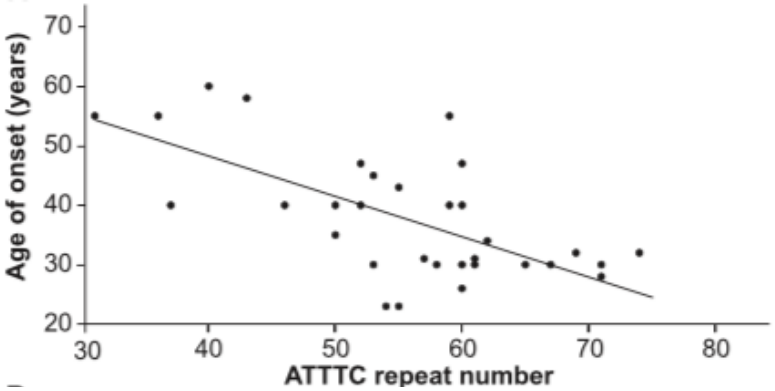

Figure 1.6 - Genetic instability. Graphic representation of the inverse correlation between the length of the $(ATTTC)_n$ insertion and age of onset. Affected individuals with larger insertion sizes have earlier onset ($r = -0.68$, $p < 0.001$, $n = 33$). *Adapted from Seixas et al., 2017.*

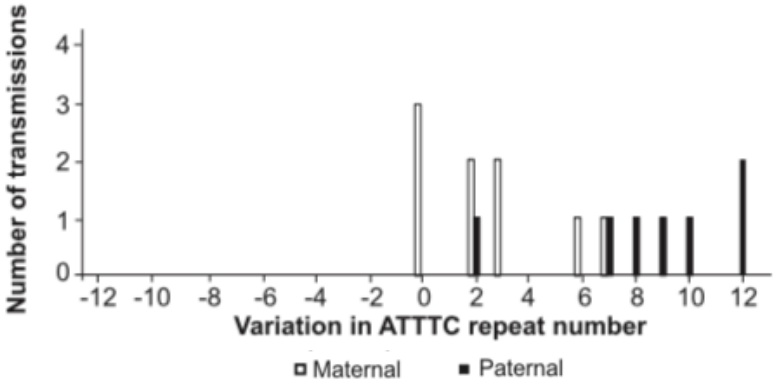

Figure 1.7 - Intergenerational instability of the ATTTC repeat insertion. Schematic representation of the variations in ATTTC repeat number throughout parent-to-offspring transmissions with maternal (white bars) or paternal (black bars) origins. *Adapted from Seixas et al 2017.*

Seixas et al 2017 have previously shown that the SCA37 pathogenic alleles are unstable upon transmission. However, beyond the enormous instability of the ATTTC repeat insertion, it is also extremely important to highlight the instability of the ATTTT repeat at the normal alleles and/or flanking the ATTTC repeat insertions at the pathogenic alleles. Looking at the pedigree representations in Figure 1.3, it is possible to observe that the 5' ATTTT and 3' ATTTT flanking the repeat insertion at the pathogenic allele are unstable upon transmission (e.g., Pedigree M - individual II.12 to its offspring III.15 and III.16). Additionally, it is noted a huge instability in the large nonpathogenic alleles.

Aware of this instability, it was further conducted a mutation screening aiming to investigate the genetic evolution of the region, in order to increase the understanding of how the $(ATTTC)_n$ repeat insertion occurred.

1.5.1.4 SCA37 haplotype

Next generation sequencing (NGS) of the candidate genes was carried out in both affected individuals and unaffected relatives. In this study, Loureiro et al 2019 [55] identified 20 heterozygous intergenic and intronic variants, appearing only in affected individuals and absent or with a frequency inferior to 1% in the 1000 Genomes Project or dbSNP databases. After haplotype analysis with variants with a frequency lower than 0.1%, they found three additional pedigrees with the same core haplotype as the three families used to map the gene mutation and verified that these variants were very rare in the control population, what makes the SCA37 haplotype rare in Portuguese population. Later, Loureiro et al 2019 [55] investigated the mutational mechanism behind SCA37 and constructed a haplotype network, analyzing the different haplotypes of the short pure stable with $(ATTTT)_{<100}$ (haplotypes I, II, III, V, VI and VII), large with $(ATTTT)_{>100}$ (haplotypes I, IV and V) and nucleotide interrupted alleles (haplotypes I and VIII), showing the phylogenetic relationship among them (Figure 1.8). The authors found a total of eight different haplotypes in the Portuguese population, with all the SCA37 studied individuals sharing the same haplotype (haplotype VIII). Interestingly, haplotype VIII was also found in large unstable alleles and, remarkably, haplotype I was found in short pure stable, interrupted and large alleles. This means that short pure alleles are clearly unstable and can originate the interrupted and the large alleles. This suggested that *cis-elements* present at the repeat flanking region originally at individuals with haplotype I, could have been crucial for the occurrence of the mutation leading to SCA37.

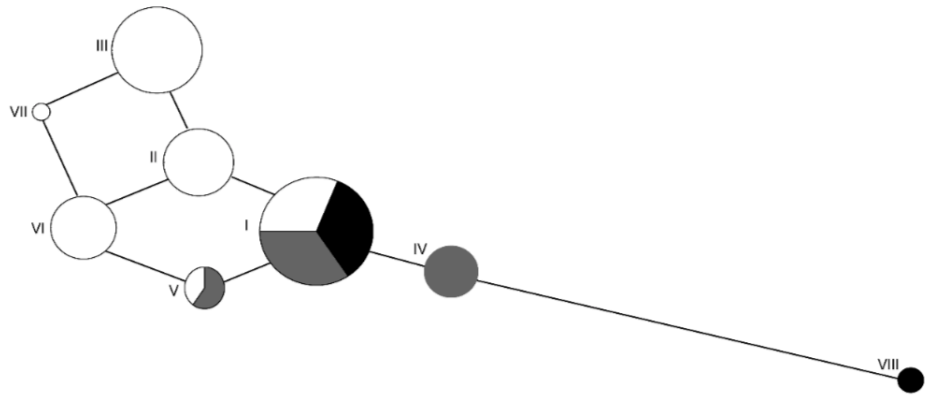

Figure 1.8 - Phylogenetic relationship between non-pathogenic *DAB1* alleles. Short pure alleles (<100 ATTTT) are represented in white, large pure alleles (>100 ATTTT) in black and interrupted alleles in grey. Circle size is proportional to number of chromosomes tested and line length is proportional to genetic distance among haplotypes. *Adapted from Loureiro et al 2019.*

1.5.1.5 The flanking region of the SCA37 repeat

We know that the $(ATTTC)_n$ repeat insertion is located in the middle polyA region of an AluJb transposable element [17]. Additionally, the previously conducted NGS studies identified in affected subjects 20 variants in the candidate region containing the *DAB1* gene, encompassing 3.4 Mb [17]. Several of these variants had a frequency lower than 0.1% and they have allowed to identify several SCA37 families. Later, additional SNPs spanning a region of 723 kb flanking the repeat have been crucial to uncover the mutational mechanism leading to SCA37, by Loureiro et al 2019 [55]. These SNPs, especially those associated with the SCA37 haplotype, might be located at important regulatory regions, like CpG dinucleotides or CTCF-binding sites (CTCF-BS), or potentially influencing the methylation status of the repeat flanking region, contributing to repeat instability. Also, the fact that there is a common haplotype found in short pure, interrupted and large unstable alleles raises the hypotheses of cis-genetic factors being responsible for that instability.

Given that, further investigations are needed, with a special focus on the analysis of the repeat flanking region, searching for possible regulatory regions that could be affected by the presence of these genetic variants.

1.6 Potential chromatin rearrangements and epigenetic mechanisms in SCA37

Epigenetics is defined as a set of modifications in DNA structure and associated histone proteins that do not involve alterations in the DNA sequence itself, representing an inherited biological phenomenon capable of producing alterations in gene expression [62]. Epigenetic alterations can influence gene expression in several repeat expansion diseases, because the DNA is wrapped around histone proteins (H2A, H2B, H3 and H4), forming the nucleosome, which undergoes structural alterations that can facilitate or inhibit transcription. When these nucleosomes are in a relaxed form (euchromatin) the transcription machinery is capable to bind and transcribe the respective genes. When the nucleosomes are in a compact state (heterochromatin) the DNA is inaccessible to the transcriptional complex and the transcription is inhibited [63].

Several epigenetic mechanisms contribute to changes in chromatin state, including different combinations of post-translational histone modifications (either methylation or acetylation of lysine and arginine amino acid residues in the N-terminal histone tails) or methylation of the 5' carbon of a cytosine, originating 5-methylcytosine (5-mC) in CpG dinucleotides (Figure 1.9) [64].

DNA methylation constitutes an epigenetic mark essential for the normal establishment of embryonic developmental programs and plays important roles in several processes like genomic imprinting, X-inactivation and gene repression [65]. Although the mechanistic association between DNA methylation and gene silencing is not completely understood, there are supportive material that associates DNA methylation in regulatory regions (like promoters or enhancers) with gene silencing [66,67], due to the functional interaction between 5-mC and specific transcription factors [68,69].

Gene silencing by DNA methylation can occur through different mechanisms like blocking transcriptional machinery or recruitment of histone deacetylases and methyl-CpG-binding proteins (MBP) that modify chromatin conformation [70].

There is another factor contributing to a higher-order structure of chromatin organization inside the nucleus, the CCTC-binding factor (CTCF). In the past years, CTCF has been shown to function both as a classical transcription factor [71] and as a chromatin insulator that blocks the interaction between a promoter and proximal enhancers or silencers [72], throughout the modulation of the tridimensional organization of chromatin [73].

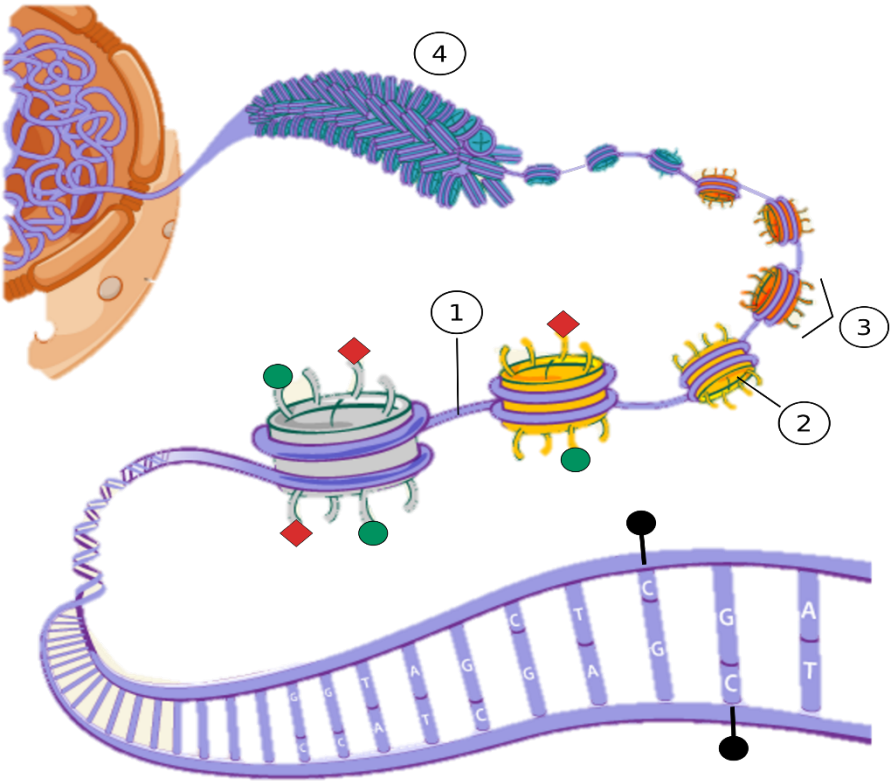

Figure 1.9 - Epigenetic organization of chromatin. In eukaryotic cells, most DNA **(1)** is wrapped around a core of histone homodimers **(2)**, forming nucleosomes **(3)**, the basic unit of chromatin **(4)**. There are different epigenetic mechanisms responsible for the alteration of chromatin conformation, from an active state (euchromatin) to an inactive state (heterochromatin) or *vice-versa* altering, consequently, gene expression. DNA of a gene promoter can be methylated (**black circles**) at the CpG dinucleotides, normally leading to gene silencing. However, DNA is not independent of its associated histones, existing also several post-translation modifications in histones that alter chromatin conformation, as histone methylation (**green circles**) or acetylation (**red diamonds**) in the N-terminal histone tails. *Edited from a broadinstitute.org image*.

The understanding of the higher order chromatin organization increased the importance of studying the flanking regions of the repeat, searching for flanking elements that might possibly contribute to repeat instability.

In repeat diseases, there are several factors known to influence the repeat instability. Firstly, the nucleotide sequence of the repeat and the repeat tract length. Secondly, the genetic context of the repeat is extremely important and definitely contributes to the complexity of the repeat diseases, once they can comprise origins of replication [74], TFBS [75], sense or antisense promoters [76], CTCF-BS and epigenetic mechanisms as CpG methylation [77], histone modifications or alterations in chromatin structure [78,79].

Thus, the repeat expansion instability might result not only from the size of the repeat itself, but also from the genetic variants present at the repeat flanking regions.

1.6.1 DNA methylation

DNA methylation is an epigenetic mechanism in which a methyl group is covalently transferred from S-adenosyl-methionine to the 5' carbon of a cytosine, originating 5-mC [80,81]. During development, the methylation process is dynamic and catalyzed by *de novo* methyltransferases DNMT3A and DNMT3B, and the 5-mC state is maintained and stabilized throughout cell division by other methyltransferase, DNMT1 [82].

Normally, DNA methylation occurs in the cytosine-phosphate-guanine (5'-CpG-3') dinucleotides and genomic DNA regions with 200 to 500 bp having a GC content higher than 50% and a CpG observed/expected ratio of more than 0.6 are defined as CpG island (CGI) [83]. Approximately 50%-60% of genes contain CpG islands, normally proximal to the promoter [84,85]. Most of CpG sites randomly distributed in the human genome are methylated, but CGIs are mostly unmethylated in normal tissues [86]. Nowadays, the mechanistic explanation for the maintenance of a methylation-free state in CpG islands in a globally methylated genome remains unknown [87]. These facts can lead us to the assumption that hypomethylated CGIs are the normal somatic cell state and acquire aberrant hypermethylation in disease [88,89]. Although it is known that CGIs are involved in transcription, its functional significance is just beginning to emerge and further studies are necessary to understand its functionalities [90].

Some repeat expansions are non-methylatable, as the case of $(CAG)_n$ (causing HD, SCA 1, SCA2, SCA3, SCA6, SCA7 and SCA17), $(CTG)_n$ (causing DM1 and SCA8), $(GAA)_n$ (causing FRDA) or $(ATTTT)_n/(ATTTC)_n$ (causing SCA37 and FAME). However, some repeats can be methylated and this methylation contributes to the pathogenesis of several repeat diseases (e.g CGG repeats in FXS) [88]. Although the epigenetic state of the repeat itself plays a crucial role in disease pathogenesis, the methylation state of the repeat flanking regions might also be implicated in disease. In fact, DNA methylation defects have been linked to some repeat expansion diseases, and it is known that many disease-causing repeats have a high percentage of CpG within the repeat, or have in its proximity CpG islands [77]. Consequently, the disease locus may be regulated,

at least in part, by DNA methylation. Interestingly, the methylation pattern of the repeat flanking region has been reported as a contributing factor for repeat instability in several repeat diseases, mainly at diseases in which the repeats are located in noncoding regions, like FXS [91,92], DM1 [93,94], FRDA [95,96], FTD/ALS [97–99], Fascioscapulohumeral dystrophy (FSHD) [100,101] or BSS [18]. However, DNA methylation can contribute to repeat instability in diseases in which the repeat is located in coding regions, as the case of SCA1 [78] and SCA7 [79]. The effects of DNA methylation in several repeat diseases are summarized in Table 1.1.

Moreover, DNA methylation is also involved in the regulation of the tridimensional chromatin structure by controlling the CTCF insulator activity, because CpG methylation of the CTCF-BS can inhibit CTCF binding [73]. Besides, CTCF-binding can also prevent the spreading of DNA methylation to the surrounding regions, playing an important role in the maintenance of DNA sequences as methylation-free [102], protecting the nearby promoters from silencing. Interestingly, CTCFs have also been implicated in the regulation of noncoding transcription and it is known that the expansion of STRs can interfere with insulator function. It has been identified CTCF-BS flanking the repeats in several repeat diseases, like DM1 [40,103], FRDA [104], FXS [39], FSHD [105], SCA7 [79], and HD [103], suggesting that epigenetic alterations in CTCF-BS might contribute to the repeat instability verified at these diseases.

Table 1.1 - Effects of DNA methylation in different repeat expansion diseases.

Repeat Diseases	Region of DNA methylation	Effects	References
Fragile X syndrome (FXS)	• CGG expanded repeat itself. • CpG island within *FMR1* promoter.	DNA methylation results in heterochromatin formation at *FMR1* promoter, leading to gene silencing.	[106] [107] [91] [108] [92].
Myotonic dystrophy 1 (DM1)	• Region upstream of the CTG repeat expansion.	Alterations in DNA conformation leads to a reduction of gene expression. It also affects the expression of two proximal genes (*SIX5* and *DMWD*).	[93,94,109,110].
Friedreich ataxia (FRDA)	• Region upstream of GAA repeat expansion.	Direct correlation between DNA methylation and repeat size. Inverse correlation between DNA methylation and age of onset.	[95,96].
Frontotemporal lobar degeneration and amyotrophic lateral sclerosis (FTLD/ALS)	• G4C2 repeat expansion. • CpG island located at *C9orf72* promoter.	It may have a protective effect against the RNA foci formation and dipeptide repeat protein aggregates. Higher methylation rate is correlated with a shorter disease duration.	[97–99].
Facioscapulohumeral muscular dystrophy (FSHD)	• D4Z4 repeat tandem array at 4q35.	The D4Z4 hypomethylation at the contracted alleles might be the consequence of nucleosome loss at the contractions.	[100,101].
Baratella-Scott syndrome (BSS)	• GGC repeat expansion located at *XYLT1* promoter region.	DNA methylation causes the transcription repression of *XYLT1* gene. It might be the cause of missing heritability patterns of the disease.	[18].

1.7 Genome vs epigenome

In the past few years, it has been investigated the interaction between the genome and the epigenome. Recent progress in this field showed that these interactions can occur in cis – where the DNA sequence and the specific haplotype can influence the pattern of DNA methylation (or even other epigenetic marks) at the locus. It has been reported by Weksberg et al 2019 [64] that, at a given allele-specific methylated gene, there is a strong correlation between CpG methylation patterns and local SNPs. Consequently, it can be assumed that epigenetic phenomena can be regulated/influenced by genetic variants [64].

Therefore, it is extremely important to know all the variants present at the repeat flanking region, once it is known that i) point mutations in the CTCF-BS can inhibit CTCF binding, resulting in conformational alterations in chromatin that leads to alterations in gene expression [79]; ii) point mutations in CpG dinucleotides can lead to alterations in the percentage of CpG in the region of interest and, consequently, alterations in CpG island methylation status; iii) there are evidences for the existence of SNPs overlapping CTCF-BS, abrogating CTCF-binding and, also, originating preferential CpG methylation [111,112].

2. AIMS

In SCA37, there is an increase in $(ATTTC)_n$ size during transmission to the next generation, with larger increases when the father is the transmitting parent. Moreover, the $(ATTTT)_n$ is unstable in both unaffected alleles and in flanking ATTTTs of pathogenic alleles. To identify putative cis elements that may be implicated in repeat instability and in the origin of pathological SCA37 allele, my master project aims to investigate:

1) Single nucleotide polymorphisms (SNPs) flanking the repeat region, that create or eliminate CpG dinucleotides and, consequently, have the potential to influence regulatory regions or the methylation rate of the SCA37 locus.

2) DNA methylation patterns of the SCA37 repeat flanking region in cells from different tissues, with variable repeat sizes, configurations and disease status.

To achieve my goals, the identification of SNPs will be performed through the analysis of the repeat and its flanking sequences. DNA methylation profiles will be assessed at each CpG dinucleotide upstream and downstream the repeat.

3. METHODOLOGY

3.1 Biological samples from subjects or transgenic zebrafish

This study was carried out using DNA samples previously extracted from peripheral blood cells of affected individuals, as well as anonymized control subjects [17,55]. DNA was also extracted from primary cultures of fibroblasts previously established from skin biopsies or control fibroblasts obtained from the biobanks of Target ALS and NINDS, USA. Absence of mycoplasma contamination in fibroblast cell lines was confirmed by PCR at the CCGen core, i3S – U.Porto. Zebrafish (*Danio rerio*) DNA was extracted from larvae resulting from crossing transgenic fish lines, engineered by Joana R. Loureiro, in order to express the pathogenic $(ATTTC)_{120}$ or the nonpathogenic $(ATTTT)_{7}$ repeat in the cerebellum. Approval for these studies was obtained from the i3S Ethics Committees.

3.2 Genotyping

3.2.1 Amplification of *DAB1* pentanucleotide repeat and flanking regions

The DNA sequence corresponding to the repeat and its flanking region was amplified by standard PCR. The amplification of normal alleles (with $(ATTTT)_{<50}$), large nonpathogenic (with $(ATTTT)_{>50}$), interrupted alleles (with $(ATTTT)_{n}$ interrupted by other nucleotide motifs) and mutant alleles (with $(ATTTC)_{n}$ insertion) was performed using the PCR mix and thermocycler conditions specified in Appendix A. The primers used and the respective sequence are discriminated at Appendix B.

Amplified PCR products were visualized in a 1% agarose gel, stained with GreenSafe Premium (Nzytech) and under UV transilluminator. Molecular weight was assessed by comparisons with Gene Ruller DNA Ladder Mix (ThermoScientific).

3.2.2 Sanger sequencing of flanking region in short normal alleles

After short normal alleles amplification, the PCR products were cleared by two enzymes. While exonuclease I is responsible for the removal of the remaining single stranded primers, the alkaline phosphatase removes the remaining dNTPs in the mixture, both interfering with the further sequencing reaction. The ExoProStar reaction is carried out in a total volume of 5 μL, containing

0.5 μL of exonuclease I (Illustra™; Stock concentration: 10U/μL), 0.5 μL of alkaline phosphatase (Illustra™; Stock concentration: 1U/μL) and 4 μL of PCR volume. The enzymatic reaction is conducted at 37ºC for 15 minutes, followed by 15 minutes at 80ºC. Purified PCR products were then sequenced by Sanger sequencing.

