Universidade de Aveiro Departamento de Ciências Médicas

2021

Ângela Sofia Moreira Marques

Investigação da tradução independente de AUG de um pentanucleótido repetitivo em SCA37

Investigation of non-AUG dependent pentanucleotide repeat translation in SCA37

Universidade de Aveiro Departamento de Ciências Médicas

2021

Ângela Sofia Moreira Marques

Investigação da tradução independente de AUG de um pentanucleótido repetitivo em SCA37

Investigation of non-AUG dependent pentanucleotide repeat translation in SCA37

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de mestre em Biomedicina Molecular, realizada sob a orientação científica da doutora Joana Loureiro, investigadora do grupo *Genetics of Cognitive Dysfunction* do Instituto de Investigação e Inovação em Saúde – Universidade do Porto, da doutora Gabriela Moura, professora auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro e da doutora Isabel Silveira, líder do grupo *Genetics of Cognitive Dysfunction* do Instituto de Investigação e Inovação em Saúde – Universidade do Porto.

This work was financed by FEDER - Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by Portuguese funds through FCT - Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior in the framework of the project POCI-01-0145-FEDER-029255 (PTDC/MED-GEN/29255/2017).

o júri

presidente	Professor Doutor Ramiro Daniel Carvalho de Almeida Professor Auxiliar, Departamento de Ciências Médicas, Universidade de Aveiro
vogal - arguente principal	Professora Doutora Paula Maria Vieira Jorge Investigadora Principal, Centro de Genética Médica Jacinto Magalhães - Porto
vogal - orientador	Doutora Joana Maria Geraldes da Rocha Loureiro Investigadora Júnior, Instituto de Investigação e Inovação em Saúde (i3S) - Universidade do Porto

agradecimentos

À doutora Joana Loureiro: obrigada por me guiares com paciência, rigor e empatia, e por me incentivares a fazer um trabalho cada vez melhor. Vou levar os teus ensinamentos e o teu exemplo para a vida.

À doutora Isabel Silveira: obrigada pela confiança e pela oportunidade de integrar o seu grupo de investigação, por acompanhar o meu trabalho e melhorá-lo com o seu conhecimento.

À doutora Gabriela Moura: obrigada pela ajuda a resolver as questões burocráticas com a Universidade de Aveiro.

Às minhas colegas de laboratório, Sofia Figueiredo e Filipa Castro: obrigada pela companhia, pelas dúvidas esclarecidas e por ajudarem a caloira a orientar-se no laboratório (e no i3S em geral).

A tantos outros investigadores do i3S, sobretudo à doutora Mariana Graça do grupo UnIGENe: obrigada pela troca de conhecimentos e materiais que fizeram toda a diferença no desenvolvimento deste trabalho.

Ao grupo de investigação Stem Cells & Neurogenesis liderado pelo doutor Diogo Castro: obrigada por partilharem a linha de "human neural stem cells" e os protocolos usados na sua cultura.

Aos meus pais, Deolinda e Carlos: obrigada por tudo, em especial por terem como lema "Pela tua saúde e pela tua educação, fazemos tudo".

Ao meu irmão Miguel, à minha cunhada Carla, e à minha madrinha Andreia: obrigada pelo vosso apoio, presença e conselhos ao longo de toda ou grande parte da minha vida, sobretudo durante os últimos meses.

Ao meu namorado, Pedro: obrigada por seres um bom cheerleader, por partilhares do meu entusiasmo quando as coisas correm bem e ouvires os meus lamentos com paciência quando correm mal.

Aos meus amigos, de Lordelo e de Aveiro: obrigada por estarem presentes mesmo quando estão longe e por acreditarem mais em mim do que eu mesma.

A todas as pessoas que, de alguma forma, me ajudaram nesta jornada e não estão nesta página, saibam que estão no meu coração. O meu sincero obrigada.

palavras-chave

Neurodegeneração, ataxias espinocerebelosas, ataxia espinocerebelosa tipo 37 (SCA37), expansões de repetição microssatélite, doenças causadas por expansão de repetições, doenças causadas por de inserção de $(ATTTC)_n$, toxicidade mediada por RNA, tradução associada a repetições independente de AUG (tradução RAN)

resumo

As ataxias espinocerebelosas (SCAs) são um grupo de doenças neurodegenerativas geralmente com início tardio, para as quais não há cura. As SCAs caracterizam-se por degeneração cerebelar progressiva que leva ao desequilíbrio da marcha, perda de coordenação dos membros e disartria, sintomas geralmente associados a demência, epilepsia e enxaqueca. A SCA37 é uma doença hereditária autossómica dominante causada por uma inserção de $(ATTTC)_n$ num intrão localizado na 5'-UTR do gene *DAB1*. Em Portugal e Espanha, há centenas de indivíduos afetados ou em risco de ter SCA37. Indivíduos não afetados têm dois alelos com 7-400 repetições de ATTTT, enquanto indivíduos afetados têm um alelo com a configuração $[(ATTTT)_{60–79}(ATTTC)_{31–75}(ATTTT)_{58–90}]$. O RNA $(AUUUC)_n$ forma agregados nucleares em linhas celulares humanas transfetadas e é tóxico quando injetado em embriões de zebrafish, o que indica que este RNA repetitivo inicia uma cascata de mecanismos patogénicos que ainda não são totalmente compreendidos. Muitas doenças causadas por transcritos repetitivos caracterizam-se pela tradução não canónica da expansão repetitiva, por um mecanismo chamado tradução associada a repetições independente de AUG (RAN). A tradução RAN leva à produção de polipéptidos repetitivos a partir das três fases de leitura da expansão repetitiva, e estes polipéptidos geralmente são tóxicos e contribuem para o fenótipo neurodegenerativo. Como tal, neste trabalho, o meu objetivo foi investigar se a inserção repetitiva $(ATTTC)_n$ que causa a SCA37 é aberrantemente traduzida por RAN.

Para tal, gerei diferentes construtos para deteção da tradução RAN a partir das três fases de leitura do $(ATTTC)_n$, sendo que todas elas codificavam poli(ISFHF). Primeiro, clonei o alelo patogénico da SCA37 e alelos não patogénicos em vetores de expressão de mamíferos, depois inseri 2 codões de terminação em cada fase de leitura a montante da sequência repetitiva e 3 epitopos (HA, myc e flag), um em cada fase de leitura, a jusante da sequência repetitiva. Confirmei *in silico* que os vetores desenvolvidos eram capazes de levar à tradução do pentapéptido RAN putativo fundido com um epitopo. Transfetei células estaminais neurais humanas (hNSC) e HEK293T com estes vetores e confirmei que o RNA $(AUUUC)_n$ era transcrito, por hibridização fluorescente *in situ*. Por fim, analisei a produção de poli(ISFHF) 72h pós-transfeção, por dot blot e western blot com anticorpos α-HA, α-myc, α-flag, e α-poli(ISFHF).

Apesar de ter usado diferentes linhas celulares humanas e vetores, não detetei péptidos RAN traduzidos *in vitro* a partir de nenhuma das fases de leitura da inserção $(ATTTC)_n$. Uma vez que a SCA37 é uma doença de manifestação tardia, talvez sejam necessárias décadas de acumulação em neurónios cerebelares para o poli(ISFHF) se tornar detetável, o que não é possível mimetizar com ensaios de transfeção em linhas celulares proliferativas.

keywords

Neurodegeneration, spinocerebellar ataxia, spinocerebellar ataxia type 37 (SCA37), microsatellite repeat expansions, repeat expansion disease, $(ATTTC)_n$ insertion disease, RNA-mediated toxicity, repeat associated non-AUG dependent (RAN) translation

abstract

Spinocerebellar ataxias (SCAs) are a heterogeneous group of neurodegenerative diseases, usually with late onset, for which there is no cure. SCAs are mainly characterized by progressive cerebellar degeneration leading to symptoms of gait imbalance, limb incoordination and dysarthria, often associated with dementia, epilepsy and migraine. SCA37 is an autosomal dominant inherited disease caused by an $(ATTTC)_n$ insertion in a 5'-UTR intron of *DAB1* gene. SCA37 is a disease with hundreds of affected and at-risk individuals in Portugal and Spain. Unaffected individuals have two $(ATTTT)_n$ alleles ranging from 7-400 repeat units, while affected individuals carry one allele with the configuration $[(ATTTT)_{60–79}(ATTTC)_{31–75}(ATTTT)_{58–90}]$. The $(AUUUC)_n$ RNA forms nuclear aggregates in transfected human cell lines and is toxic when injected in zebrafish embryos, indicating that this repetitive RNA mediates a cascade of pathogenic mechanisms not yet fully understood. Many diseases caused by pathogenic transcribed repeats are characterized by non-canonical translation of the repeat expansion, by a mechanism known as repeat associated non-AUG dependent (RAN) translation. RAN translation leads to the production of repetitive polypeptides from the three reading frames of the repeat expansion, which are often toxic being implicated in the neurodegenerative phenotype. Therefore, in this work, I aimed to investigate if the $(ATTTC)_n$ insertion causing SCA37 is abnormally translated by RAN.