3.2.3 Sanger sequencing of flanking region in large normal and mutant alleles

After agarose gel electrophoresis, the PCR products corresponding to large normal or mutant alleles were sliced from the gel and extracted using the Zymoclean™ Gel DNA Recovery Kit, according to manufacturer's protocol (Appendix C).

DNA extracted from the gel was quantified with NanoDrop 1000 Spectrophotometer and the 260/280 and 260/230 ratios are annotated. To validate the DNA integrity, 1 μL of DNA is loaded at a 1% agarose gel, in a mixture containing 1μL of DNA, 1 μL of Loading Die and 5 μL of ultra-pure water, and the gel electrophoresis is conducted at 90V. Purified DNA was then sequenced by Sanger sequencing.

3.2.4 DNA analysis by Sanger sequencing

Sanger sequencing is a technique developed in 1977 by Frederick Sanger and colleagues [113]. This technique relies on the utilization of chain-terminating dideoxynucleotides (ddNTP) fluorescently labelled and lacking a 3'-OH group, that results in the extension stop when the ddNTP is incorporated. The amplified products (with a chain-terminating ddNTP at each nucleotide throughout the sequence) are then separated by size using capillary electrophoresis that detects and record the fluorescence, resulting in an output in the form of a chromatogram (with different colors for each nucleotide).

In this work, Sanger sequencing was performed with 2.5 μL of DNA template, 1 μL of BigDye™ Terminator v3.1 plus 1 μL of BigDye sequencing buffer (Applied Biosystems), 0.5 μM of primer and water up to 10 μL. In order to analyze the presence of SNPs at the repeat and its flanking region, the reactions were performed with two external primers (Alu CTCF 2F and Alu CTCF 2R) and five internal primers (Alu 24F, Alu 24R, Frag 2F, SeqLocusA (F) and SeqLocusB (R)) (Appendix B) (Figure 3.1). The sequencing conditions were carried out with an initial denaturation step at 95ºC for 1 minute, 35 cycles of 95ºC for 10 seconds, 56ºC for 30 seconds and 60ºC for 4 minutes, and a final extension step of 60ºC for 10 minutes. The reaction was purified using a Sephadex Column (detailed protocol in Appendix D).

The templates were then analyzed on a Capillary Electrophoresis Sequencer (Applied Biosystems 3130 xl and Applied Biosystems 3130 genetic analyzers).

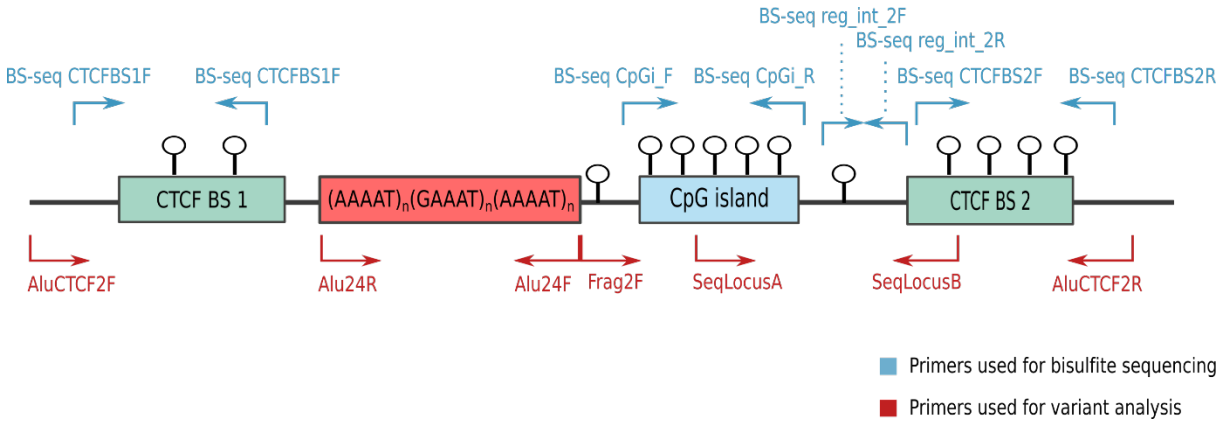

Figure 3.1 - Schematic representation of the repeat and its flanking region, with the location of the primers used for both SNP and DNA methylation analysis. In blue (at the top), are the primers used for Bisulfite sequencing analysis and its respective binding location (approximately). In red (at the bottom), are the primers used for the repeat and flanking region PCR amplification and Sanger sequencing for SNP analysis, as well as its respective binding locations. The primer ID and sequence can be consulted in Table B.1 and Table B.2 in Appendix B.

3.3 Sequence analysis and SNP annotation

The sequencing results are analyzed in SeqScape® v 2.7 (Applied Biosystems). This genetic analyzer software includes a Variant Reporter Software, a tool designed for reference-based analysis in which sequence comparisons are made for variant identification, such as SNP detection and validation (detailed analysis protocol available in Appendix E).

The statistical analysis of the nucleotide variants by Fisher's exact test was performed in R v.3.5.1. (R Core Team, 2018).

3.4 DNA CpG methylation analysis by bisulfite sequencing

In this work, DNA was isolated from fibroblast cell lines using QIAamp® DNA Mini Kit (QIAGEN), according to manufacturer procedures (Appendix F). DNA was quantified with NanoDrop 1000 Spectrophotometer.

3.4.1 Bisulfite treatment

The analysis of DNA CpG methylation was performed by Bisulfite Sequencing (BS-seq). BS-seq is probably the most accurate protocol and, consequently, the most widely used for analyzing DNA methylation. The first reported method using bisulfite-treated DNA to obtain the methylation

profile was elaborated by Frommer et al in 1992 [114]. This method relies on the assumption that the bisulfite treatment chemically converts cytosine residues to uracil, while the 5-methylcytosines remain unaffected [115] (Figure 3.2). Bisulfite treatment was performed using EZ DNA Methylation Direct™ Kit (Zymo Research Corp) in three basic steps: (1) Denaturation – incubation at 98ºC during 8 minutes to transform dsDNA into ssDNA; (2) Conversion – Incubation with sodium bisulfite at 64ºC and low pH during 3.5 hours and deamination of the cytosine residues in the fragmented DNA; (3) Desulphonation – incubation at high pH at room temperature for 20 minutes to remove the sulfite, originating the uracil (detailed protocol in Appendix G).

Bisulfite-converted DNA is quantified with NanoDrop 1000 Spectrophotometer. After bisulfite treatment, DNA is single stranded, once the original base-pairing no longer exists (because the unmethylated cytosines are converted into uracil). Given that, the absorption coefficient at 260 nm will be the same of RNA. When choosing the type of analyte to quantify in NanoDrop specifications, RNA-40 is selected (40 μg/ml for Abs_{260} = 1).

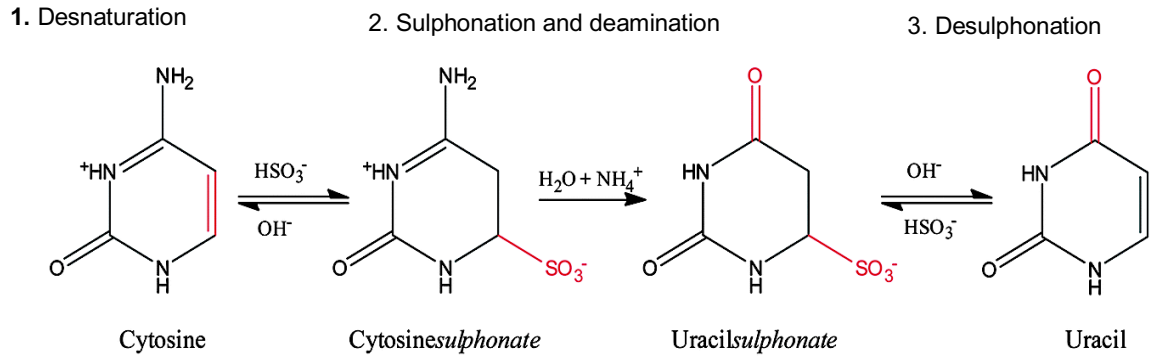

Figure 3.2 – Chemical representation of bisulfite conversion reaction of genomic DNA.

3.4.2 PCR amplification of bisulfite treated DNA

The repeat and its flanking region have approximately 1939 bp and contains 13 CpG dinucleotides. However, fragments amplified by BS-seq specific primers should not have more than 300 bp due to the high fragmentation of converted DNA. Given that, the repeat and its flanking region were divided into several fragments.

Designing primers against the region of interest is considered a critical step in obtaining adequate DNA methylation results. As mentioned before, after the bisulfite treatment all the unmethylated cytosines are transformed into uracils. Given that, we cannot design primers considering the pre-treated genomic DNA sequence (the original) as template, because they will not be 100% complementary. There are several rules that must be followed in order to design functional primers for bisulfite treated DNA amplification. First, the size of the primer should be between 26 and 30, to contend the loss of all cytosines; Second, the size of the amplicon should

be between 150-300 because the bisulfite converted DNA is highly fragmented, which difficult the amplification of large fragments. Third, cytosines at CpG dinucleotides within the primer sequence should be avoided, because each target bisulfite-treated DNA sequence can have a T (if unmethylated) or a C (if methylated), leading to a mismatch at the primer binding. If not possible to avoid, they should be placed in the 5' end and substituted by a mixed base at that position (Forward Primer: **Y** = C/T; Reverse Primer: **R** = G/A), in order to allow the complementarity with either methylated or unmethylated cytosines. The bisulfite specific primers in this work were design using MethPrimer [116] or MethylPrimer Software ® (Thermo Fisher).

Distribution of CpGs and the different sets of primers are illustrated in Figure 3.1. The CTCF-BS 1 contains two CpG dinucleotides, CpG island has five CpG dinucleotides, the region between CpG island and CTCF-BS 2 has one CpG dinucleotide and CTCF-BS 2 contains four CpG dinucleotides.

PCR reaction mixtures for bisulfite-sequencing of each region are specified in Appendix H and the primers and its respective sequence in Appendix B. For each PCR product, the clean-up was performed as specified in Section 3.2.2.

Due to the intronic nature of the region under study and also the degradation and high fragmentation of bisulfite treated DNA (huge reduction of DNA sequence complexity and quality decrease), the amplification and respective sequencing of the sub-regions specified above had to bypass some technical issues. Therefore, all the optimization steps for the PCR amplifications and sequencing reactions, for the four sub-regions, are available in Tables I.1, I.2, I.3 and I.4 at Appendix I.

3.4.3 Sanger sequencing of the bisulfite converted DNA

Sequencing analysis of the bisulfite converted DNA is performed with the same mixture and conditions specified in Section 3.3.4 and the primers used are detailed in the bisulfite sequencing table at the Appendix B. The sequencing template is purified using a Sephadex Column (detailed protocol in Appendix D) and the products were analyzed on a Capillary Electrophoresis Sequencer.

3.4.4 Analysis of the bisulfite converted sequences

The analysis of the sequences was performed with Epigenetic Sequencing Methylation Analysis Software (ESME) ®, based on the algorithm created by Lewin et al in 2004 [117]. ESME ® software is written in C++ and the current format available is a precompiled version for Debian/Linux 6.0.

There are a lot of challenges regarding the direct sequencing of bisulfite treated DNA (e.g., the sequence signal is weak when compared to genomic sequencing, the cytosine signal is normally

overscaled compared to the other bases, and it may have base caller artifacts interfering in the sequence analysis). Lewin et al 2004 developed an algorithm that is capable to overpass these obstacles and calculate a quantitative methylation information directly from the sequencing output (.abi files). Details of the ESME ® algorithm can be found at Appendix I.

To analyze the regions with less than 3 CpG dinucleotides, a modified version of ESME ® software was used (Specified in Appendix I).

ESME ® output files were graphically analyzed in R v.3.5.1. (R Core Team, 2018) and differences in methylation rate were accessed by Mann-Whitney-U tests.

4. RESULTS

4.1 Flanking $(ATTTT)_n$ variants potentially influencing repeat instability

To gain insight into the mechanisms of instability that originated the nonpathogenic unstable chromosomes followed by mutant $(ATTTC)_n$ alleles, I analyzed the sequencing data resulting from previously performed NGS of the SCA37 candidate region [17]. To confirm the previously obtained sequencing data, I sequenced, by Sanger sequencing, the $(ATTTT)_n$ flanking region in additional individuals. After comparing the repeat and its flanking region with the reference genome (according to GRCh37/hg19), I identified the presence of 9 SNPs, being one of them located upstream the repeat and the remaining eight downstream the repeat. I then analyzed these SNPs in three different categories of alleles: normal alleles with less than 50 repeats (n=23), large nonpathogenic with more than 50 repeat units and interrupted alleles (n=5) and mutant SCA37 alleles (n=8) from the three families. The results are summarized in Table 4.1 and the extended version of the table is at Appendix K.

Table 4.1 – SNPs identified flanking the *DAB1* $(ATTTT)_n$ region.

SNP ID	Position according to $(ATTTT)_n$	Ancestral allele frequency (dbSNP)	Variant (MAF*) (dbSNP)	Variant frequency (%)		
				Normal alleles (n=23)	Large and interrupted alleles (n=5)	Mutant alleles (n=8)
SNP 1	Upstream	**G:** 0.941	**A:** 0.059	30.43	0.00	0.00
(ATTTT)n	0	$(ATTTT)_{15}$				
SNP 2	Downstream	**A:** 0.969	**G:** 0.031	4.35	0.00	0.00
SNP 3		**T:** 0.941	**A:** 0.059	26.09	0.00	0.00
SNP 4		**G:** 0.970	**A:** 0.030	4.35	0.00	0.00
SNP 5		**C:** 0.941	**T:** 0.059	26.09	0.00	0.00
SNP 6		**T:** 0.584	**C:** 0.416	60.87	100.00	100.00
SNP 7		**C:** 0.875	**T:** 0.125	4.35	100.00	100.00
SNP 8		**G:** 0.941	**A:** 0.060	26.09	0.00	0.00
SNP 9		**C:** 0.783	**T:** 0.217	8.70	100.00	100.00

*MAF: Minor allele frequency

In the Alu orientation, an *in silico* analysis previously performed in the UCSC genome browser allowed us to identify several genetic elements present at the repeat flanking region, namely a CpG island downstream the repeat and two putative CTCF-BS flanking the $(ATTTT)_n$ repeat (Figure 4.1). Observing the positions of the SNPs found, it is possible to verify that four SNPs are located at CpG dinucleotides in the reference sequence (SNP 5, SNP 7, SNP 8 and SNP 9), meaning that the individuals presenting these variants will have the respective CpG disrupted. Also,

SNP 6 is located at a TpG dinucleotide that, when the variant is present, is transformed into a new CpG dinucleotide. Furthermore, three SNPs are positioned in the CTCF-BS downstream the repeat (SNP 7, SNP 8 and SNP 9). All SNPs found are annotated in SNP databases (dbSNP) and SNP 6 and SNP 9 have previously been identified by Loureiro et al 2019 [55]. The minor allele frequency (MAF) values show that the majority of the SNPs are rare variants (Table 4.1), with MAF < 12.5%, except SNP 6 (MAF 41.6%) and SNP 9 (MAF 21.7%).

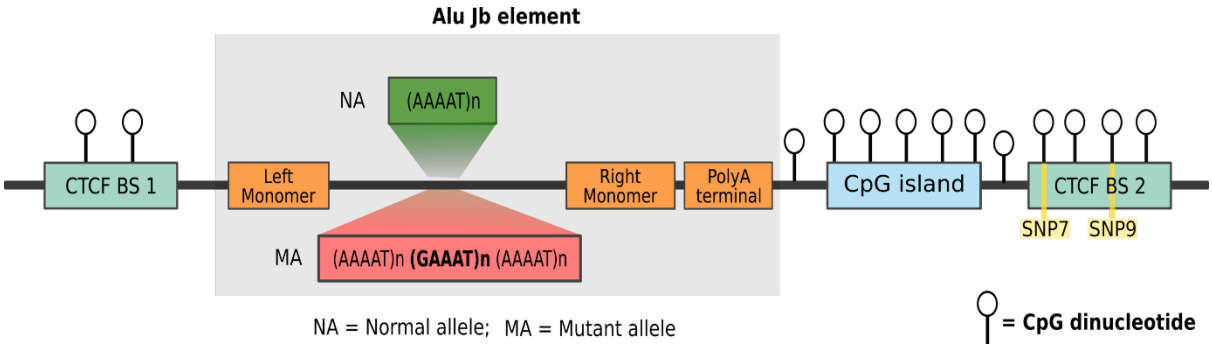

Figure 4.1 - Schematic representation of the genomic context of the repeat flanking regions in AluJb-oriented strand. The pathogenic $(ATTTC)_n$ repeat insertion is located in the middle poly A of an AluJb element. Downstream of the repeat is a CpG island and the repeat is flanked by two putative CTCF-binding sites (CTCF BS). CpG dinucleotides are represented by a white circle pins; SNP 7 and SNP 9 that abolish CpG10 and CpG12 are represented by yellow bars.

I calculated the relative frequencies of the variants in the three groups of alleles (Table 4.1). The results showed that six of the nine variants (SNP 1, SNP 2, SNP 3, SNP 4, SNP 5 and SNP 8) were found in normal alleles (with frequencies of 30.4%, 4.4%, 26.1%, 4.4%, 26.1% and 26.1%, respectively), but not in the large nonpathogenic or interrupted alleles or in mutant alleles. SNP 6 was found in 60.9% of the normal alleles and in all large and interrupted nonpathogenic alleles and mutant alleles. The frequency of these SNPs was not different for the groups of alleles studied as Fisher's exact tests showed no statistically significant differences in frequencies of these SNPs between normal vs large nonpathogenic and interrupted alleles or normal vs mutant alleles. However, there is a small number of alleles analyzed.

In normal alleles, SNP 7 had a frequency of 4.4% and SNP 9 of 8.7%, whereas all nonpathogenic large and interrupted alleles and mutant alleles presented both SNPs. The frequency of SNP 7 was higher in large nonpathogenic and interrupted than in normal alleles (Fisher's exact test, $p<0.0001$) and in mutant chromosomes compared with normal alleles (Fisher's exact test, $p<0.0001$). The SNP 9 showed also a higher frequency in nonpathogenic large and

interrupted alleles (Fisher's exact test, p<0.0001) and mutant alleles (Fisher's exact test, p<0.0001) compared with normal chromosomes.

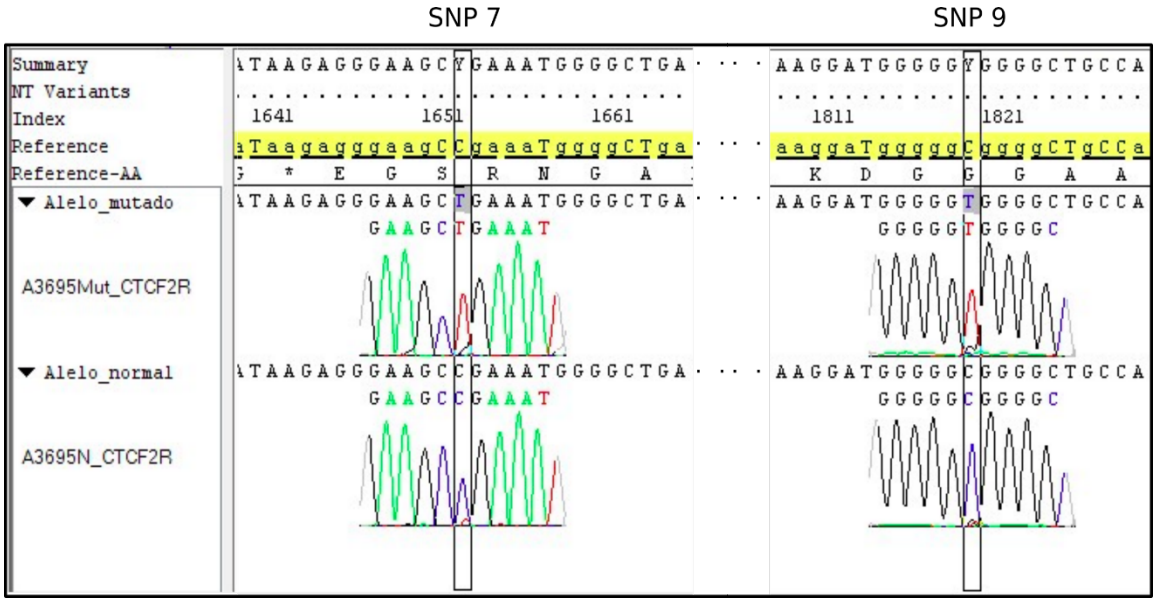

Figure 4.2 – SeqScape® output of an affected individual with both SNP 7 and SNP 9. Representative analysis of an affected individual, showing both alleles: at the top, the mutant allele having both SNP 7 and SNP 9; at the bottom, the normal allele with no differences comparing to the reference sequence (that is underlined in yellow).

4.2 Binding activities in CTCFs flanking the $(ATTTT)_n$

CTCFs can modulate genome tridimensional structure, being a contributing factor for repeat instability. Two SNPs, SNP7 and SNP9, are located in CpGs in putative CTCF-BS (Figure 4.1) and the variant identified in mutant chromosomes abolishes the CpG dinucleotides. To evaluate CTCF-binding at the repeat flanking regions, I searched across online databases for evidences of CTCF-binding at different cell types. First, I searched in CTCFBSDB2.0, a database that comprises a collection of experimentally determined and computationally predicted CTCF-BS from the literature [118]. Results show that CTCF effectively binds to the $(ATTTT)_n$ repeat flanking region in different cell types (Table 4.2). Results of ChiP-seq experiments for CTCF-binding at the repeat flanking region were available for Caco-2, HepG2, H1-hESC and fibroblast cell lines, resulting in different occupancy percentages, with all of them having a score for CTCF-binding above the average (average value = 122.98).

Table 4.2 - CTCF binding at the repeat flanking regions in different cell types.

Cell type (ID) [a]	Location	Score for CTCF binding [b]	Occupancy	Experiment	Source
Caco-2	chr1: 57832940-57833090	255	8%	ChiP-seq	wgEncodeEH000404
HepG2_13496	chr1:57833584-57834139	555	25%		wgEncodeEH000080
HepG2_2219	chr1:57833781-57833971	255	7%		wgEncodeEH001516
HepG2_98940	chr1: 57833820-57833970	255	25%		wgEncodeEH000401
HepG2_3498	chr1: 57833840-57833990	255	50%		wgEncodeEH000401
H1-hESC_3407	chr1: 57833869-57833923	255	20%		wgEncodeEH000560
Fibroblast_2175	chr1: 57833931-57833934	255	7%		wgEncodeEH001127

[a] **Caco-2**: immortalized cell line of human colorectal adenocarcinoma cells; **HepG2**: human liver cancer cell line; **H1-hESC**: Human embryonic stem cell line H1; **Fibroblast_2175**: fibroblast cell line. [b] Score for CTCF binding was accessed in UCSC genome browser and it ranges from 100 and 1000, with an average value of 122.98.

Subsequently, I performed a search in the Functional Annotation of Mammalian Genomes 5 (FANTOM5)[119]. This interface contains data from more than 1000 human and mouse samples, allowing to understand transcription regulation across different cells, giving the opportunity to search for active genes or transcription factors at different locations and biological contexts. Using ZENBU interactive interface to explore the available ChIP-seq data, I selected the region flanking the $(ATTTT)_n$ repeat at the *DAB1* gene and found ChIP-seq experiments performed in K562 (human immortalized myelogenous leukemia line) and HL-60 (established from a patient with acute myeloid leukemia) cell types. At the last release of ENCODE-TF-ChIP-seq (version 10 modified), it is reported CTCF binding with a quality score (q-value) superior to 3.4, being higher in K562 cell type than in HL-60 (Figure 0.1, Appendix M).

The ZENBU interface allow us to calculate the statistical enrichment of the TF binding by performing a Wilcoxon-Mann-Whitney rank-sum enrichment test, with all the TF binding data at a given region. The statistical analysis results demonstrate that there is an enrichment of CTCF-binding at the repeat flanking region both in K562 and HL-60, reflected on the Z-scores (Figure 0.2, Appendix M). A “high signal” or high enrichment of CTCF-binding is considered when Z-score > 1.64 [120].

These *in silico* results shows that CTCF effectively binds to the predicted location in those cell types. However, further experimental analysis needs to be performed in order to understand the CTCF-binding at neuronal cell types and, also, if there is differential binding throughout developmental stages.

4.3 SNPs predict disruption of TF binding

The results obtained above from the variant analysis shows that three of the SNPs found downstream the repeat are located within the predicted CTCF-BS. To investigate if those SNPs are predicted to disrupt CTCF-binding, I performed a search throughout several databases. The identification code of the three SNPs was loaded into SNP2TFBS (Mapping SNPs to Transcription Factor Binding Sites) [121], a web interface that allows to verify if a variant affects transcription factor binding in the human genome. The output results show that none of the three have been predicted to affect CTCF-binding, but SNP 9 is predicted to affect the ligation of several transcription factors, as SP2, SP1, EGR1, KLF4 and KLF5, based on the alteration of the PWM (position weight matrix) scores (Table 4.3). However, as we have few experimental data available for this specific region, it is necessary to conduct experimental assays to verify if those SNPs can disrupt those TF-binding sites and, more specifically, CTCF-BS.

Table 4.3 - Predicted binding of transcription factors affected by SNP 9.

TF name	PWM score on Ref [a]	PWM score on Alt [b]	Score difference [c]
SP2	1782	1494	-288
SP1	1751	1575	-176
EGR1	1496	1373	-123
KLF4	1472	1488	16
KLF5	1555	1551	-4

[a] Match score on reference genome (SNPs that change the PMW score above the higher threshold are retained) [b] Match score on the alternate genome (SNPs that change the PMW score above the higher threshold are retained) [c] PWM score difference between alternate and reference genome.

4.4 Changes in CpG methylation in SCA37 cells

Because two SNPs eliminated CpG dinucleotides and DNA methylation at their location, I analyzed the methylation profile of the region available in the UCSC database. The MeDIP-seq CpG Score (MCS), generated from postmortem human frontal cortex gray matter of a 57 year-old male [122], was evaluated for each CpG. MeDIP-seq uses immunoprecipitation to extract the methylated fraction of the genome. An antibody against 5-methylcytosine is used to immunoprecipitate methylated DNA that are then sequenced and mapped to the genome. Consequently, a higher MeDIP-seq score means a high methylation rate for the CpG. Within this region there are 13 CpG dinucleotides shown to be methylated with different MCS scores (Figure

4.3). The MeDIP-seq CpG Scores (MCS) at the **CpG island,** namely CpG 4 (MCS=22), CpG 5 (MCS=36), CpG 6 (MCS=39), CpG 7 (MCS=39) and CpG 8 (MCS=26)), are higher when compared to the CpG dinucleotides located at the **CTCF-BS-1,** CpG 1 (MCS=8) and CpG 2 (MCS=7), and **CTCF-BS-2,** CpG 10 (MCS=3), CpG 11 (MCS=2) CpG 12 (MCS=3) and CpG 13 (MCS=3) (Figure 4.3).

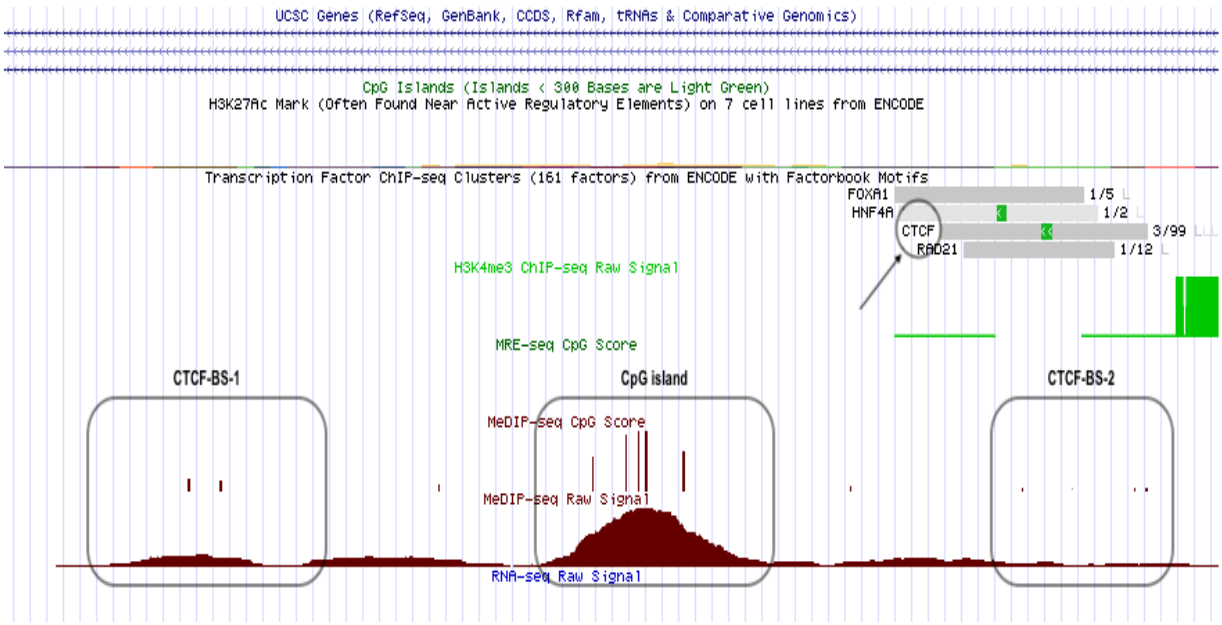

Figure 4.3 - Representation of the MeDIP-seq CpG Scores (MCS) of the 13 CpG dinucleotides at the repeat flanking region in UCSC genome browser. The observed window corresponds to the following tracks: ENCODE regulation, ENC chromatin, ENC DNA methylation, CpG islands and UCSF Brain methylation. The boxes highlight the methylation profile of the three main regions analyzed: CTCF binding site 1 (CTCF-BS 1), CpG island and CTCF binding site 2 (CTCF-BS 2). Above the boxes, it is possible to see the transcription factor binding site (TFBS) with the predicted binding of FOXA1, HNF1A, CTCF and RAD21.

The methylation profile of the *DAB1* $(ATTTT)_n$ repeat flanking region in brain cells shown above indicates that changes in DNA methylation of this region could influence binding of TFs or modify the architecture of the region. To investigate if the $(ATTTC)_n$ in SCA37 cells leads to changes in the methylation pattern, I used direct bisulfite sequencing of DNA from leukocytes (Figure 4.4), fibroblasts (Figure 4.5) and transgenic zebrafish with the $(ATTTC)_n$ and flanking region. The repeat flanking region was divided into four sub-regions with less than 300 bp for successful amplification of bisulfite treated DNA, and the sequences resulting from the amplification of each region were loaded into the program ESME ®, that calculates the methylation rate at each CpG dinucleotide, as specified in section 3.4.4 and Appendix J. The output of the analysis was divided into 11 different tables, each one containing the methylation rate in all CpGs for the respective samples, is available in Appendix L. In Figure 4.6 it is possible to visualize the methylation rate of each CpG at the respective sample, represented by a gradient of grays, in which the white color corresponds to 0% methylation and the black color to 100% methylation.

Table 4.4 – Methylation rate in blood cells for each CpG.

Region	CpG	Group (n)	Methylation Rate	
			Mean ± SD	p-value
CTCF BS 1	CpG 1	Normal (n=10)	0.851 ± 0.05	p-value = 0.036
		Affected (n=8)	0.725 ± 0.04	
	CpG 2	Normal (n=10)	0.737 ± 0.11	p-value = 0.016 *
		Affected (n=8)	0.850 ± 0.80	
CpG island	CpG 4	Normal (n=10)	0.896 ± 0.02	p-value = 0.471
		Affected (n=8)	0.910 ± 0.04	
	CpG 5	Normal (n=10)	0.848 ± 0.02	p-value = 0.010 *
		Affected (n=8)	0.884 ± 0.03	
	CpG 6	Normal (n=10)	0.682 ± 0.22	p-value = 0.077
		Affected (n=8)	0.913 ± 0.05	
	CpG 7	Normal (n=10)	0.914 ± 0.02	p-value = 0.750
		Affected (n=8)	0.916 ± 0.02	
	CpG 8	Normal (n=10)	0.717 ± 0.06	p-value = 0.560
		Affected (n=8)	0.689 ± 0.09	
CTCF BS 2	CpG 10	Normal (n=10)	0.534 ± 0.25	p-value = 0.045 *
		Affected (n=8)	0.240 ±0.20	
	CpG 11	Normal (n=10)	0.862 ± 0.06	p-value = 0.290
		Affected (n=8)	0.891 ±0.04	
	CpG 12	Normal (n=10)	0.520 ± 0.38	p-value = 0.210
		Affected (n=8)	0.301 ± 0.17	
	CpG 13	Normal (n=10)	0.638 ±0.44	p-value = 0.800
		Affected (n=8)	0.725 ± 0.04	

In peripheral blood, CpG 1 and CpG 2, located at CTCF-BS 1, are hypermethylated (methylation rate >50%) in both unaffected and affected individuals. Even so, for **CpG 2** (Table 4.4) affected individuals showed significantly higher methylation levels than unaffected individuals (Mann-Whitney-U test; $p < 0.05$) (Figure 4.4). For the CpGs located at the CpG island, no significant differences in methylation rate between unaffected and affected individuals have been found for CpG 4, CpG 6, CpG 7 and CpG 8. It is important to highlight that CpG 6 is overlapped by SNP 5. For CpG5, I found a significantly higher methylation rate in affected individuals compared with unaffected (Mann-Whitney-U test, $p < 0.010$). For the CpGs located at CTCF-BS 2, the methylation rate of CpG 10 is significantly higher in unaffected individuals than in affected individuals (Mann-Whitney-U test, $p < 0.05$). No significant differences in methylation rate have been found for CpG 11, CpG 12 and CpG 13. CpG 10 is overlapped by SNP 7, resulting in the elimination of this CpG in the alleles where SNP 7 is present (all affected individuals and individuals with large and

nonpathogenic alleles). Although there are no differences in methylation of CpG 12, this CpG is overlapped by SNP 9, resulting in no CpG present in the alleles with SNP 9 (all the affected individuals and individuals with large and nonpathogenic alleles). Consequently, I knew that the methylation rate obtained at CpG 10 and CpG 12 for the individuals carrying the SNP 7 and SNP 9, corresponds exclusively to the normal allele.

Concerning CpG 9, I could not assess methylation rate because, as described in Appendix J, the ESME® software performs a normalization of cytosine signal based on at least two CpG dinucleotides present in the sequence, but the fragment with CpG 9 had only one CpG dinucleotide (in less than 300 bp, as needed to guarantee successful bisulfite sequencing). However, with the analysis of the electropherograms from Sanger sequencing, it was possible to conclude that CpG 9 is methylated in both normal and affected individuals.

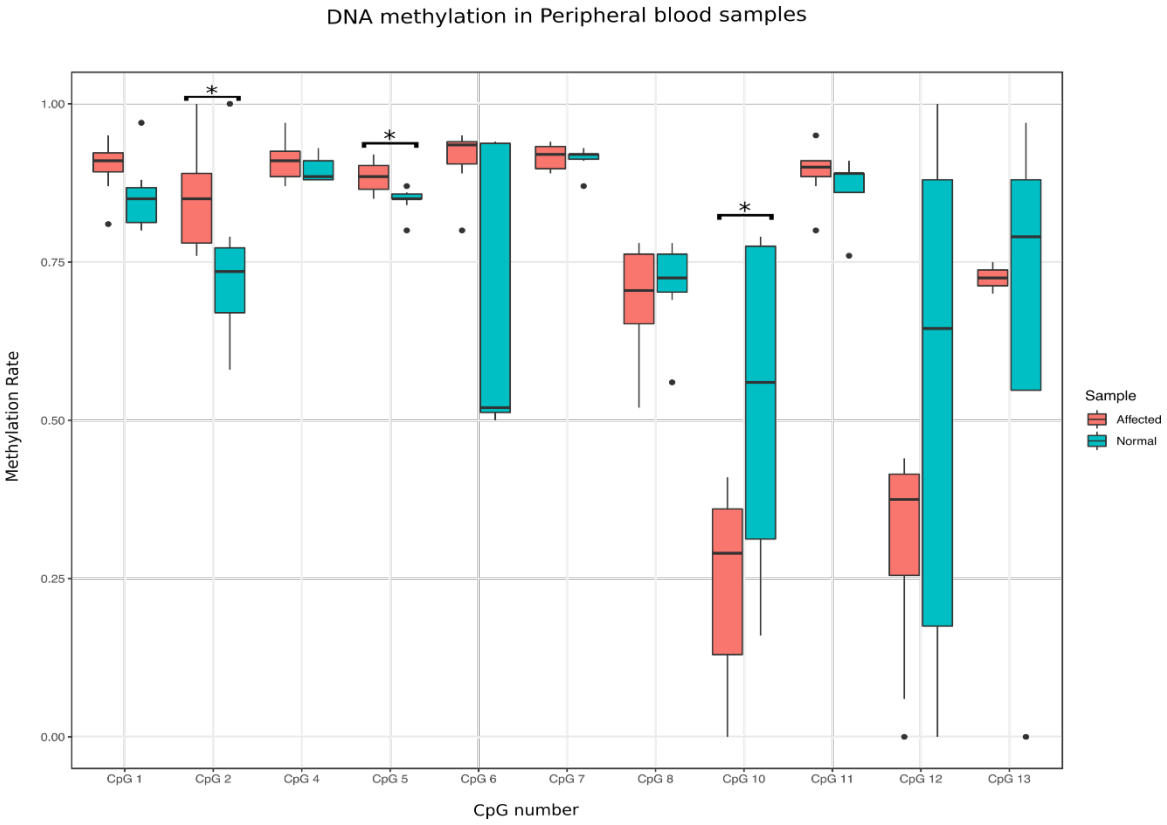

Figure 4.4 - Boxplot representation of the methylation rates in peripheral blood samples of normal and affected individuals, for CpGs located at the repeat and its flanking region. Non-parametric Mann-Whitney-U tests were performed to find statistically significant differences between normal and affected individuals. * represents a significance level of α=0.05 (CI=95%).

As referred in section 4.1, **SNP 6** (present in 69.6% of normal alleles, 100% of large nonpathogenic alleles and 100% of mutant alleles), creates a new CpG dinucleotide (TpG→CpG), located between the CpG island and the CTCF-BS 2 region. This new CpG is methylated in all the

affected individuals. For the same region specified above, it was not possible to calculate the methylation rate of this new CpG.

In fibroblast cell lines, for the CpGs studied there was no statistically significant differences in methylation rate between unaffected and affected cell lines. (Figure 4.5 and Table N.1 – Appendix N) probably due to the small number of cell lines examined.

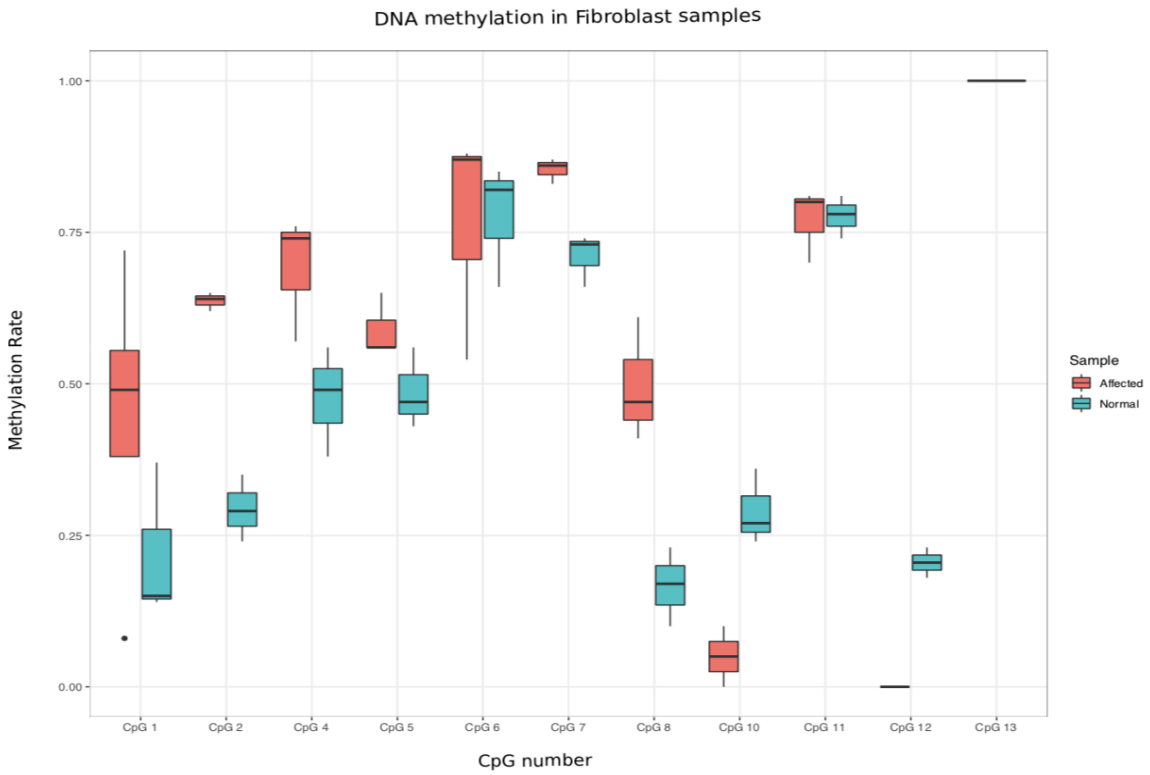

Figure 4.5 - Boxplot representation of the methylation rates in fibroblast cell lines with normal and affected phenotypes, for CpGs located at the repeat and its flanking region. Non-parametric Mann-Whitney-U tests were performed to find statistically significant differences between normal and affected fibroblast cell lines.

In zebrafish embryos no statistical analysis of the methylation rate is possible (n=1), however observing the heatmap (Figure 4.7) it is possible to visualize methylation rate differences for example, CpG 1 and CpG 2, where the transgenic zebrafish are hypermethylated comparing to unaffected zebrafish embryos or at CpG 10 and CpG 12, where the unaffected zebrafish embryos are hypermethylated compared to the affected zebrafish.