To achieve this aim, I engineered different constructs to detect RAN translation from the three $(ATTTC)_n$ reading frames, all encoding poly(ISFHF). I cloned the SCA37 pathogenic and nonpathogenic alleles in mammalian expression vectors, then I placed 2 stop codons in each frame upstream the repeat and 3 epitopes (HA, myc and flag), one in each reading frame, downstream the repeat. I confirmed *in silico* that the developed constructs were able to lead to the translation of the putative RAN pentapeptide fused with a tag epitope. Then, I transfected HEK293T and human neural stem cells (hNSC) with these constructs and I showed that the $(AUUUC)_n$ RNA was being transcribed, by fluorescence *in situ* hybridization. After this, I analysed the production of poly(ISFHF) 72h post-transfection, by dot blot and western blot using α-HA, α-myc, α-flag and α-poly(ISFHF) antibodies.

Despite having used different human cell lines and genetic backbones, I did not detect SCA37 RAN pentapeptides translated *in vitro* from any frame of the $(ATTTC)_n$ insertion. Since SCA37 is a late-onset disease, poly(ISFHF) may require decades of accumulation in cerebellar neurons to become detectable, which is not mimicked in transfection assays in proliferative cell lines.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1. Repetitive DNA sequences in the human genome

Approximately a half of the human genome is composed of repetitive DNA sequences[1]. Depending on their structure and ability to move in genome, from one place to another, these repetitive sequences are classified into interspersed repeats or tandem repeats (Figure 1).

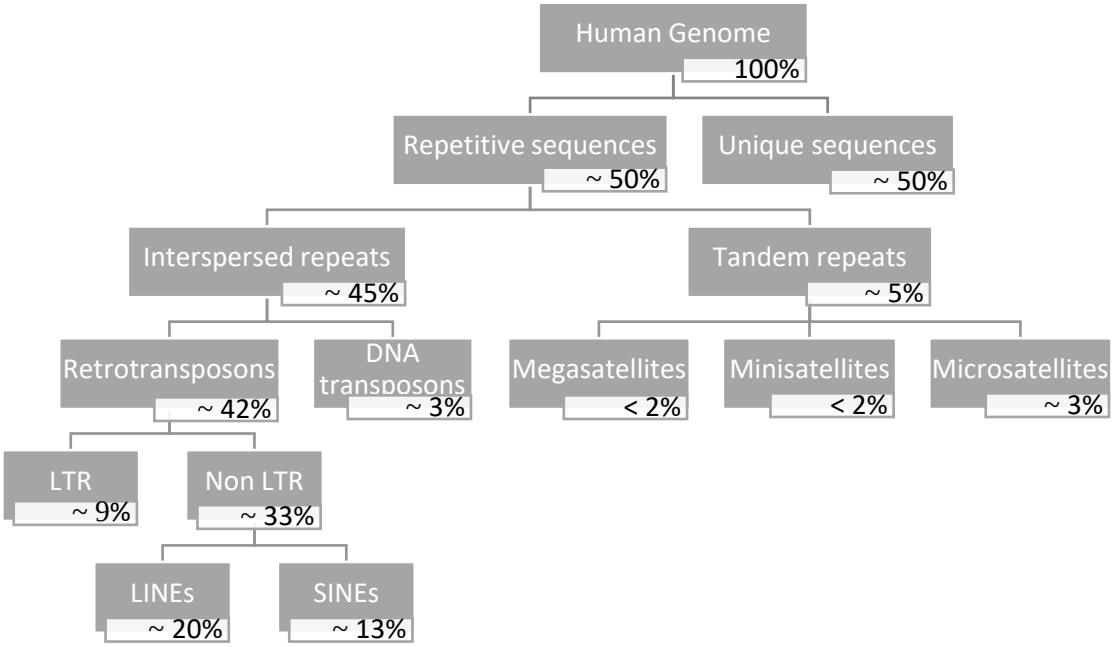


Figure 1. Schematic diagram of the classes of DNA sequences and respective contribution (%) for the human genome[1]. Approximately a half of the human genome is composed of repetitive DNA sequences. Repetitive DNA can be classified as interspersed or tandem repeats. Interspersed repeats include both retrotransposons and DNA transposons. Retrotransposons are divided into LTRs (long terminal repeats) and non-LTRs. Non-LTRs can be further divided into LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements). Tandem repeats include mega-, mini- and microsatellites.

Interspersed repeats, or transposable elements, are repetitive DNA sequences found scattered through the genome, due to their ability to jump from one location to another[1]. Depending on the mechanism used to move in the genome, interspersed repeats are divided into transposons and retrotransposons[1].

The transposons are elements that use a “cut-and-paste” mechanism to change their location in the genome. These elements encode an enzyme named transposase that excises the element from the original position and reinserts it in a new location in the

genome[1]. Although these elements compose approximately 3% of the human genome, they are inactive, as they have lost their ability to transpose[1].