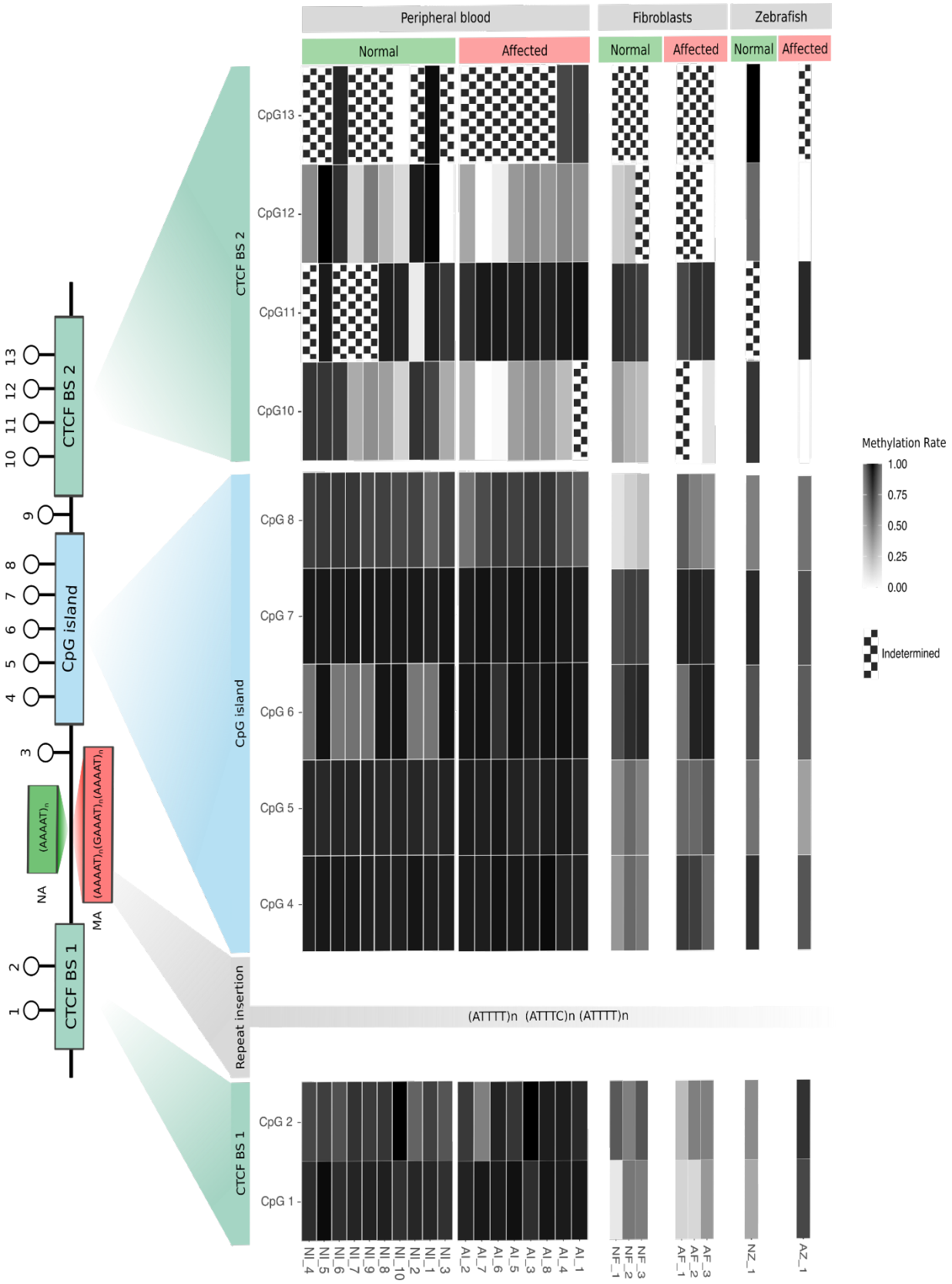

Figure 4.6 - Heatmap representation of the methylation rate of a) Peripheral blood samples; b) Fibroblast cell lines and c) zebrafish embryos, at the eleven CpGs analyzed. The CpGs are located at the repeat flanking region (numbered from 1 to 13 at the left panel scheme). For each tissue type, the green section corresponds to unaffected individuals (controls) and the red section corresponds to the affected individuals. The gradient of greys represents the methylation rate at each CpG, with the white color corresponding to 0% methylation and the black color corresponding to 100% methylation.

5. DISCUSSION

This master thesis allowed the identification of two SNPs, SNP7 and SNP9, in a potential CTCF-binding region closely adjacent to the *DAB1* $(ATTTC)_n$ repeat, being part of the SCA37 haplotype. These SNPs are only present in a small number of normal alleles, whereas all nonpathogenic large and interrupted chromosomes studied carried these SNPs. Thus, SNP 7 and SNP 9 together could be implicated in the mechanism of *DAB1* $(ATTTT)_n$ repeat instability, contributing to the generation of large alleles prone to nucleotide substitutions.

These two SNPs are both located at CpG dinucleotides in CTCF-BS 2, causing the elimination of both CpGs. This CTCF-BS shows an occupancy score for CTCF-binding above the average in several cancer and embryonic human cells, based in the CTCFBSDB2.0 database. ChIP-seq experiments from ENCODE-TF-ChIP-seq also show an enrichment of CTCF-binding at this CTCF-BS in at least two cancer cell types. These bioinformatic analyses suggest that this region may, in fact, be a CTCF-BS in neuronal cells and more precisely in cerebellar neurons, but experimental work is needed to confirm this.

CTCF-BS 2 could potentially be disrupted either by methylation or by SNP variants. In unaffected chromosomes the methylation score in public databases for CpG 10 and CpG 12, where SNP 7 and SNP 9 are located, is low. On the other hand, our results show that CpG 10 and CpG 12 in unaffected chromosomes are hypermethylated, while in affected chromosomes those CpGs are disrupted by SNP 7 and SNP 9 and, consequently, the methylation rate is low. This suggests a potential role for *DAB1* $(ATTTT)_n$ flanking CTCF-BS in instability.

STRs like the $(ATTTT)_n$ play an important role in our genome, namely by interfering with binding of transcription factors, causing alterations in chromatin 3D organization or interfering with epigenetic regulation [123,124]. Our bioinformatic search detected cis-regulatory elements flanking the $(ATTTT)_n$ region, similar to repeat regions in other repeat diseases like DM1[40,103] and SCA7[79]. These genetic elements contribute to repeat instability in that diseases.

The results obtained with this work revealed important genetic characteristics of the region, with the discovery of a total of nine different SNPs in a short region of about 1.3 kb containing the *DAB1* $(ATTTT)_n$. First of all, it served to highlight the already suspected hypothesis that the repeat and its flanking region is highly polymorphic [55]. Considering the relative frequencies of the nine genetic variants present at the region in study, it is possible to see that the majority of the variants occur in normal alleles with a high frequency. However, SNP 6, SNP 7 and SNP 9 are present in 100% of the unaffected individuals with large nonpathogenic or interrupted alleles, as well as in 100% affected individual mutant alleles, while its presence in the normal alleles is rare.

SNP 7 and SNP 9 are transmitted with all nonpathogenic large and interrupted alleles and mutant alleles studied, raising the hypothesis of being involved in repeat size and nucleotide instability. Small stable normal alleles in our study rarely carry SNP 7 or SNP 9. Following these SNPs and the number of repeat units at each allele, it is possible to observe a pattern. For example, individual NI_1 does not have SNP 7 or SNP 9 and its alleles have a low number of ATTTT repeat

units (RU). Individual NI_4 has SNP 9, with a slight increase in the ATTTT RU at that allele. Interestingly, the unaffected individual NI_7 has the combination of SNP 7 and SNP 9 and the ATTTT RU increased 10 units. This hypothesis is strengthened by the presence of these two SNPs at nonpathogenic large and interrupted alleles that have 51-120 and 50 ATTTT RU, respectively (Figure 5.1). In affected individuals, these two SNP variants are present in all mutant alleles and the number of ATTTT RU are >100 (Figure 5.1). However, this need to be confirmed in a larger number of alleles.

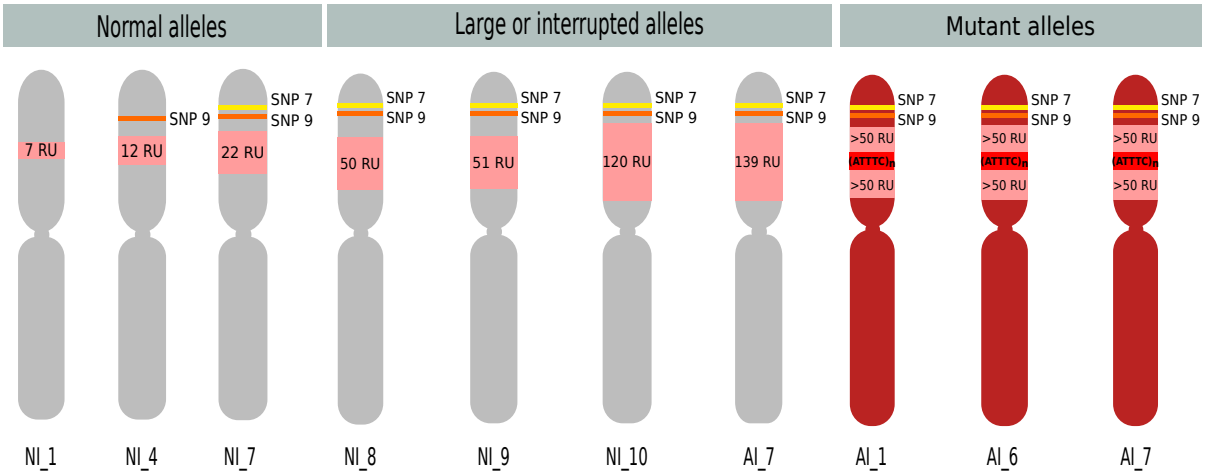

Figure 5.1 – Schematic representation of the observed pattern of repeat size associated with SNP 7 and SNP 9. The normal individual NI_1 does not carry SNP 7 or SNP 9, and the number of ATTTT repeat units (RU) is low. The normal individual NI_4 has SNP 9 and the ATTTT RU increased to 12. The normal individual NI_7 carries the two variants (SNP 7 and SNP 9) and the number of RU increased for 22. For the normal individuals NI_8, NI_9 and NI_10 and the affected individual AI_7, their large or interrupted alleles have the two variants (SNP 7 and SNP 9), and the number of ATTTT RU increased for 50, 51, 120 and 139, respectively. The two variants are also transmitted together in all the mutant alleles, in which the number of ATTTT RU is >100.

This interesting association can lead us to the assumption that the combined transmission of these two SNPs might contribute to a significant bias towards expansion of the $(ATTTT)_n$ repeat, contributing also to the allelic instability of the repeat and its flanking region. This increase in repeat size instability probably originated the ideal conditions for the occurrence of the nucleotide substitution (T→C) and posterior expansion of the ATTTC motif, that led to the formation of the pathogenic insertion responsible for SCA37 [55]. Additionally, we know that SNP 9 is present in all the large alleles studied from the Portuguese population, so it would be important to confirm if the same occurs in large alleles from other populations. SNP 7 and SNP 9, together with other genetic elements present at the repeat flanking region, might act as regulatory cis-elements and predisposing factors for repeat instability. For a better evaluation of the relationship between these two SNPs and repeat instability, it is necessary to increase the number of alleles analyzed to corroborate the association between those two SNPs and an increase in the repeat instability.

Why these SNPs can have influence in such genetic instability? There are several hypotheses justifying this instability:

(a) These SNP variants might lead to alterations in chromatin structure, by inhibiting CTCF binding [79]. We have reasons to think that CTCF effectively binds to the predictive region, at least in some cell types. In Table 4.2, it is possible to observe ChIP-seq results for CTCF binding in colorectal adenocarcinoma (Caco-2), human liver (HepG2), human embryonic stem cells (hESC) and fibroblast cell lines, with different occupancy levels. Additionally, ChIP-seq results in K562 (myelogenous leukemia) and HL-60 (acute myeloid leukemia) show a statistically significant enrichment in CTCF binding at the repeat flanking region. Although we don't have any experimental assay to prove CTCF-binding to the predictive regions in cerebellar cells, the ENCODE ChIP-Seq results increase the strength of that hypothesis. However, it is necessary to experimentally verify CTCF-binding at the repeat flanking region in these cells. Considering that CTCF effectively binds to the predictive sites, the occurrence of SNP 7 and SNP 9 might interfere with CTCF recognition of its binding site and, consequently, inhibit its binding. There are previous studies reporting the abrogation of CTCF-binding caused by point mutations at its binding site [79,125,126].

The role for CTCF in repeat instability has previously been described in SCA7, an autosomal dominant cerebellar ataxia caused by a CAG repeat expansion in the *ataxin-7* (*ATXN7*) gene, characterized by a marked genetic instability and a strong anticipation. A study conducted by Libby et al., 2008 [79] demonstrated that when the CTCF-BS is mutated inhibiting CTCF-binding, there is an increase in repeat instability both in germline and in somatic tissues. Additionally, they reported that CpG methylation at CTCF-BS can inhibit CTCF-binding and, consequently, lead to destabilization of the repeat expansion. Thus, the inhibition of CTCF-binding in regions proximal to the repeat promotes the expansion in size of repeat alleles, meaning that the occurrence of these chromatin rearrangements might be an important contributing factor for repeat instability. Therefore, similar to SCA7, it is possible that the SNPs found at the putative CTCF-BS flanking the SCA37 repeat inhibit CTCF-binding, leading to chromatin rearrangements that might contribute to SCA37 repeat instability.

The contribution to repeat instability of chromatin organization, mediated by CTCF-BS, is not perfectly clear in FXS, though it has been speculated that alteration in chromatin conformation caused by expanded repeats might contribute to repeat size instability [127]. For example, in FXS, the CGG repeat size at the *FMR1* gene is associated with heterochromatin formation, decrease in CTCF binding and disruption of topological-associated domains (TADs)[127,128], increasing its repeat instability.

Given the involvement of CTCF-BS in other repeat diseases, it would be important to investigate the role of CTCF-BS at the SCA37 locus and, specially, its contribution for repeat instability.

(b) Those genetic variants might affect transcription factor binding [103]. There is a TFBS downstream the repeat, colocalizing with SNP 7 and SNP 9. Thus, after running an *in silico*

analysis for understanding the interaction of SNP 7 and SNP 9 with transcription factor binding (Section 4.3), I verified that SNP 9 is reported as influencing the binding of several TFs in different cell types. SP1, one of the TFs whose binding might be affected by SNP 9, plays an important role in recruiting TATA-binding protein for promoters without TATA-box [90], allowing the transcription to occur. If SNP 9 inhibits SP1 binding, TATA-binding protein will not be recruited to the putative promoter, interfering with transcription. On the other hand, RAD21, a subunit of the cohesion complex that colocalizes with CTCF, is also predicted to bind this TFBS location. Transcription of repeats are normally associated with instability[103], so the presence of SNPs at TFBS affecting its binding might somehow contribute to this instability.

(c) SNPs might be located at binding sites of repair proteins. If repair proteins are no longer able to bind to that locations, transcription errors or DNA breaks will not be correctly repaired, leading consequently to an increase in instability. This mechanism has already been reported in other repeat diseases such as DM1, MJD/SCA3 and FXS [129]. However, with the *in silico* analysis made at the region in public databases, we have only evidence for binding of RAD21, which is a protein also involved in DNA double stand break repair.

(d) SNPs are located at CpG dinucleotides, leading to its disruption. The disruption of those CpGs located at CTCF-BS 2 lead to an alteration of the methylation pattern. DNA methylation has been reported as having a regulatory role in several repeat diseases [18,92,93,97,100,101,130], and the alteration of the methylation pattern can have influence at the transcription level, increasing repeat instability.

There are evidences sustaining that approximately half of the CpG islands (CGIs) are associated with annotated transcription start sites (TSS) and the majority of them are close to promoters in the human genome [90]. Consequently, DNA methylation at CGIs is an important regulatory factor for the repeat flanking region and it can influence proximal promoter activity. CGIs are generally unmethylated in a heavily methylated genome, but some CGIs can become methylated during normal development, resulting in stabilization of proximal promoter silencing [131,132]. This silencing can occur due to DNA methylation blocking of transcription factor binding or mediated by methyl-binding domain proteins that recruit chromatin remodeling proteins that cause DNA changes inhibiting promoter activity [70].

In FXS, the expanded CGG repeat leads to a hypermethylation of the CGG repeat itself and the CGI located within the promoter. This hypermethylation leads to heterochromatin formation at *FMR1* promoter, resulting in the silencing of the gene and consequently loss of FMRP protein production [92,108]. Interestingly, there are rare cases in which individuals having more than 200 CGG repeat units but an unmethylated CGI, presents a normal phenotype [133].

Regarding technical challenges, I found difficult assessing the methylation rate for the CpGs located at CTCF-BS 2, especially for CpG 11 and CpG 13, due to the repeat complexity of

this intronic region. Beyond the intronic nature intrinsic at this region, which makes the PCR amplification and Sanger sequencing difficult, the template is highly fragmented by the bisulfite treatment. Thus, even after many attempts at optimization (Appendix I), it was not possible to obtain sequences with sufficient quality to calculate the percentage of methylation for some individuals at these CpGs.

Observing CpG 10 and CpG 12, the lower methylation rate seems to be a direct consequence of the presence of SNP 7 and SNP 9. If looking closer to the heatmap (Figure 4.6), almost all individuals with less than 50% methylation have the same haplotype (with the presence of SNP 7 and SNP 9). More specifically, individuals AI_6 and AI_7 are homozygous for SNP 7 and SNP 9 (Table K.1 at Appendix K) and, consequently, their methylation rate at CpG 10 and CpG 12 is 0%.

Interestingly, it has already been reported some studies of SNPs at CTCF-BS abrogating CTCF-binding in a haplotype-dependent manner and, consequently, originate preferential CpG methylation of the unoccupied allele [112,126]. This seems not to be our case, once the methylation profile at the remaining CpGs of the region does not significantly vary between individuals with and without SNP 7 and SNP 9. However, to assess allele-specific methylation (ASM) at the SCA37 locus, further studies applying different methods have to be conducted, as the bisulfite conversion method degrades DNA into short fragments, making it impossible to design primers for long amplicons with sufficient CpGs and variants to discriminate the allele.

The investigation of the methylation pattern at three different sample types gives us a general overview of methylation status of this region in these cells. However, it is widely known that DNA methylation is cell- and/or tissue-specific [134,135]. This can be corroborated by the difference found between our results (in peripheral blood samples) and the methylation score (MCS) calculated in brain autopsies. Consequently, the methylation pattern results obtained for peripheral blood and fibroblast samples, do not correspond to what happens specifically in the Purkinje cells of the cerebellum, where the pathogenic mechanisms of SCA37 is crucial. These results are extremely important to understand the overall picture of the epigenetic alterations at the region under study, but it cannot be used for predicting what happens in the cerebellum. The ideal scenario would be performing the methylation analysis at Purkinje cells and/or other affected cell types at cerebellum. Unfortunately, we all know how difficult it is to obtain post-mortem brain tissue to be able to conduct such analysis.

The DNA methylation analysis was conducted by direct-bisulfite sequencing. The critical part of bisulfite sequencing is the analysis of the converted sequences because, when extracting genomic DNA from a tissue (e.g., peripheral blood), that DNA sample will be a mixture of different cells, each one of them with a different methylation profile. For a correct quantification of the methylation rate at each CpG, it is necessary to calculate the proportion of methylated templates and the unmethylated templates. The most common method to perform the calculation at bisulfite converted sequences is the cloning-based bisulfite sequencing. With this method, PCR products are cloned into a vector and transformed into competent *E.coli* and the clones are further sequenced [114]. However, to obtain a sensitivity higher than the direct-bisulfite sequencing, it is

necessary to perform, at least, 6 sequencing reactions per amplicon and per individual, which makes this cloning-based method more expensive and time consuming, when comparing to direct-bisulfite sequencing analysis [136]. Given that, ESME® software became extremely useful to calculate the methylation rate at each CpG, applying all the corrections and normalizations needed[117].

Although the direct-bisulfite sequencing enables the analysis of DNA methylation rate in a cost-effective way, this method does not allow to distinguish between 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5-hmC). 5-hmC is the first oxidative product of TET-mediated 5-mC demethylation pathway[137]. In the past few years, it has been increasing the evidences for the importance of 5-hmC in gene regulation. 5-hmC was thought to act only as an intermediary of the demethylation process, but it is now considered a stable epigenetic mark associated with transcription regulation [138]. Comparative analysis indicates that the level of 5-hmC in the genome is lower than 5-mC level. However, its abundance in brain tissue and the special enrichment in Purkinje cells and granular cells at the cerebellum [139] are a strong indicative of the important role that 5-hmC plays at several neurodegenerative disorders [140,141]. There are recent studies reporting the possible regulatory role of 5-hmC in repeat expansion diseases, like FXS [142] and FRDA[143]. Given that, perhaps it would be interesting to investigate the presence of 5-hmC at the repeat flanking region in brain tissue.

Furthermore, it is known that in mammals, the methylation occurs almost exclusively in the CpG dinucleotides. However, it is also possible that the methylation occurs in non-CpG context (CpH, being H = A, T or C) [144–146] , but it has recently been proven that non-CpG methylation might be the result of potentially erroneous or nonspecific methylation from the methylation machinery, that was supposed to act at neighboring CpGs [147]. For that reason, in this study only the DNA methylation at CpG dinucleotides was considered.

In summary, this work served to identify, in first place, several nucleotide variants present at the repeat flanking region. Interestingly, two of the variants found are located at predicted CTCF-BS and, consequently, could disrupt CTCF-binding, leading to chromatin rearrangements associated with repeat instability. Also, the co-transmission of those two variants is associated to an increase of the ATTTT repeat units, what led us to hypothesized that this event contributes to the repeat instability that helped to create the ideal conditions for the occurrence of the dynamic mutation responsible for the $(ATTTC)_n$ repeat insertion formation.

The second part of the work served to analyze the methylation pattern of the repeat flanking region. It was observed some significant differences between unaffected individuals and affected individuals and, also, that the CGI located downstream the repeat is hypermethylated, suggesting that the DNA methylation plays an important regulatory role at this region.

6. FUTURE AVENUES

Regarding the influence of the haplotype in the CTCF-binding raised by this work, it is necessary to conduct experimental assays to confirm CTCF-binding at the two CTCF-BS flanking the repeat. This can be performed throughout chromatin-immunoprecipitation using antibodies against CTCF, with further sequencing (ChiP-seq). However, in order to obtain more information with a single experiment, it could be conducted a chromosome conformation capture-on-chip (4C). 4C can give us the contact profiles of the repeat and its flanking region, e.g., between promoters and enhancers, and also the capture of the conformational changes possibly caused by CTCF insulation[148]. Additionally, it can allow us to understand if the repeat interacts with other regions in the genome, resulting in a general view of the conformational regulatory changes occurring at this highly unstable region. Additionally, the predicted disruption of TF binding by SNP 9 (shown in section 4.3) should also be experimentally proved, by comparing ChIP-seq results between cells from individuals with and without SNP 9.

Later, in order to understand if the hypermethylation of the CGI influences the activity of a close promoter, we should perform an expression study, in which the promoter activity would be measured (e.g., by a vector-based promoter assay) in sequences with hypermethylated CGI compared with promoter activity in sequences with unmethylated CGI. Different levels of expression would imply that CGI methylation effectively plays a regulatory role. These experimental studies must be conducted in cerebellar tissue, where the repeat is effectively expressed.

The same principle should be applied for the DNA methylation analysis. As discussed before, DNA methylation is cell-specific and, consequently, the DNA methylation pattern found in peripheral blood samples, fibroblasts or zebrafish embryos cannot be extrapolated for affected cells of the cerebellum. Given that, it would be important to perform this analysis in human cerebellar samples, the most affected tissue in SCA37 patients. To conduct such experiments, it is necessary to obtain cerebellar samples from SCA37 affected individuals for further analysis by bisulfite sequencing. The ideal scenario would be the DNA methylation analysis specifically at Purkinje and granular cells from the cerebellum. This analysis could be achieved by single-cell bisulfite sequencing, a recent molecular method that allows the detection of the methylated cytosines in genomic DNA from single cells [149,150], after the isolation of Purkinje and/or granular cells from cerebellar slices by, for example, laser capture microdissection (LCM) [151–153]. Unfortunately, beyond the high difficulty in obtaining brain samples from SCA37 affected individuals, this approach requires expensive materials, very sophisticated equipment and highly qualified training.

I have highlighted the importance of DNA methylation for gene regulation, but it is not the only epigenetic mark with regulatory roles, and/or capable of causing chromatin rearrangements responsible for repeat instability. There are several studies suggesting that the interaction between DNA methylation and histone modifications is very important for chromatin dynamics [154]. Therefore, it would be interesting to conduct ChiP-seq experiments in order to understand if the presence of other epigenetic marks (e.g., histone methylation – normally associated to gene silencing – or histone acetylation – present at euchromatin regions [155]) contribute to the high instability of this

repeat region. Additionally, the presence of 5-hmC should also be investigated. Although the gold standard bisulfite conversion techniques are not capable of distinguish between 5-mC and 5-hmC there are several methods allowing that distinction as, for example, the antibody-based hydroxymethylated DNA immunoprecipitation following by sequencing (hMeDIP-seq) [156,157] or single nucleotide 5-hmC mapping approaches like whole-genome oxidative bisulfite in combination with conventional bisulfite sequencing (WG-Bis/ox-Bis-seq)[158] or TET-assisted bisulfite sequencing (TAB-seq)[159].

Many advances have been made in recent years to understand the pathogenic molecular mechanisms responsible for SCA37. However, the path to a better knowledge of the molecular mechanisms contributing to SCA37 disease is still very long. Therefore, the hard work carried so far must resume.

7. WEB RESOURCES

Primer 3 Software (v 0.4.0)	https://bioinfo.ut.ee/primer3-0.4.0/
MethPrimer (v 2.0)	http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi
1000 Genomes	http://www.1000genomes.org/
UCSC Genome Browser	http://genome.ucsc.edu/
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
SNP2TFBS	https://ccg.epfl.ch/snp2tfbs/
CTCFBSDB 2.0	http://insulatordb.uthsc.edu
ZENBU	https://fantom.gsc.riken.jp/zenbu/gLyphs/
ENCODE (SCREEN)	https://screen.wenglab.org
R Project	https://www.r-project.org/index.html

8. REFERENCES

1. Strachan, T. & Read, A. *Human molecular genetics*. (New York : Garland Science/Taylor & Francis Group, c2011., 2011).
2. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
3. Rohilla, K. J. & Gagnon, K. T. RNA biology of disease-associated microsatellite repeat expansions. *Acta Neuropathol. Commun.* **5**, 63 (2017).
4. Gemayel, R., Vinces, M. D., Legendre, M. & Verstrepen, K. J. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
5. Polak, U., McIvor, E., Dent, S. Y. R., Wells, R. D. & Napierala, M. Expanded complexity of unstable repeat diseases. *BioFactors* **39**, 164–175 (2013).
6. Heitz, D. *et al.* Isolation of sequences that span the fragile X and identification of a fragile X-related CpG island. *Science* **251**, 1236–1239 (1991).
7. Spada, A. R. L., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
8. Paulson, H. Repeat expansion diseases. in *Handbook of Clinical Neurology* vol. 147 105–123 (Elsevier, 2018).
9. Macdonald, M. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
10. David, G. Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). *Hum. Mol. Genet.* **7**, 165–170 (1998).
11. De Baere, E. *et al.* FOXL2 and BPES: Mutational Hotspots, Phenotypic Variability, and Revision of the Genotype-Phenotype Correlation. *Am. J. Hum. Genet.* **72**, 478–487 (2003).
12. Brais, B. *et al.* Short GCG expansions in the *PABP2* gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.* **18**, 164–167 (1998).
13. Barasoain, M. *et al.* Study of the Genetic Etiology of Primary Ovarian Insufficiency: *FMR1* Gene. *Genes* **7**, 123 (2016).
14. Holmes, S. E. *et al.* Expansion of a novel CAG trinucleotide repeat in the 5′ region of *PPP2R2B* is associated with SCA12. *Nat. Genet.* **23**, 391–392 (1999).
15. Mahadevan, M. *et al.* Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255**, 1253–1255 (1992).
16. Liquori, C. L. Myotonic Dystrophy Type 2 Caused by a CCTG Expansion in Intron 1 of *ZNF9*. *Science* **293**, 864–867 (2001).
17. Seixas, A. I. *et al.* A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of *DAB1* , Mapping to SCA37 , Causes Spinocerebellar Ataxia. *Am. J. Hum. Genet.* **101**, 87–103 (2017).
18. LaCroix, A. J. *et al.* GGC Repeat Expansion and Exon 1 Methylation of *XYLT1* Is a Common Pathogenic Variant in Baratela-Scott Syndrome. *Am. J. Hum. Genet.* **104**, 35–44 (2019).
19. Orr, H. T. & Zoghbi, H. Y. Trinucleotide Repeat Disorders. *Annu. Rev. Neurosci.* **30**, 575–621 (2007).
20. Khristich, A. N. & Mirkin, S. M. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295**, 4134–4170 (2020).
21. Sato, N. *et al.* Spinocerebellar Ataxia Type 31 Is Associated with "Inserted" Penta-Nucleotide Repeats Containing (TGGAA)n. *Am. J. Hum. Genet.* **85**, 544–557 (2009).
22. Cen, Z. *et al.* Intronic pentanucleotide TTTCA repeat insertion in the *SAMD12* gene causes familial cortical myoclonic tremor with epilepsy type 1. *Brain* **141**, 2280–2288 (2018).
23. Lei, X. X. *et al.* TTTCA repeat expansion causes familial cortical myoclonic tremor with epilepsy. *Eur. J. Neurol.* **26**, 513–518 (2019).
24. Florian, R. T. *et al.* Unstable TTTTA/TTTCA expansions in *MARCH6* are associated with Familial Adult Myoclonic Epilepsy type 3. *Nat. Commun.* **10**, 4919 (2019).

25. Corbett, M. A. *et al.* Intronic ATTTC repeat expansions in *STARD7* in familial adult myoclonic epilepsy linked to chromosome 2. *Nat. Commun.* **10**, 4920 (2019).

26. Yeetong, P. *et al.* TTTCA repeat insertions in an intron of *YEATS2* in benign adult familial myoclonic epilepsy type 4. *Brain* **142**, 3360–3366 (2019).

27. Brouwer, J. R., Willemsen, R. & Oostra, B. A. Microsatellite repeat instability and neurological disease. *BioEssays* **31**, 71–83 (2009).

28. Nelson, D. L., Orr, H. T. & Warren, S. T. The Unstable Repeats—Three Evolving Faces of Neurological Disease. *Neuron* **77**, 825–843 (2013).

29. Oberle, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097–1102 (1991).

30. Pieretti, M. *et al.* Absence of expression of the *FMR1* gene in fragile X syndrome. *Cell* **66**, 817–822 (1991).

31. Kumari, D., Biacsi, R. E. & Usdin, K. Repeat Expansion Affects Both Transcription Initiation and Elongation in Friedreich Ataxia Cells. *J. Biol. Chem.* **286**, 4209–4215 (2011).

32. Balendra, R. & Isaacs, A. M. *C9orf72*-mediated ALS and FTD: multiple pathways to disease. *Nat. Rev. Neurol.* **14**, 544–558 (2018).

33. Loureiro, J. R., Oliveira, C. L. & Silveira, I. Unstable repeat expansions in neurodegenerative diseases: nucleocytoplasmic transport emerges on the scene. *Neurobiol. Aging* **39**, 174–183 (2016).

34. Sznajder, Ł. J. & Swanson, M. S. Short Tandem Repeat Expansions and RNA-Mediated Pathogenesis in Myotonic Dystrophy. *Int. J. Mol. Sci.* **20**, 3365 (2019).

35. Sopher, B. L. *et al.* CTCF Regulates *Ataxin-7* Expression through Promotion of a Convergently Transcribed, Antisense Noncoding RNA. *Neuron* **70**, 1071–1084 (2011).

36. Ikeda, Y., Daughters, R. S. & Ranum, L. P. W. Bidirectional expression of the SCA8 expansion mutation: One mutation, two genes. *The Cerebellum* **7**, 150–158 (2008).

37. Chung, D. W., Rudnicki, D. D., Yu, L. & Margolis, R. L. A natural antisense transcript at the Huntington's disease repeat locus regulates *HTT* expression. *Hum. Mol. Genet.* **20**, 3467–3477 (2011).

38. Wilburn, B. *et al.* An Antisense CAG Repeat Transcript at *JPH3* Locus Mediates Expanded Polyglutamine Protein Toxicity in Huntington's Disease-like 2 Mice. *Neuron* **70**, 427–440 (2011).

39. Ladd, P. D. *et al.* An antisense transcript spanning the CGG repeat region of *FMR1* is upregulated in premutation carriers but silenced in full mutation individuals. *Hum. Mol. Genet.* **16**, 3174–3187 (2007).

40. Cho, D. H. *et al.* Antisense Transcription and Heterochromatin at the DM1 CTG Repeats Are Constrained by CTCF. *Mol. Cell* **20**, 483–489 (2005).

41. Mori, K. *et al.* Bidirectional transcripts of the expanded *C9orf72* hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. *Acta Neuropathol. (Berl.)* **126**, 881–893 (2013).

42. Lapidot, M. & Pilpel, Y. Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep.* **7**, 1216–1222 (2006).

43. Lavorgna, G. *et al.* In search of antisense. *Trends Biochem. Sci.* **29**, 88–94 (2004).

44. Morris, K. V., Santoso, S., Turner, A.-M., Pastori, C. & Hawkins, P. G. Bidirectional Transcription Directs Both Transcriptional Gene Activation and Suppression in Human Cells. *PLoS Genet.* **4**, e1000258 (2008).

45. Durr, A. Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. *Lancet Neurol.* **9**, 885–894 (2010).

46. Coutinho, P. *et al.* Hereditary Ataxia and Spastic Paraplegia in Portugal: A Population-Based Prevalence Study. *JAMA Neurol.* **70**, 746 (2013).

47. Sequeiros, J., Martins, S. & Silveira, I. Epidemiology and population genetics of degenerative ataxias. in *Handbook of Clinical Neurology* vol. 103 227–251 (Elsevier, 2012).

48. Paulson, H. L., Shakkottai, V. G., Clark, H. B. & Orr, H. T. Polyglutamine spinocerebellar ataxias — from genes to potential treatments. *Nat. Rev. Neurosci.* **18**, 613–626 (2017).

49. Taroni, F. & DiDonato, S. Pathways to motor incoordination: the inherited ataxias. *Nat. Rev. Neurosci.* **5**, 641–655 (2004).

50. Klockgether, T., Mariotti, C. & Paulson, H. L. Spinocerebellar ataxia. *Nat. Rev. Dis. Primer* **5**, 24 (2019).

51. Carlson, K. M., Andresen, J. M. & Orr, H. T. Emerging pathogenic pathways in the spinocerebellar

ataxias. *Curr. Opin. Genet. Dev.* **19**, 247–253 (2009).

52. Paulson, H. L. The Spinocerebellar Ataxias: *J. Neuroophthalmol.* **29**, 227–237 (2009).

53. Howell, B. W. Mouse disabled (*mDab1*): a Src binding protein implicated in neuronal development. *EMBO J.* **16**, 121–132 (1997).

54. Sheldon, M. *et al.* Scrambler and yotari disrupt the disabled gene and produce a reeler -like phenotype in mice. *Nature* **389**, 730–733 (1997).

55. Loureiro, J. R. *et al.* Mutational mechanism for *DAB1* $(ATTTC)_n$ insertion in SCA37: ATTTT repeat lengthening and nucleotide substitution. *Hum. Mutat.* **40**, 404–412 (2019).

56. Chauhan, C., Dash, D., Grover, D., Rajamani, J. & Mukerji, M. Origin and Instability of GAA Repeats: Insights from Alu Elements. *J. Biomol. Struct. Dyn.* **20**, 253–263 (2002).

57. Clark, R. M. *et al.* Expansion of GAA triplet repeats in the human genome: unique origin of the FRDA mutation at the center of an Alu. *Genomics* **83**, 373–383 (2004).

58. Kurosaki, T., Matsuura, T., Ohno, K. & Ueda, S. Alu-Mediated Acquisition of Unstable ATTCT Pentanucleotide Repeats in the Human *ATXN10* Gene. *Mol. Biol. Evol.* **26**, 2573–2579 (2009).

59. Ishiura, H. *et al.* Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* **50**, 581–590 (2018).

60. Serrano-Munuera, C. *et al.* New Subtype of Spinocerebellar Ataxia With Altered Vertical Eye Movements Mapping to Chromosome 1p32. *JAMA Neurol.* **70**, 764 (2013).

61. Corral-Juan, M. *et al.* Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. *Brain* **141**, 1981–1997 (2018).

62. Berger, S. L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. An operational definition of epigenetics. *Genes Dev.* **23**, 781–783 (2009).

63. Lee, J. Y. & Orr-Weaver, T. L. Chromatin. in *Encyclopedia of Genetics* 340–343 (Elsevier, 2001).

64. Weksberg, R. *et al.* Epigenetics. in *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics* 79–123 (Elsevier, 2019).

65. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).

66. Iurlaro, M., von Meyenn, F. & Reik, W. DNA methylation homeostasis in human and mouse development. *Curr. Opin. Genet. Dev.* **43**, 101–109 (2017).

67. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).

68. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).

69. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).

70. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).

71. Filippova, G. N. *et al.* An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian *c-myc* oncogenes. *Mol. Cell. Biol.* **16**, 2802–2813 (1996).

72. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485 (2000).

73. Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol. Cell* **66**, 711-720.e3 (2017).

74. Pearson, C. E., Nichol Edamura, K. & Cleary, J. D. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* **6**, 729–742 (2005).

75. Greene, E., Mahishi, L., Entezam, A., Kumari, D. & Usdin, K. Repeat-induced epigenetic changes in intron 1 of the *frataxin* gene and its consequences in Friedreich ataxia. *Nucleic Acids Res.* **35**, 3383–3390 (2007).

76. Lin, Y., Dion, V. & Wilson, J. H. Transcription promotes contraction of CAG repeat tracts in human cells. *Nat. Struct. Mol. Biol.* **13**, 179–180 (2006).

77. Brock, G. Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. *Hum. Mol. Genet.* **8**, 1061–1067 (1999).

78. Dion, V., Lin, Y., Hubert, L., Waterland, R. A. & Wilson, J. H. *Dnmt1* deficiency promotes CAG repeat expansion in the mouse germline. *Hum. Mol. Genet.* **17**, 1306–1317 (2008).

79. Libby, R. T. *et al.* CTCF cis-Regulates Trinucleotide Repeat Instability in an Epigenetic Manner: A Novel Basis for Mutational Hot Spot Determination. *PLoS Genet.* **4**, e1000257 (2008).

80. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9**, 2395–2402 (2000).

81. Cheng, X. Structure and Function of DNA Methyltransferases. *Annu. Rev. Biophys. Biomol. Struct.* **24**, 293–318 (1995).

82. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases DNM3A and DNMT3B Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99**, 247–257 (1999).

83. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).

84. Costello, J. F. Methylation matters. *J. Med. Genet.* **38**, 285–303 (2001).

85. Straussman, R. *et al.* Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.* **16**, 564–571 (2009).

86. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. DNA motifs associated with aberrant CpG island methylation. *Genomics* **87**, 572–579 (2006).

87. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**, 245–254 (2003).

88. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).

89. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).

90. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).

91. Naumann, A., Hochstein, N., Weber, S., Fanning, E. & Doerfler, W. A Distinct DNA-Methylation Boundary in the 5′- Upstream Sequence of the *FMR1* Promoter Binds Nuclear Proteins and Is Lost in Fragile X Syndrome. *Am. J. Hum. Genet.* **85**, 606–616 (2009).

92. Colak, D. *et al.* Promoter-Bound Trinucleotide Repeat mRNA Drives Epigenetic Silencing in Fragile X Syndrome. *Science* **343**, 1002–1005 (2014).

93. López Castel, A. *et al.* Expanded CTG repeat demarcates a boundary for abnormal CpG methylation in myotonic dystrophy patient tissues. *Hum. Mol. Genet.* **20**, 1–15 (2011).

94. Steinbach, P., Gläser, D., Vogel, W., Wolf, M. & Schwemmle, S. The *DMPK* Gene of Severely Affected Myotonic Dystrophy Patients Is Hypermethylated Proximal to the Largely Expanded CTG Repeat. *Am. J. Hum. Genet.* **62**, 278–285 (1998).

95. Al-Mahdawi, S. *et al.* The Friedreich ataxia GAA repeat expansion mutation induces comparable epigenetic changes in human and transgenic mouse brain and heart tissues. *Hum. Mol. Genet.* **17**, 735–746 (2007).

96. Castaldo, I. *et al.* DNA methylation in intron 1 of the *frataxin* gene is related to GAA repeat length and age of onset in Friedreich ataxia patients. *J. Med. Genet.* **45**, 808–812 (2008).

97. Liu, E. Y. *et al. C9orf72* hypermethylation protects against repeat expansion-associated pathology in ALS/FTD. *Acta Neuropathol. (Berl.)* **128**, 525–541 (2014).

98. Xi, Z. *et al.* Hypermethylation of the CpG Island Near the G4C2 Repeat in ALS with a *C9orf72* Expansion. *Am. J. Hum. Genet.* **92**, 981–989 (2013).

99. Xi, Z. *et al.* The *C9orf72* repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathol. (Berl.)* **129**, 715–727 (2015).

100. Himeda, C. L. & Jones, P. L. The Genetics and Epigenetics of Facioscapulohumeral Muscular Dystrophy. *Annu. Rev. Genomics Hum. Genet.* **20**, 265–291 (2019).

101. Salsi, V., Magdinier, F. & Tupler, R. Does DNA Methylation Matter in FSHD? *Genes* **11**, 258 (2020).

102. Filippova, G. N. *et al.* Boundaries between Chromosomal Domains of X Inactivation and Escape Bind CTCF and Lack CpG Methylation during Early Development. *Dev. Cell* **8**, 31–42 (2005).

103. Filippova, G. N. *et al.* CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat. Genet.* **28**, 335–343 (2001).

104. De Biase, I., Chutake, Y. K., Rindler, P. M. & Bidichandani, S. I. Epigenetic Silencing in Friedreich

Ataxia Is Associated with Depletion of CTCF (CCCTC-Binding Factor) and Antisense Transcription. *PLoS ONE* **4**, e7914 (2009).

105. Ottaviani, A. *et al.* The D4Z4 Macrosatellite Repeat Acts as a CTCF and A-Type Lamins-Dependent Insulator in Facio-Scapulo-Humeral Dystrophy. *PLoS Genet.* **5**, e1000394 (2009).

106. Coffee, B., Zhang, F., Warren, S. T. & Reines, D. Acetylated histones are associated with *FMR1* in normal but not fragile X-syndrome cells. *Nat. Genet.* **22**, 98–101 (1999).

107. Coffee, B., Zhang, F., Ceman, S., Warren, S. T. & Reines, D. Histone Modifications Depict an Aberrantly Heterochromatinized *FMR1* Gene in Fragile X Syndrome. *Am. J. Hum. Genet.* **71**, 923–932 (2002).

108. Avitzour, M. *et al. FMR1* Epigenetic Silencing Commonly Occurs in Undifferentiated Fragile X-Affected Embryonic Stem Cells. *Stem Cell Rep.* **3**, 699–706 (2014).

109. Alwazzan, M., Newman, E., Hamshere, M. G. & Brook, J. D. Myotonic Dystrophy Is Associated with a Reduced Level of RNA from the *DMWD* Allele Adjacent to the Expanded Repeat. *Hum. Mol. Genet.* **8**, 1491–1497 (1999).

110. Klesert, T. R., Otten, A. D., Bird, T. D. & Tapscott, S. J. Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of DMAHP. *Nat. Genet.* **16**, 402–406 (1997).

111. Paliwal, A. *et al.* Comparative Anatomy of Chromosomal Domains with Imprinted and Non-Imprinted Allele-Specific DNA Methylation. *PLoS Genet.* **9**, e1003622 (2013).

112. Tycko, B. Allele-specific DNA methylation: beyond imprinting. *Hum. Mol. Genet.* **19**, R210–R220 (2010).

113. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**, 5463–5467 (1977).

114. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.* **89**, 1827–1831 (1992).

115. Hajkova, P. *et al.* DNA-Methylation Analysis by the Bisulfite-Assisted Genomic Sequencing Method. in *DNA Methylation Protocols* vol. 200 143–154 (Humana Press, 2002).

116. Li, L.-C. & Dahiya, R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18**, 1427–1431 (2002).

117. Lewin, J., Schmitt, A. O., Adorjan, P., Hildmann, T. & Piepenbrock, C. Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics* **20**, 3005–3012 (2004).

118. Ziebarth, J. D., Bhattacharya, A. & Cui, Y. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res.* **41**, D188–D194 (2012).

119. the FANTOM consortium *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).

120. The ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

121. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).

122. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).

123. Borel, C. *et al.* Tandem repeat sequence variation as causative Cis-eQTLs for protein-coding gene expression variation: The case of CSTB. *Hum. Mutat.* **33**, 1302–1309 (2012).

124. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).

125. Do, C. *et al.* Genetic–epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biol.* **18**, 120 (2017).

126. Do, C. *et al.* Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. *Am. J. Hum. Genet.* **98**, 934–955 (2016).

127. Sun, J. H. *et al.* Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224-238.e15 (2018).

128. Buendía, G. A. *et al.* Three-dimensional chromatin interactions remain stable upon CAG/CTG repeat expansion. *Sci. Adv.* **6**, eaaz4012 (2020).

129. Martins, S. *et al.* Modifiers of (CAG)n instability in Machado–Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. *Hum. Genet.* **133**, 1311–1318 (2014).

130. Ishiura, H. *et al.* Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat. Genet.* **51**, 1222–1232 (2019).

131. Mohn, F. & Schübeler, D. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.* **25**, 129–136 (2009).

132. Mohn, F. *et al.* Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Mol. Cell* **30**, 755–766 (2008).

133. Peprah, E. Fragile X Syndrome: The *FMR1* CGG Repeat Distribution Among World Populations: *FMR1* Prevalence in World Populations. *Ann. Hum. Genet.* **76**, 178–191 (2012).

134. Edgar, R. D., Jones, M. J., Meaney, M. J., Turecki, G. & Kobor, M. S. BECon: a tool for interpreting DNA methylation findings from blood in the context of brain. *Transl. Psychiatry* **7**, e1187–e1187 (2017).

135. Farré, P. *et al.* Concordant and discordant DNA methylation signatures of aging in human blood and brain. *Epigenetics Chromatin* **8**, 19 (2015).

136. Hernández, H. G., Tse, M. Y., Pang, S. C., Arboleda, H. & Forero, D. A. Optimizing methodologies for PCR-based DNA methylation analysis. *BioTechniques* **55**, (2013).

137. Lin, I.-H., Chen, Y.-F. & Hsu, M.-T. Correlated 5-Hydroxymethylcytosine (5hmC) and Gene Expression Profiles Underpin Gene and Organ-Specific Epigenetic Regulation in Adult Mouse Brain and Liver. *PloS One* **12**, e0170779 (2017).

138. Hardwick, J. S., Lane, A. N. & Brown, T. Epigenetic Modifications of Cytosine: Biophysical Properties, Regulation, and Function in Mammalian DNA. *BioEssays* **40**, 1700199 (2018).

139. Kriaucionis, S. & Heintz, N. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science* **324**, 929–930 (2009).

140. Cheng, Y., Bernstein, A., Chen, D. & Jin, P. 5-Hydroxymethylcytosine: A new player in brain disorders? *Exp. Neurol.* **268**, 3–9 (2015).

141. Al-Mahdawi, S., Virmouni, S. A. & Pook, M. A. The emerging role of 5-hydroxymethylcytosine in neurodegenerative diseases. *Front. Neurosci.* **8**, (2014).

142. Brasa, S. *et al.* Reciprocal changes in DNA methylation and hydroxymethylation and a broad repressive epigenetic switch characterize *FMR1* transcriptional silencing in fragile X syndrome. *Clin. Epigenetics* **8**, 15 (2016).

143. Al-Mahdawi, S., Sandi, C., Mouro Pinto, R. & Pook, M. A. Friedreich Ataxia Patient Tissues Exhibit Increased 5-Hydroxymethylcytosine Modification and Decreased CTCF Binding at the *FXN* Locus. *PLoS ONE* **8**, e74956 (2013).

144. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci.* **97**, 5237–5242 (2000).

145. Tomizawa, S. -i. *et al.* Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**, 811–820 (2011).

146. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* **13**, 97–109 (2012).

147. Busche, S. *et al.* Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol.* **16**, 290 (2015).

148. Krijger, P. H. L., Geeven, G., Bianchi, V., Hilvering, C. R. E. & de Laat, W. 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. *Methods* **170**, 17–32 (2020).

149. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).

150. Kunowska, N. Studying DNA Methylation in Single-Cell Format with scBS-seq. in *Single Cell Methods* (ed. Proserpio, V.) vol. 1979 235–250 (Springer New York, 2019).

151. Simone, N. L., Bonner, R. F., Gillespie, J. W., Emmert-Buck, M. R. & Liotta, L. A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 (1998).

152. Ruegsegger, C. *et al.* Impaired mTORC1-Dependent Expression of Homer-3 Influences SCA1 Pathophysiology. *Neuron* **89**, 129–146 (2016).

153. Martuscello, R. T., Louis, E. D. & Faust, P. L. A Stainless Protocol for High Quality RNA Isolation from Laser Capture Microdissected Purkinje Cells in the Human Post-Mortem Cerebellum. *J. Vis. Exp.* 58953 (2019)

154. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10**, 295–304 (2009).

155. Li, E. & Zhang, Y. DNA Methylation in Mammals. *Cold Spring Harb. Perspect. Biol.* **6**, a019133–a019133 (2014).

156. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).

157. Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev.* **25**, 679–684 (2011).

158. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).

159. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).

Appendix A

Table A.1 - PCR mix reaction for the amplification of *DAB1* pentanucleotide repeat and flanking region.

PCR Mix:	
5X PrimeStar GxLBuffer (Takara; Mg2+ Plus)	4.00 μL
PrimeSTAR GxL DNA Polymerase (Takara; Stock concentration: 1.25 U/ μL)	0.25 μL
dNTP mixture (Takara; 2.5mM each dNTP)	0.80 μM
Primer F	0.20 μM
Primer R	0.20 μM
DNA	50 ng
Water up to	20 μL

Table A.2 - Thermocycler conditions for the amplification of *DAB1* pentanucleotide repeat and flanking region.

PCR conditions	**Short normal alleles**	**Large, interrupted and mutant alleles**
1st denaturation	98ºC for 1 minute	
2nd denaturation	98ºC for 10 seconds	
Annealing	68ºC for 1 minute and 45 seconds	68ºC for 4 minutes
Extension		
Cycles	Back to step 2 for 27 times	Back to step 2 for 30 times
Final extension	68ºC for 10 minutes	

Appendix B

Table B.1 - Primer identification and respective sequences for SNP analysis.

Position	Primer ID	Primer sequence	Design Tool
External	Alu CTCF 2F	5'-GCATAGTCGACTTGTTTCACACATGGCAGTGG-3'	Primer 3 Software (v 0.4.0)
	Alu CTCF 2R	5'-GCACAGAATTCCTGAGCTTGGGGCTTTTATGG-3'	
Internal	Alu 24F	5' – ATTTGCCCTTTGCTGATTGA - 3'	
	Alu 24R	5' - TGAAACTGAGGCTCAAAAATGA - 3'	
	Frag 2F	5' – AAAAAAGTAAGATTCGGAAGTCTG - 3'	
	Seqlocus A (F)	5' – TGTAGTTTTGAGTTTCCATCTCCTC - 3'	
	Seqlocus B (R)	5' – CAGGCTCACCCCACTTTG - 3'	

Table B.2 - Primer identification and respective sequence for Bisulfite Sequencing analysis.

Amplicon	Primer ID	Primer sequence	Design Tool
CTCF BS 1	BS-Seq CTCFBS1_F	5' - ATGAAATTGAGGTTTAAAATGA – 3’	MethylPrimer ® (ThermoFisher)
	BS-Seq CTCFBS1_R	5' - CAAAATCTCACTATATTACCCAA - 3'	
CpG island	BS-Seq CpGi_F	5’- GTTTGTAGTAAGAATTAATTAGTAAAGGGT -3’	MethPrimer 2.0 (The Li Lab)
	BS-Seq CpGi_R	5’- GTTTGTAGTAAGAATTAATTAGTAAAGGGT -3’	
Intermediate region	BS-Seq Reg_int_2F	5’ - AAATTATGATTGTAGATTTGGGTTT - 3’	MethylPrimer ® (ThermoFisher)
	BS-Seq Reg_int_2R	5’ - CAACAATACCCTAAATATATACACACC - 3’	
CTCF BS 2	BS-Seq CTCFBS2_F	5’- TTTTGGATTTTGAGTATTGATGAG -3’	MethPrimer 2.0 (The Li Lab)
	BS-Seq CTCFBS2_R	5’- AAAACACTACAAATTCCCACTCTAA -3’	

Appendix C

ZymoClean ™ Gel DNA Recovery Kit

Large nonpathogenic alleles and mutant alleles were separated in a 1.5% agarose gel. DNA on the bands corresponding to each allele are excised from the gel using a blade at the UV transilluminator, then transfered to a 1.5 ml Eppendorf tube.

After measuring the agarose slice weight, it was added three volumes of Agarose Dissolving Buffer (ADB) to each 100 mg of agarose excised from the gel and the mixture was Incubated at 50ºC for 10 minutes (until its completely dissolved).

The melted agarose was transfered to the ZymoSpin™ Column in a Collection tube and centrifuged at 13 000 rpm for 1 minute. The flow-through was discarded.

It was added 200 ml of DNA Wash Buffer to the column and centrifuged at 13 000 rpm for 30 seconds. This step was repeated, increasing to 1 minute of centrifugation.

The column was placed in a labeled 1.5 ml Eppendorf tube, 20 μL ultra-pure water (at 65ºC) was added and the tube was centrifuged at 13 000 rpm for 1 minute.

The flow-through was collected with a pipette, loaded again into the column and centrifuged at 13 000 rpm for 1 minute.

Appendix D

Illustra ™ Sephadex G-50 Fine DNA Grade (Particle size dry 50 μm – 150 μm)

The columns were washed adding 700 μL of ultra-pure water and centrifuged at 2 000 rcf for 4 minutes. The flow-through was discarded.

It was added 700 μL of Sephadex G-50 to the column and centrifuged at 2 000 rcf for 4 minutes. A Sephadex column was created at the tube and the flow-through is discarded.

The columns were placed into a 1.5 ml Eppendorf and 10 μL of sequencing template was added to the center of the Sephadex column, followed by a centrifugation at 2 000 rcf for 4 minutes.

In order to increase the stability of the single stranded DNA molecules and to reach the minimum volume of 12 μL, 3-6 μL of Hi-Di™ Formamide (Applied Biosystems) was added to the eluted DNA.

Appendix E

Creation of an analysis protocol

It is necessary to create an analysis protocol in order to define the conditions to be applied equally to all samples. The analysis protocol settings include 1) basecaller; 2) mixed bases; 3) clear range; 4) filtering and the following parameters were stablished for each:

1. Basecalling
    a. Basecaller: KB.bcp;
    b. DyeSet/Primer: KB_3100_POP7_BDTv3.mob
    c. Processed Data: True Profile
    d. Ending base: At PCR stop
    e. Quality threshold: Call al bases and assign quality values (QV).
2. Mixed bases
    a. Selected “Use Mixed Base Identification” and “Call IUB if 2nd highest peak is ≥ 15% of the highest peak”.
3. Clear range
    a. Selected “Use quality values” and “Remove bases from the ends until fewer than **4** bases out of **20** have QVs less than **20**”;
    b. Selected “Use reference trimming”.
4. Filter

    The criteria used for rejecting sequences if they do not meet minimum standard was:

    i. Maximum mixed bases (%): 20.0
    ii. Maximum Ns (%): 10.0
    iii. Minimum Clear Length (bp): 50
    iv. Minimum Sample Score: 25

It is necessary to define settings for Gap and Extension Penalties to ensure that the Gap and Extension Penalties added allows the alignment to be extended into regions where one sequence may have lost or gained characters. The settings used in the analysis defaults were:

1. Gap penalty: 22.5
2. Extension penalty: 8.5
3. Library matches: 20

Creation of a reference data group (RDG)

The Reference Data Group (RDG) defines the sequence and the known nucleotide variants to which SeqScape® Software compares the consensus segments.

The reference sequence is a contiguous DNA sequence comprising the region of interest, and it was downloaded from the UCSC Genome Browser (Human Assembly: Feb. 2009 (GRCh37/hg19)).

Creation and analysis of a project

In order to facilitate the analysis process, for each DNA sample (ID), a New Project Wizard was created. For each Project created, the next steps were followed:

1. Creation of a New Project Wizard;
2. Select the Besecalling settings;
3. Select Ending Base at PCR stop;
4. Select the RDG created before;
5. It is necessary to differentiate the sequences from normal alleles and the mutant alleles:
    a. For unaffected individuals a single Specimen is created, and the sequences belonging to that individual are loaded into that file.
    b. For affected individuals, it is necessary to create two Specimens – one corresponding to normal allele and other to mutant allele.

Analysis of data

After the importation of all data, at the Project Navigator it is possible to view all the variants found at the alignment to the reference sequence and the electropherogram fragment correspondent to each one. Furthermore, it is also possible to download a report from each Project.

Appendix F

QiAamp ® DNA Mini Kit (QiAGEN)

DNA of fibroblast cell lines was extracted according to the following protocol:

1. An approximate number of 5 x 10^6 cells was detached from the culture flask by trypsinization.
   1.1. To trypsinize cells, the medium was discarded and the cells were washed with PBS. The PBS was discarded, and 0.10-0.25% of trypsin was added. After cells have detached from the flask, the medium was collected, and cells were transferred to a 1.5 ml microcentrifuge tube and centrifuged at 300g for. 5 minutes, followed by supernatant removal.
   1.2. Several washes were performed in order to eliminate the remaining medium from the pellet, resuspending cell pellet in 200 μL of PBS followed by a centrifugation step at 10 000 rpm for 5 minutes. The supernatant was discarded without disturbing the cell pellet and step 1.2 is repeated until media was completely removed.

2. Cell pellet was resuspended in PBS to a final volume of 200 μL.
3. It was added 20 μL of proteinase K.
4. 200 μL of Buffer AL was added to the sample and mixed by pulse-vortexing for 15 seconds until it gets a homogeneous solution.
5. It was followed by an incubation at 56ºC for 10 minutes.
6. To remove drops from the top of the lid, a spin-down centrifugation was performed.
7. It was added 200 μL of ethanol (96-100%) to the sample, and vortexed for 15 seconds, followed by a spin-down.
8. The mixture was transferred to the QIAamp Mini spin column (placed in a 2 ml collection tube) and centrifuged at 8000 rpm for 1 minute. The QIAamp Mini column was placed in a clean 2 ml collection tube and the previous collection tube with the filtrate was discarded.
9. After opening the QIAamp Mini spin column, 500 μL Buffer AW1 was added, followed by a centrifugation at 8000 rpm for 1 minute. The collection tube with the filtrate wass discarded and the QIAamp Mini spin column was placed in a clean 2 ml collection tube.
10. It was added 500 μL of Buffer AW2 and after closing the tap, centrifugated at full speed for 3 minutes.
11. The collection tube with the filtrate was discarded and the QIAamp Mini spin column was placed in a clean 2 ml collection tube, followed by a dry centrifugation at full speed for 1 minute, in order to eliminate the chance of possible Buffer AW2 carryover.

12. The QIAamp Mini spin column was placed in a clean 1.5 ml microcentrifuge tube and the collection tube with the filtrate was discarded.
13. The DNA was eluted from the QIAamp spin column with 150 μL distilled water, previously heated at 60ºC. After 1 minute of incubation at room temperature (15-25ºC), it was centrifuged at 8000 rpm for 1 minute.
14. After the first elution, the eluted DNA was again placed in the QIAamp Mini spin column and centrifuged at 8000 rpm for 1 minute.

Appendix G

EZ DNA Methylation Direct ™ (Zymo Research Corp)

The Bisulfite conversion of DNA was performed according to the following steps:

1. 500 ng of genomic DNA and 130 μL of CT Conversion Reagent solution was added to a PCR tube. If the volume of DNA is less than 20 μL, it should be compensated with water to a final volume of 150 μL.
2. The PCR tubes were placed in a thermal cycler and the following program was run:
    i. 98ºC for 8 minutes;
    ii. 64ºC for 3.5 hours;
    iii. 4ºC storage for up to 20 hours (optional)
3. It was added 600 μL of M-Binding Buffer into a Zymo-Spin™ IC Column.
4. The mixture from step 2 was loaded into the Zymo-Spin™ IC Column containing the M-Binding Buffer and, after closing the cap, the column was inverted several times to mix the solution.
5. The tube was centrifuged at full speed for 30 seconds and the flow-through was discarded.
6. 100 μL of M-Wash Buffer was loaded into the column and centrifuged at full speed for 30 seconds.
7. It was added 200 μL of M-Desulphonation Buffer to the column, followed by an incubation at room temperature (20-30ºC) for 20 minutes. After the incubation, the tube was centrifuged at full speed for 30 seconds.
8. 200 μL of M-Wash Buffer was added to the column and centrifugated at full speed for 30 seconds. This step was repeated by adding 200 μL of M-Wash Buffer and centrifuge for an additional 30 seconds.
9. The flow-through was discarded and a dry centrifugation step (with empty column) at full speed for 30 seconds was performed.
10. The column was placed into a 1.5 ml centrifuge tube and the bisulfite-converted DNA was eluted by adding 10 μL of water (previously heated at 60ºC), followed by a centrifugation at full speed for 1 minute.
11. After the first elution, the eluted DNA was again placed in the Zymo-Spin™ IC Column and centrifuged at full speed for 1 minute.

Appendix H

Table H.1 - PCR mix for the region of interest amplification for Bisulfite Sequencing analysis.

PCR Mix	
2X GoTaq® Master Mix	6.25 μL
Primer Forward	0.20 μM
Primer Reverse	0.20 μM
DNA	10.00 ng
Water up to	12.50 μL

Table H.2 - PCR conditions for the region of interest amplification for Bisulfite Sequencing analysis, specified for each amplicon.

PCR Program	CTCF Binding Site 1	CpG island	Region between CpG island and CTCF BS 2	CTCF Binding Site 2
1st denaturation	95ºC for 2 minutes			
2nd denaturation	95ºC for 45 seconds			
Annealing	55ºC for 40 seconds	57ºC for 40 seconds	57ºC for 40 seconds	59ºC for 40 seconds
Extension	72ºC for 45 seconds			
Cycles	Back to step 2 for 35 times	Back to step 2 for 32 times	Back to step 2 for 30 times	Back to step 2 for 35 times
Final extension	72ºC for 10 minutes			

Appendix I

Table I.1 – PCR mixtures and thermocycler conditions for the optimization of the amplification and Sanger sequencing reactions of the CTCF-BS 1 region.

	PCR reaction					Sequencing reaction			
CTCF BS 1	1st try (Gradient for the annealing temperature)					1st try			
	Mix		Conditions		Results	Mix		Conditions	Results
	H_2O	3.25 µL	95ºC - 2'		No amplification	H2O	5 µL	95ºC - 1'	Clean chromatograms
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35		BigDye 1:1	2 µL	95ºC - 30'' x 35	
	Primer F (10 µM)	1 µL	56-57ºC - 40''			Primer (10 µM)	0.5 µL	56ºC - 10''	
	Primer R (10 µM)	1 µL	72ºC - 45''			PCR Template	2.5 µL	60ºC - 2'	
	DNA [25ng/µL]	1 µL	72ºC - 10'					60ºC - 10'	
	2nd try (↓ annealing temperature gradient)								
	Mix		Conditions		Results				
	H_2O	3.25 µL	95ºC - 2'		At 54ºC amplifies, with a smooth smear				
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35					
	Primer F (10 µM)	1 µL	50-54ºC - 40''						
	Primer R (10 µM)	1 µL	72ºC - 45''						
	DNA [25ng/µL]	1 µL	72ºC - 10'						
	3rd try (↑ annealing temperature)								
	Mix		Conditions		Results				
	H_2O	3.25 µL	95ºC - 2'		Good amplification				
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35					
	Primer F (10 µM)	1 µL	55ºC - 40''						
	Primer R (10 µM)	1 µL	72ºC - 45''						
	DNA [25ng/µL]	1 µL	72ºC - 10'						
	4rd try (↓ DNA concentration)					2nd try			
	Mix		Conditions		Results	Mix		Conditions	Results
	H_2O	3.25 µL	95ºC - 2'		Good amplification	H2O	5 µL	95ºC - 1'	Clean chromatograms
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35		BigDye 1:1	2 µL	95ºC - 30'' x 35	
	Primer F (10 µM)	1 µL	55ºC - 40''			Primer (10 µM)	0.5 µL	56ºC - 10''	
	Primer R (10 µM)	1 µL	72ºC - 45''			PCR Template	2.5 µL	60ºC - 2'	
	DNA [10ng/µL]	1 µL	72ºC - 10'					60ºC - 10'	

Table I.2 - PCR mixtures and thermocycler conditions for the optimization of the amplification and Sanger sequencing reactions of CpG island.

	PCR reaction					Sequencing reaction			
CpG island	1st try					1st try			
	Mix		Conditions		Results	Mix		Conditions	Results
	H_2O	3.25 µL	95ºC - 2'		Low intensity band with the expected size	H2O	5 µL	95ºC - 1'	Clean chromatograms
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35		BigDye 1:1	2 µL	95ºC - 30'' x 35	
	Primer F (10 µM)	1 µL	55ºC - 40''			Primer (10 µM)	0.5 µL	56ºC - 10''	
	Primer R (10 µM)	1 µL	72ºC - 45''			PCR Template	2.5 µL	60ºC - 2'	
	DNA [25ng/µL]	1 µL	72ºC - 10'					60ºC - 10'	
	2nd try								
	Mix		Conditions		Results				
	H_2O	3.25 µL	95ºC - 2'		High intensity band with the expected size				
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35					
	Primer F (10 µM)	1 µL	57ºC - 40''						
	Primer R (10 µM)	1 µL	72ºC - 45''						
	DNA [25ng/µL]	1 µL	72ºC - 10'						

Table I.3 - PCR mixtures and thermocycler conditions for the optimization of the amplification and Sanger sequencing reactions of the region between the CpG island and the CTCF-BS 2.

	PCR reaction					
Region between CpG island and CTCF BS 2	1st try					
	Mix		Conditions		Results	
	H_2O	3.25 µL	95ºC - 2'		High intensity band, with a smear of <100 bp>	
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35		
	Primer F (10 µM)	1 µL	57ºC - 40''			
	Primer R (10 µM)	1 µL	72ºC - 45''			
	DNA [10 ng/µL]	1 µL	72ºC - 10'			
	2nd try (↑Annealing temp)					
	Mix		Conditions		Results	
	H_2O	3.25 µL	95ºC - 2'		Low intensity of the band with the expected size; Inespecific band with 100 bp with low intensity	
	GoTaq 2x	6.25 µL	95ºC - 45''	x 35		
	Primer F (10 µM)	1 µL	59ºC - 40''			
	Primer R (10 µM)	1 µL	72ºC - 45''			
	DNA [10 ng/µL]	1 µL	72ºC - 10'			
	3rd try (↓ cycles)					
	Mix		Conditions		Results	
	H_2O	3.25 µL	95ºC - 2'		High intensity of the band with the expected size; Low intensity inespecific band with 100 bp	
	GoTaq 2x	6.25 µL	95ºC - 45''	x 30		
	Primer F (10 µM)	1 µL	57ºC - 40''			
	Primer R (10 µM)	1 µL	72ºC - 45''			
	DNA [10 ng/µL]	1 µL	72ºC - 10'			
	4rd try (↓cycles + ↓ annealing temperature)					
	Mix		Conditions		Results	
	H_2O	3.25 µL	95ºC - 2'		Low intensity of the band with the expected size;	
	GoTaq 2x	6.25 µL	95ºC - 45''	x 30		
	Primer F (10 µM)	1 µL	58ºC - 40''			
	Primer R (10 µM)	1 µL	72ºC - 45''			
	DNA [10 ng/µL]	1 µL	72ºC - 10'			
	5ft try (↓cycles + ↓ annealing temperature)					
	Mix		Conditions		Results	
	H_2O	3.25 µL	95ºC - 2'		No amplification	
	GoTaq 2x	6.25 µL	95ºC - 45''	x 28		
	Primer F (10 µM)	1 µL	57ºC - 40''			
	Primer R (10 µM)	1 µL	72ºC - 45''			
	DNA [10 ng/µL]	1 µL	72ºC - 10'			
	6th try (New enzyme: NzyTaq II)					
	Mix		Conditions		Results	
	H_2O	20.25 µL	95ºC - 3'		Good amplification of the band with the expected size	
	NzyTaqII	25 µL	94ºC - 30''	x 35		
	Primer F (10 µM)	1 µL	57ºC - 30''			
	Primer R (10 µM)	1 µL	72ºC - 30''			
	DNA [10 ng/µL]	2 µL	72ºC - 10'			
	7th try (↓ total volume)					
	Mix		Conditions		Results	
	H_2O	3.25 µL	95ºC - 3'		Good amplification of the band with the expected size; Little shadow at the 100 bp	
	NzyTaqII	6.25 µL	94ºC - 30''	x 35		
	Primer F (10 µM)	1 µL	57ºC - 30''			
	Primer R (10 µM)	1 µL	72ºC - 30''			
	DNA [10 ng/µL]	1 µL	72ºC - 10'			

Sequencing reaction				
1st try				
Mix		Conditions		Results
H2O	5 µL	95ºC - 1'		Chromatograms with background noise, probably due to the shadow with 100 bp at the PCR
BigDye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR Template	2.5 µL	60ºC - 2'		
		60ºC - 10'		

*Table I.3 continues on the next page

8th try (↓ primer concentration)				
Mix		Conditions		Results
H_2O	4.65 µL	95ºC - 3'		Very low amplification.
NzyTaqII	6.25 µL	94ºC - 30''	x 35	
Primer F (10 µM)	0.5 µL	57ºC - 30''		
Primer R (10 µM)	0.5 µL	72ºC - 30''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

9th try (↓ primer concentration)				
Mix		Conditions		Results
H_2O	4.65 µL	95ºC - 3'		Good amplification of the band with the expected size; Little shadow at the 100 bp
NzyTaqII	6.25 µL	94ºC - 30''	x 35	
Primer F (10 µM)	0.4 µL	57ºC - 30''		
Primer R (10 µM)	0.4 µL	72ºC - 30''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

Sequencing reaction				
2nd try				
Mix		Conditions		Results
H2O	5 µL	95ºC - 1'		Chromatograms with background noise, probably due to the shadow with 100 bp at the PCR
BigDye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR Template	2.5 µL	60ºC - 2'		
		60ºC - 10'		

Table I.4 - PCR mixtures and thermocycler conditions for the optimization of the amplification and Sanger sequencing reactions of the CTCF-BS 2 region.

PCR reaction				
1st try				
Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		High intensity band with the expected size; little primer dimers
GoTaq 2x	6.25 µL	95ºC - 45''	x 35	
Primer F (10 µM)	1 µL	57ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [50 ng/µL]	1 µL	72ºC - 10'		

2nd try (↓ cycles; ↓ DNA concentration)				
Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		Good intensity band with the expected size
GoTaq 2x	6.25 µL	95ºC - 45''	x 32	
Primer F (10 µM)	1 µL	57ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [25 ng/µL]	1 µL	72ºC - 10'		

3rd try (↓ cycles; ↓ DNA concentration)				
Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		Good intensity band with the expected size
GoTaq 2x	6.25 µL	95ºC - 45''	x 32	
Primer F (10 µM)	1 µL	57ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [25 ng/µL]	1 µL	72ºC - 10'		

4rd try (↓ cycles; ↓ DNA concentration)				
Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		Good intensity band with the expected size
GoTaq 2x	6.25 µL	95ºC - 45''	x 30	
Primer F (10 µM)	1 µL	57ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [25 ng/µL]	1 µL	72ºC - 10'		

5ft try (↓ cycles; ↓ DNA concentration)				
Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		Good intensity band with the expected size
GoTaq 2x	6.25 µL	95ºC - 45''	x 32	
Primer F (10 µM)	1 µL	57ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

Sequencing reaction				
1st try				
Mix		Conditions		Results
H_2O	5 µL	95ºC - 1'		Primer Reverse works fine; Primer Forward does not work
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR template	2.5 µL	60ºC - 4'		
		60ºC - 10'		

Sequencing reaction (only for primer forward)				
2nd try (↑ cycles)				
Mix		Conditions		Results
H_2O	5 µL	95ºC - 1'		Better than the 1st try, but low signal
Big Dye 1:1	2 µL	95ºC - 30''	x 40	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR template	2.5 µL	60ºC - 4'		
		60ºC - 10'		

[a] Table I.4 continues on the next page

CTCF BS 2

6th try (10 ng/uL DNA x 32 cycles)

Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		Good intensity band with the expected size
GoTaq 2x	6.25 µL	95ºC - 45''	x 32	
Primer F (10 µM)	1 µL	57ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

3rd try (↓ template from PCR)

Mix		Conditions		Results
H_2O	6.25 µL	95ºC - 1'		Better than the 1st try, but low signal
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR template	1.25 µL	60ºC - 4'		
		60ºC - 10'		

4rd try (↓ template from PCR)

Mix		Conditions		Results
H_2O	6.75 µL	95ºC - 1'		Drastic decrease of the signal at the 100 bp
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR template	0.75 µL	60ºC - 4'		
		60ºC - 10'		

5ft try (↑ primer concentration)

Mix		Conditions		Results
H_2O	8 µL	95ºC - 1'		Drastic decrease of the signal at the 100 bp
Big Dye 1:1	4 µL	95ºC - 30''	x 35	
Primer (10 µM)	4 µL	56ºC - 10''		
PCR template	4 µL	60ºC - 4'		
		60ºC - 10'		

6th try (↑ annealing time)

Mix		Conditions		Results
H_2O	5 µL	95ºC - 1'		Better, but with background noise and decrease in the signal at the 100 bp
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 15''		
PCR template	2.5 µL	60ºC - 4'		
		60ºC - 10'		

7th try (↑ initial denaturation temperature)

Mix		Conditions		Results
H_2O	5 µL	96ºC - 1'		Low signal and background noise
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	56ºC - 10''		
PCR template	2.5 µL	60ºC - 4'		
		60ºC - 10'		

8th try (↑ BigDye concentration)

Mix (vt=20 µL)		Conditions		Results
H_2O	9 µL	95ºC - 1'		Drastic decrease of the signal at the 80 bp, with background noise
Big Dye	4 µL	95ºC - 30''	x 35	
Big Dye Buffer	2 µL	56ºC - 10''		
Primer (10 µM)	1 µL	60ºC - 4'		
PCR template	4 µL	60ºC - 10'		

7th try (↑ annealing temperature)

Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		At 59ºC, good amplification; At 61ºC no amplification.
GoTaq 2x	6.25 µL	95ºC - 45''	x 32	
Primer F (10 µM)	1 µL	59-61ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

9th try (↑ primer concentration - vt= 20 µL)

Mix		Conditions		Results
H_2O	8 µL	96ºC - 1'		Low signal after the 80th bp and background noise
Big Dye 1:1	4 µL	95ºC - 30''	x 35	
Primer (10 µM)	4 µL	56ºC - 10''		
PCR template	4 µL	60ºC - 4'		
		60ºC - 10'		

8th try (other DNA sample)

Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 2'		Good amplification
GoTaq 2x	6.25 µL	95ºC - 45''	x 35	
Primer F (10 µM)	1 µL	59ºC - 40''		
Primer R (10 µM)	1 µL	72ºC - 45''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

10th try (↑ annealing temperature)

Mix		Conditions		Results
H_2O	8 µL	96ºC - 1'		It worked with this DNA sample; other samples did not worked.
Big Dye 1:1	4 µL	95ºC - 30''	x 35	
Primer (10 µM)	4 µL	58ºC - 10''		
PCR template	4 µL	60ºC - 4'		
		60ºC - 10'		

11th try (add DMSO at 2%, 5% and 10%)

Mix		Conditions		Results
H_2O	x	96ºC - 1'		It worked with 2% DMSO ifor 1 individual;
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	58ºC - 10''		
DMSO	x	60ºC - 4'		
PCR template	2.5 µL	60ºC - 10'		

*Table I.4 continues on the next page

9th try (add DMSO to PCR reaction)				
Mix		Conditions		Results
H_2O	3 µL	95ºC - 3'		Good amplification; little shadow at 400 bp
NzyTaq II	6.25 µL	94ºC - 30''	x 35	
Primer F (10 µM)	1 µL	59ºC - 30''		
Primer R (10 µM)	1 µL	72ºC - 30''		
DMSO 2%	0.25 µL			
DNA [10 ng/µL]	1 µL	72ºC - 10'		

10th try				
Mix		Conditions		Results
H_2O	3.25 µL	95ºC - 3'		Good amplification
NzyTaq II	6.25 µL	94ºC - 30''	x 35	
Primer F (10 µM)	1 µL	59ºC - 30''		
Primer R (10 µM)	1 µL	72ºC - 30''		
DNA [10 ng/µL]	1 µL	72ºC - 10'		

11th try (other enzyme: GXL Prime Star)				
Mix		Conditions		Results
H_2O	32 µL	98ºC - 1'		No amplification
Buffer 5x	10 µL	98ºC - 10''	x 35	
dNTP	4 µL	57ºC - 15''		
Primer F (10 µM)	1 µL			
Primer R (10 µM)	1 µL	68ºC - 30''		
GXL Prime Star	1 µL	72ºC - 10'		
DNA [10 ng/µL]	1 µL			

12th try				
Mix		Conditions		Results
H_2O	5 µL	96ºC - 1'		Drastic decrease in the signal after the 50 bp
Big Dye 1:1	2 µL	95ºC - 30''	x 35	
Primer (10 µM)	0.5 µL	58ºC - 10''		
PCR template	2.5 µL	60ºC - 4'		
		60ºC - 10'		

12th try (↑ Big Dye concentration + DMSO 2%)				
Mix		Conditions		Results
H_2O	3.8 µL	96ºC - 1'		
Big Dye	2 µL	95ºC - 30''	x 35	
Buffer Big Dye	1 µL	58ºC - 10''		
Primer (10 µM)	0.5 µL			
DMSO 2%	0.2 µL	60ºC - 4'		
PCR template	2.5 µL	60ºC - 10'		

Appendix J

ESME ® Algorithm

After inputing the .abi files resulting from bisulfite-converted DNA sequencing and the corresponding reference sequence (.fa file), ESME® runs the algorithm that processes the data according to the following steps:

1. **Entropy-based clipping** – removes long stretches of DNA with high background signals, present at the end of the amplicon.
2. **Signal detection** – it calculates the corresponding base intensities of each dNTP, estimating the base proportions in the molecular mixture.
3. **Alignment** – it aligns sample sequences (.abi files) with the reference sequence. Before the alignment, the reference sequence loaded to the program is bisulfite-converted *in silico*. The converted sequence has the non-CpG cytosines transformed to *t* in order to distinguish the thymines derived from the uracil (BS-converted) from the thymines originally present at the genomic sequence. The cytosines present at a CpG, corresponding to the positions where the quantification occurs, are maintained as *C*.
4. **Trace correction** – In standard analysis, when the same position has mixed signals (e.g., C and T signals), the program interprets the data as being two adjacent bases (because the standard sequencers expect one homogeneous DNA population to be sequenced). In direct bisulfite-treated DNA analysis, it is necessary to correct this interpretation. The algorithm calculates the signal from each one of the mixed bases and then fuses them into a single base.
5. **Alignment clipping** – removes the poor signal from the flanking regions after the alignment, in order to maintain the total alignment error below 10%.
6. **Signal normalization** – The cytosine picks are normally over-scaled when compared to the rest of the sequence (it may be the result of the standard basecaller software compensating for the low frequency of cytosines). The software normalizes cytosines trace curves prior to methylation rate calculation (the standard minimum CpG number required for cytosines normalization is 3), applying these formulas:

(1) $$\overline{T_T^{\text{norm}}} \equiv \overline{T_C^{\text{norm}} + C_C^{\text{norm}}}$$

(2) $$F_C = \frac{\overline{T_T^{\text{int}}} - \overline{T_C^{\text{int}}}}{\overline{C_C^{\text{int}}}}$$

$$C_b^{\text{norm}} = F_C C_b^{\text{int}}, b \in \{C, t, A, G, T\} \quad (3)$$

*B_b^{norm} represents the normalized base intensities: $B \in$ {A,C,G,T};

7. **Compensation of incomplete conversion** – The software applies a formula that compensate the *C* trace signals (through the calculation of a conversion rate (R), to consider the scenario of an incomplete conversion.

$$R = \frac{T_t^{\text{norm}}}{T_t^{\text{norm}} + C_t^{\text{norm}}} \quad (4)$$

8. **Methylation estimation** – After the bisulfite conversion rate estimation, the algorithm applies the formula that correctly calculates the methylation rate in each sample of DNA, by incorporating the previous correction:

$$M = 1 - \frac{T_C^{\text{norm}}}{(C_C^{\text{norm}} + T_C^{\text{norm}}) R_{\text{glob}}} \quad (5)$$

For the regions with less than 3 CpG dinucleotides (e.g., sequences corresponding to CTCF BS 1 having only 2 CpGs), a modified version of ESME ® software was used for the analysis. As the ESME ® settings at the executable program does not allow the alteration of the minimum CpG for normalization, I modified the running code of ESME® at Linux terminal, using *esme-h* command, and the minimum CpG number for normalization was set to 2. It is important to have into account that changing the default parameters might lead to a higher error rate, leading to a possible misinterpretation of the data.

Appendix K

Table K.1 – Genotype information of the peripheral blood samples analyzed by Sanger sequencing.

SNP information				Genotype																									
SNP ID	Position according to (ATTTT)n	Ancestral allele	Variant (MAF AI)	NI 1	NI 2	NI 3	NI 4	NI 5	NI 6	NI 7	NI 8		NI 9		NI 10		AI 1		AI 2		AI 3		AI 4		AI 5		AI 6		AI 7		AI 8	
											N	Int	N	G	N	G	N	M	N	M	N	M	N	M	N	M	N	M	N	M	N	M
SNP 1	Upstream	G (0.941)	A (0.059)	A/G	A/G	A/G	A/G	G/G	G/G	A/G	G	G	A	G	G	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	G
(ATTTT)n	-	(ATTTT)15	-	(ATTTT)7/11	(ATTTT)7/8	(ATTTT)14,/25	(ATTTT)7/12	(ATTTT)8/14	(ATTTT)8/12	(ATTTT)7/22	(ATTTT)12	(ATTTT)50	(ATTTT)7	(ATTTT)51	(ATTTT)14	(ATTTT)120	(ATTTT)14	(ATTTC)n	(ATTTT)11	(ATTTC)n	(ATTTT)14	(ATTTC)n	(ATTTT)14	(ATTTC)n	(ATTTT)10	(ATTTC)n	(ATTTT)51	(ATTTC)n	(ATTTT)139	(ATTTC)n	(ATTTT)14	(ATTTC)n
SNP 2		A (0.969)	G (0.031)	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	G	A	A	A	A	A	A
SNP 3		T (0.941)	A (0.059)	A/T	A/T	A/T	A/T	T/T	T/T	A/T	T	T	A	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
SNP 4		G (0.970)	A (0.030)	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G	G	G	G	G	G
SNP 5		C (0.941)	T (0.059)	C/T	C/T	C/T	C/T	C/C	C/C	C/T	C	C	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
SNP 6	Downstream	T (0.584)	C (0.416)	C/C	C/C	C/C	C/C	C/T	C/T	C/C	T	C	C	C	T	C	T	C	C	C	T	C	T	C	C	C	C	C	C	T	C
SNP 7		C (0.875)	T (0.125)	C/C	C/C	C/C	C/C	C/C	C/C	C/T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	T	T	T	T	C	T
SNP 8		G (0.940)	A (0.060)	G/A	G/A	G/A	G/A	G/G	G/G	G/A	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
SNP 9		C (0.783)	T (0.217)	C/C	C/C	C/C	C/T	C/C	C/C	C/T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	T	T	T	T	C	T

* NI = Normal individual; AI = Affected individual; N = normal allele; Int = interrupted allele; G = large allele; M = Mutant allele

CpG --> TpG
CpG --> CpA
TpG --> CpG

Appendix L

Table L.1 - Methylation rate values at CpG 1.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.92	218	CpG 1
Peripheral blood	AI 2	0.87	218	CpG 1
Peripheral blood	AI 3	0.81	218	CpG 1
Peripheral blood	AI 4	0.9	218	CpG 1
Peripheral blood	AI 5	0.95	218	CpG 1
Peripheral blood	AI 6	0.9	218	CpG 1
Peripheral blood	AI 7	0.92	218	CpG 1
Peripheral blood	AI 8	0.93	218	CpG 1
Fibroblasts	AF 1	0.72	218	CpG 1
Fibroblasts	AF 2	0.5	218	CpG 1
Fibroblasts	AF 3	0.48	218	CpG 1
Zebrafish	AZ 1	0.7	218	CpG 1
Peripheral blood	NI 1	0.85	218	CpG 1
Peripheral blood	NI 2	0.85	218	CpG 1
Peripheral blood	NI 3	0.80	218	CpG 1
Peripheral blood	NI 4	0.82	218	CpG 1
Peripheral blood	NI 5	0.97	218	CpG 1
Peripheral blood	NI 6	0.81	218	CpG 1
Peripheral blood	NI 7	0.86	218	CpG 1
Peripheral blood	NI 8	0.88	218	CpG 1
Peripheral blood	NI 9	0.87	218	CpG 1
Peripheral blood	NI 10	0.80	218	CpG 1
Fibroblasts	NF 1	0.15	218	CpG 1
Fibroblasts	NF 2	0.14	218	CpG 1
Fibroblasts	NF 3	0.37	218	CpG 1
Zebrafish	NZ 1	0.31	218	CpG 1

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.2 - Methylation rate values at CpG 2.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.83	266	CpG 2
Peripheral blood	AI 2	0.76	266	CpG 2
Peripheral blood	AI 3	1	266	CpG 2
Peripheral blood	AI 4	0.89	266	CpG 2
Peripheral blood	AI 5	0.78	266	CpG 2
Peripheral blood	AI 6	0.87	266	CpG 2
Peripheral blood	AI 7	0.76	266	CpG 2
Peripheral blood	AI 8	0.89	266	CpG 2
Fibroblasts	AF 1	0.62	266	CpG 2
Fibroblasts	AF 2	0.75	266	CpG 2
Fibroblasts	AF 3	0.64	266	CpG 2
Zebrafish	AZ 1	0.79	266	CpG 2
Peripheral blood	NI 1	0.73	266	CpG 2
Peripheral blood	NI 2	0.58	266	CpG 2
Peripheral blood	NI 3	0.64	266	CpG 2
Peripheral blood	NI 4	0.7	266	CpG 2
Peripheral blood	NI 5	0.74	266	CpG 2
Peripheral blood	NI 6	0.66	266	CpG 2
Peripheral blood	NI 7	0.79	266	CpG 2
Peripheral blood	NI 8	0.78	266	CpG 2
Peripheral blood	NI 9	0.75	266	CpG 2
Peripheral blood	NI 10	1	266	CpG 2
Fibroblasts	NF 1	0.24	266	CpG 2
Fibroblasts	NF 2	0.41	266	CpG 2
Fibroblasts	NF 3	0.36	266	CpG 2
Zebrafish	NZ 1	0.42	266	CpG 2

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.3 - Methylation rate values at CpG 4.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.87	58	CpG 4
Peripheral blood	AI 2	0.9	58	CpG 4
Peripheral blood	AI 3	0.92	58	CpG 4
Peripheral blood	AI 4	0.87	58	CpG 4
Peripheral blood	AI 5	0.92	58	CpG 4
Peripheral blood	AI 6	0.94	58	CpG 4
Peripheral blood	AI 7	0.89	58	CpG 4
Peripheral blood	AI 8	0.97	58	CpG 4
Fibroblasts	AF 1	0.76	58	CpG 4
Fibroblasts	AF 2	0.74	58	CpG 4
Fibroblasts	AF 3	0.57	58	CpG 4
Zebrafish	AZ 1	0.64	58	CpG 4
Peripheral blood	NI 1	0.88	58	CpG 4
Peripheral blood	NI 2	0.93	58	CpG 4
Peripheral blood	NI 3	0.89	58	CpG 4
Peripheral blood	NI 4	0.88	58	CpG 4
Peripheral blood	NI 5	0.92	58	CpG 4
Peripheral blood	NI 6	0.91	58	CpG 4
Peripheral blood	NI 7	0.91	58	CpG 4
Peripheral blood	NI 8	0.88	58	CpG 4
Peripheral blood	NI 9	0.88	58	CpG 4
Peripheral blood	NI 10	0.88	58	CpG 4
Fibroblasts	NF 1	0.38	58	CpG 4
Fibroblasts	NF 2	0.56	58	CpG 4
Fibroblasts	NF 3	0.49	58	CpG 4
Zebrafish	NZ 1	0.8	58	CpG 4

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.4 - Methylation rate values at CpG 5.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.85	108	CpG 5
Peripheral blood	AI 2	0.90	108	CpG 5
Peripheral blood	AI 3	0.85	108	CpG 5
Peripheral blood	AI 4	0.91	108	CpG 5
Peripheral blood	AI 5	0.87	108	CpG 5
Peripheral blood	AI 6	0.92	108	CpG 5
Peripheral blood	AI 7	0.88	108	CpG 5
Peripheral blood	AI 8	0.89	108	CpG 5
Fibroblasts	AF 1	0.56	108	CpG 5
Fibroblasts	AF 2	0.56	108	CpG 5
Fibroblasts	AF 3	0.65	108	CpG 5
Zebrafish	AZ 1	0.34	108	CpG 5
Peripheral blood	NI 1	0.8	108	CpG 5
Peripheral blood	NI 2	0.85	108	CpG 5
Peripheral blood	NI 3	0.87	108	CpG 5
Peripheral blood	NI 4	0.85	108	CpG 5
Peripheral blood	NI 5	0.84	108	CpG 5
Peripheral blood	NI 6	0.86	108	CpG 5
Peripheral blood	NI 7	0.85	108	CpG 5
Peripheral blood	NI 8	0.85	108	CpG 5
Peripheral blood	NI 9	0.85	108	CpG 5
Peripheral blood	NI 10	0.86	108	CpG 5
Fibroblasts	NF 1	0.43	108	CpG 5
Fibroblasts	NF 2	0.56	108	CpG 5
Fibroblasts	NF 3	0.47	108	CpG 5
Zebrafish	NZ 1	0.52	108	CpG 5

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.5 - Methylation rate values at CpG 6.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.91	126	CpG 6
Peripheral blood	AI 2	0.93	126	CpG 6
Peripheral blood	AI 3	0.89	126	CpG 6
Peripheral blood	AI 4	0.95	126	CpG 6
Peripheral blood	AI 5	0.94	126	CpG 6
Peripheral blood	AI 6	0.8	126	CpG 6
Peripheral blood	AI 7	0.94	126	CpG 6
Peripheral blood	AI 8	0.94	126	CpG 6
Fibroblasts	AF 1	0.54	126	CpG 6
Fibroblasts	AF 2	0.87	126	CpG 6
Fibroblasts	AF 3	0.88	126	CpG 6
Zebrafish	AZ 1	0.61	126	CpG 6
Peripheral blood	NI 1	0.51	126	CpG 6
Peripheral blood	NI 2	0.5	126	CpG 6
Peripheral blood	NI 3	0.94	126	CpG 6
Peripheral blood	NI 4	0.52	126	CpG 6
Peripheral blood	NI 5	0.94	126	CpG 6
Peripheral blood	NI 6	0.5	126	CpG 6
Peripheral blood	NI 7	0.52	126	CpG 6
Peripheral blood	NI 8	0.93	126	CpG 6
Peripheral blood	NI 9	0.52	126	CpG 6
Peripheral blood	NI 10	0.94	126	CpG 6
Fibroblasts	NF 1	0.66	126	CpG 6
Fibroblasts	NF 2	0.82	126	CpG 6
Fibroblasts	NF 3	0.85	126	CpG 6
Zebrafish	NZ 1	0.65	126	CpG 6

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.6 - Methylation rate values at CpG 7.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.89	137	CpG 7
Peripheral blood	AI 2	0.9	137	CpG 7
Peripheral blood	AI 3	0.93	137	CpG 7
Peripheral blood	AI 4	0.91	137	CpG 7
Peripheral blood	AI 5	0.93	137	CpG 7
Peripheral blood	AI 6	0.89	137	CpG 7
Peripheral blood	AI 7	0.94	137	CpG 7
Peripheral blood	AI 8	0.94	137	CpG 7
Fibroblasts	AF 1	0.86	137	CpG 7
Fibroblasts	AF 2	0.87	137	CpG 7
Fibroblasts	AF 3	0.83	137	CpG 7
Zebrafish	AZ 1	0.67	137	CpG 7
Peripheral blood	NI 1	0.87	137	CpG 7
Peripheral blood	NI 2	0.92	137	CpG 7
Peripheral blood	NI 3	0.93	137	CpG 7
Peripheral blood	NI 4	0.92	137	CpG 7
Peripheral blood	NI 5	0.92	137	CpG 7
Peripheral blood	NI 6	0.91	137	CpG 7
Peripheral blood	NI 7	0.92	137	CpG 7
Peripheral blood	NI 8	0.92	137	CpG 7
Peripheral blood	NI 9	0.91	137	CpG 7
Peripheral blood	NI 10	0.92	137	CpG 7
Fibroblasts	NF 1	0.66	137	CpG 7
Fibroblasts	NF 2	0.74	137	CpG 7
Fibroblasts	NF 3	0.73	137	CpG 7
Zebrafish	NZ 1	0.84	137	CpG 7

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.7 - Methylation rate values at CpG 8.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.6	193	CpG 8
Peripheral blood	AI 2	0.52	193	CpG 8
Peripheral blood	AI 3	0.77	193	CpG 8
Peripheral blood	AI 4	0.67	193	CpG 8
Peripheral blood	AI 5	0.73	193	CpG 8
Peripheral blood	AI 6	0.76	193	CpG 8
Peripheral blood	AI 7	0.68	193	CpG 8
Peripheral blood	AI 8	0.78	193	CpG 8
Fibroblasts	AF 1	0.61	193	CpG 8
Fibroblasts	AF 2	0.47	193	CpG 8
Fibroblasts	AF 3	0.41	193	CpG 8
Zebrafish	AZ 1	0.52	193	CpG 8
Peripheral blood	NI 1	0.56	193	CpG 8
Peripheral blood	NI 2	0.72	193	CpG 8
Peripheral blood	NI 3	0.71	193	CpG 8
Peripheral blood	NI 4	0.74	193	CpG 8
Peripheral blood	NI 5	0.77	193	CpG 8
Peripheral blood	NI 6	0.78	193	CpG 8
Peripheral blood	NI 7	0.77	193	CpG 8
Peripheral blood	NI 8	0.73	193	CpG 8
Peripheral blood	NI 9	0.7	193	CpG 8
Peripheral blood	NI 10	0.69	193	CpG 8
Fibroblasts	NF 1	0.10	193	CpG 8
Fibroblasts	NF 2	0.17	193	CpG 8
Fibroblasts	NF 3	0.23	193	CpG 8
Zebrafish	NZ 1	0.47	193	CpG 8

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.8 - Methylation rate values at CpG 10.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	NA	NA	CpG 10
Peripheral blood	AI 2	0.29	39	CpG 10
Peripheral blood	AI 3	0.41	39	CpG 10
Peripheral blood	AI 4	0.24	39	CpG 10
Peripheral blood	AI 5	0.35	39	CpG 10
Peripheral blood	AI 6	0.02	39	CpG 10
Peripheral blood	AI 7	0	39	CpG 10
Peripheral blood	AI 8	0.37	39	CpG 10
Fibroblast	AF 1	NA	NA	CpG 10
Fibroblast	AF 2	0	39	CpG 10
Fibroblast	AF 3	0.1	39	CpG 10
Zebrafish	AZ 1	0	39	CpG 10
Peripheral blood	NI 1	0.71	39	CpG 10
Peripheral blood	NI 10	0.16	39	CpG 10
Peripheral blood	NI 2	0.79	39	CpG 10
Peripheral blood	NI 3	0.3	39	CpG 10
Peripheral blood	NI 4	0.78	39	CpG 10
Peripheral blood	NI 5	0.78	39	CpG 10
Peripheral blood	NI 6	0.76	39	CpG 10
Peripheral blood	NI 7	0.3	39	CpG 10
Peripheral blood	NI 8	0.35	39	CpG 10
Peripheral blood	NI 9	0.41	39	CpG 10
Fibroblast	NF 1	0.36	39	CpG 10
Fibroblast	NF 2	0.27	39	CpG 10
Fibroblast	NF 3	0.24	39	CpG 10
Zebrafish	NZ 1	0.78	39	CpG 10

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.9 - Methylation rate values at CpG 11.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.95	113	CpG 11
Peripheral blood	AI 2	0.80	113	CpG 11
Peripheral blood	AI 3	0.89	113	CpG 11
Peripheral blood	AI 4	0.91	113	CpG 11
Peripheral blood	AI 5	0.89	113	CpG 11
Peripheral blood	AI 6	0.91	113	CpG 11
Peripheral blood	AI 7	0.91	113	CpG 11
Peripheral blood	AI 8	0.87	113	CpG 11
Fibroblast	AF 1	0.7	113	CpG 11
Fibroblast	AF 2	0.81	113	CpG 11
Fibroblast	AF 3	0.8	113	CpG 11
Zebrafish	AZ 1	0.84	113	CpG 11
Peripheral blood	NI 1	0.89	113	CpG 11
Peripheral blood	NI 2	NA	NA	CpG 11
Peripheral blood	NI 3	0.76	113	CpG 11
Peripheral blood	NI 4	NA	NA	CpG 11
Peripheral blood	NI 5	0.91	113	CpG 11
Peripheral blood	NI 6	NA	NA	CpG 11
Peripheral blood	NI 7	NA	NA	CpG 11
Peripheral blood	NI 8	0.89	113	CpG 11
Peripheral blood	NI 9	NA	NA	CpG 11
Peripheral blood	NI 10	0.86	113	CpG 11
Fibroblast	NF 1	0.81	113	CpG 11
Fibroblast	NF 2	0.78	113	CpG 11
Fibroblast	NF 3	0.74	113	CpG 11
Zebrafish	NZ 1	NA	NA	CpG 11

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.10 - Methylation rate values at CpG 12.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.41	206	CpG 12
Peripheral blood	AI 2	0.32	223	CpG 12
Peripheral blood	AI 3	0.43	206	CpG 12
Peripheral blood	AI 4	0.44	206	CpG 12
Peripheral blood	AI 5	0.37	206	CpG 12
Peripheral blood	AI 6	0.06	206	CpG 12
Peripheral blood	AI 7	0	206	CpG 12
Peripheral blood	AI 8	0.38	206	CpG 12
Fibroblast	AF 1	NA	NA	CpG 12
Fibroblast	AF 2	NA	NA	CpG 12
Fibroblast	AF 3	0	206	CpG 12
Zebrafish	AZ 1	0	206	CpG 12
Peripheral blood	NI 1	0.99	206	CpG 12
Peripheral blood	NI 2	0.9	206	CpG 12
Peripheral blood	NI 3	0	206	CpG 12
Peripheral blood	NI 4	0.82	223	CpG 12
Peripheral blood	NI 5	1	206	CpG 12
Peripheral blood	NI 6	0.80	206	CpG 12
Peripheral blood	NI 7	0.15	206	CpG 12
Peripheral blood	NI 8	0.25	206	CpG 12
Peripheral blood	NI 9	0.49	206	CpG 12
Peripheral blood	NI 10	0.13	206	CpG 12
Fibroblast	NF 1	0.18	206	CpG 12
Fibroblast	NF 2	0.23	206	CpG 12
Fibroblast	NF 3	NA	NA	CpG 12
Zebrafish	NZ 1	0.56	206	CpG 12

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Table L.11 - Methylation rate values at CpG 13.

Type of Sample	Individual	Methylation Rate	Position in the amplified sequence	CpG
Peripheral blood	AI 1	0.75	223	CpG 13
Peripheral blood	AI 2	NA	NA	CpG 13
Peripheral blood	AI 3	NA	NA	CpG 13
Peripheral blood	AI 4	0.7	223	CpG 13
Peripheral blood	AI 5	NA	NA	CpG 13
Peripheral blood	AI 6	NA	NA	CpG 13
Peripheral blood	AI 7	NA	NA	CpG 13
Peripheral blood	AI 8	NA	NA	CpG 13
Fibroblast	AF 1	1	223	CpG 13
Fibroblast	AF 2	NA	NA	CpG 13
Fibroblast	AF 3	NA	NA	CpG 13
Zebrafish	AZ 1	NA	NA	CpG 13
Peripheral blood	NI 1	0.97	223	CpG 13
Peripheral blood	NI 2	0.73	223	CpG 13
Peripheral blood	NI 3	NA	NA	CpG 14
Peripheral blood	NI 4	NA	NA	CpG 15
Peripheral blood	NI 5	NA	NA	CpG 16
Peripheral blood	NI 6	0.85	223	CpG 17
Peripheral blood	NI 7	NA	NA	CpG 18
Peripheral blood	NI8	NA	NA	CpG 19
Peripheral blood	NI 9	NA	NA	CpG 20
Peripheral blood	NI 10	0	223	CpG 21
Fibroblast	AF 1	NA	NA	CpG 13
Fibroblast	AF 2	NA	NA	CpG 13
Fibroblast	AF 3	NA	NA	CpG 13
Zebrafish	NZ 1	1	223	CpG 13

* AI = Affected individual; AF = Affected fibroblast; AZ = Affected zebrafish; NI = Normal individuals; NF = Normal fibroblasts; NZ = Normal zebrafish

Appendix M

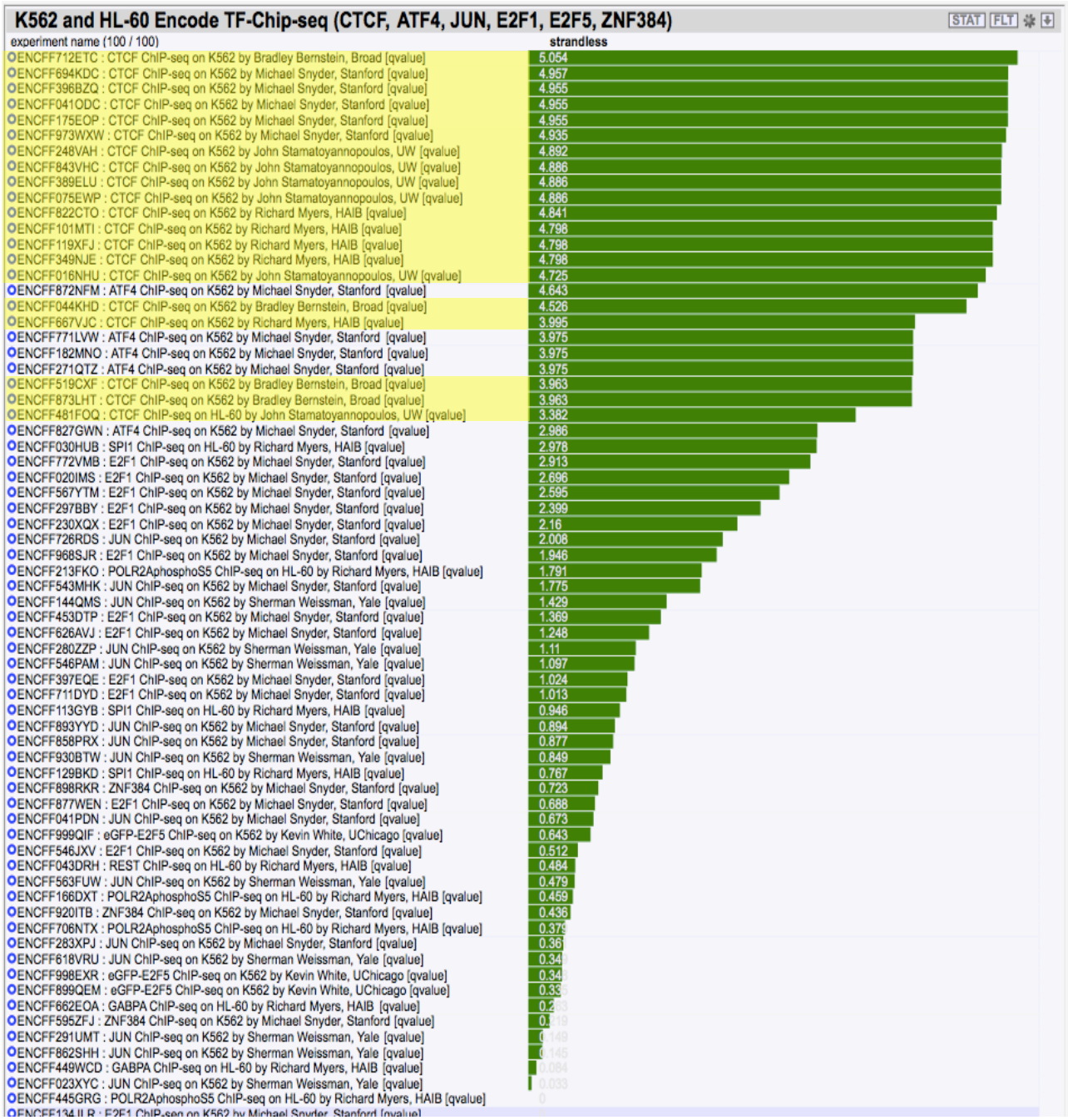

Figure 0.1 - Encode TF-ChIP-seq data available for (ATTTT)n repeat flanking region. The q-score (quality score – that shows the quality of the TF reads) are represented by the green bars. A higher q-value means that the sequencing of the immunoprecipitated region have smaller probability of errors.

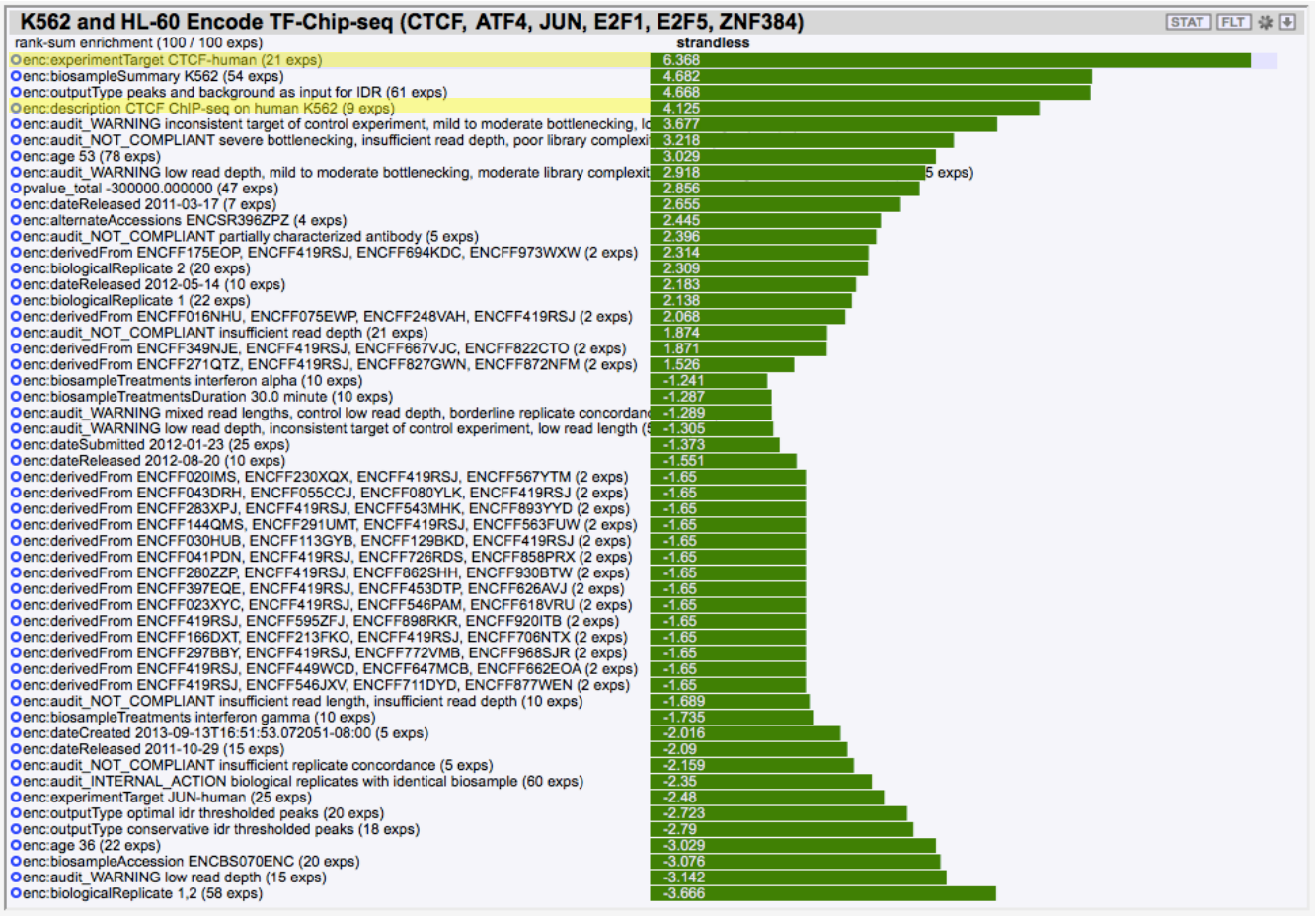

Figure 0.2 – Statistical analysis showing the Wilcox Mann-Whitney-U enrichment test results. Green bars represent the Z-score (according to the Encode definitions, a “high signal” or high enrichment in TF binding is defined by a Z-score superior to 1.64).

Appendix N

Table N.1 - Non-parametric Mann-Whitney test results for fibroblast cell lines.

			Methylation Rate	
Region	**CpG**	**Group (n)**	**Mean ± SD**	**p-value**
CTCF BS 1	CpG 1	Normal (n=3)	0.220 ± 0.13	p-value = 0.4
		Affected (n=3)	0.450 ± 0.27	
	CpG 2	Normal (n=3)	0.29 ± 0.06	p-value = 0.1
		Affected (n=3)	0.64 ± 0.02	
CpG island	CpG 4	Normal (n=3)	0.477 ± 0.09	p-value = 0.1
		Affected (n=3)	0.690 ± 0.10	
	CpG 5	Normal (n=3)	0.487 ± 0.07	p-value = 0.164
		Affected (n=3)	0.590 ± 0.05	
	CpG 6	Normal (n=3)	0.777 ± 0.10	p-value = 0.7
		Affected (n=3)	0.763 ± 0.19	
	CpG 7	Normal (n=3)	0.710 ± 0.04	p-value = 0.1
		Affected (n=3)	0.853 ± 0.02	
	CpG 8	Normal (n=3)	0.167 ± 0.07	p-value = 0.1
		Affected (n=3)	0.497 ± 0.10	
CTCF BS 2	CpG 10	Normal (n=3)	0.290 ± 0.06	p-value = 0.2
		Affected (n=3)	0.05 ± 0.07	
	CpG 11	Normal (n=3)	0.777 ± 0.04	p-value = 1
		Affected (n=3)	0.770 ± 0.06	
	CpG 12	Normal (n=3)	0.205 ± 0.04	p-value = 0.667
		Affected (n=3)	0 ± 0	
	CpG 13	Normal (n=3)	NA	p-value = 0.667
		Affected (n=3)	1 ± 0	