The retrotransposons or retroelements are the biggest subclass of interspersed repeat elements. These elements use a “copy-and-paste” mechanism to move in genome[1]. In this mechanism, the element is first transcribed, then the RNA is reverse transcribed into complementary DNA (cDNA) that is later inserted in a new location in the genome[1]. Contrarily to transposition that does not enable the multiplication of the elements, the retrotransposition increases the number of retrotransposons in the genome. Mammalian retrotransposons are subdivided into LTRs (long-terminal repeats) and non-LTRs. The non-LTR retrotransposons are classified as autonomous or non-autonomous, depending on their ability to encode the enzymes required for retrotransposition[1]. The LINEs (long interspersed nuclear elements), as the L1 element, are autonomous elements whereas the SINEs (short interspersed nuclear elements), as Alu elements, are non-autonomous elements[1]. The non-autonomous repeats use the enzymes encoded by the autonomous elements to retrotranspose[1]. Although there are many human retrotransposons, only L1, some Alu and a small group of LTR elements, called human endogenous retroviruses, retain the ability to retrotranspose[1]. The remaining elements are not active and are thus referred as fossil retrotransposons[1]. Alu elements are a primate specific class of retrotransposons[1]. There are more than 1 million Alus in the human genome[2]. Alus are divided according to their age in 3 families, AluJ is the oldest, AluS is the intermediate and AluY is the youngest family[3]. From these families, only AluY and a few subfamilies of AluS remain active in the human genome[3].

Tandem repeats are DNA sequences that are repeated several times, each motif following the previous. Based on total repeat length, tandem repeats are divided into megasatellites, minisatellites and microsatellites[1]. Megasatellites are the longest tandem repeated sequences. These repeats consist in blocks of hundreds of kilobases (kb) found mostly in centromeric heterochromatin[1]. Minisatellites, also called variable number tandem repeats (VNTR), are tandemly repeated DNA motifs typically with 10-50 base pairs (bp), ranging from 100 bp to 20 kb[1]. Microsatellites, also called short tandem repeats (STRs), are tandemly repeated DNA motifs typically with 2-10 bp spanning approximately up to 100 bp in length[1, 4].

Microsatellites are highly polymorphic repeats, meaning that different repeat size alleles may be found in a given population[5-7]. The polymorphic nature of microsatellites makes them extremely useful as genetic markers to be used in population genetics, gene mapping and forensic genetics[8-10].

2. Diseases caused by microsatellite repeat expansions and insertions

In 1991, the discovery of the trinucleotide repeat expansions causing spinal and bulbar muscular atrophy (SBMA) and fragile X syndrome (FXS) set the stage for the identification of more than 40 pathogenic repeat expansions in the following 3 decades[11] (Figure 2). These repeat expansions of tri-, tetra-, penta- and hexanucleotides are found in both coding and noncoding gene regions (Figure 2). In loci associated with repeat expansions, unaffected individuals usually carry short repeat allele sizes whereas affected individuals have alleles expanded over a threshold established for each disease[11].

In 2009, the discovery of the genetic cause of spinocerebellar ataxia type 31 (SCA31) showed that tandem repeat loci are not only associated with disease by the expansion of the polymorphic repeat, but also by the insertion of a new repetitive motif in an ancestral microsatellite[12]. SCA31 is caused by a TGGAA repeat insertion in an intronic region shared by thymidine kinase 2 (*TK2*) and brain expressed, associated with Nedd4 (*BEAN*) genes[12]. Individuals affected with SCA31 carry a 2.5–3.8 kb long repeat insertion containing $(TGGAA)_n$ whereas unaffected individuals have polymorphic pentanucleotide repeats that do not contain the $(TGGAA)_n$ pathogenic repetitive motif[12]. Recently, 7 other neurological diseases were found caused by pentanucleotide repeat insertions[13-18] (Figure 2). Interestingly, all these 7 diseases: SCA37, FAME1 (familial adult myoclonic epilepsy 1), FAME2, FAME3, FAME4, FAME6 and FAME7 are caused by a similar noncoding $(ATTTC)_n$ insertion in the middle or adjacent to normal polymorphic ATTTT repeats located in *DAB1* (DAB reelin adaptor protein 1), *SAMD12* (sterile alpha motif domain containing 12), *STARD7* (StAR related lipid transfer domain containing 7), *MARCHF6* (membrane associated ring-CH-type finger 6), *YEATS2* (YEATS domain containing 2), *TNRC6A* (trinucleotide repeat containing adaptor 6A) and *RAPGEF2* (Rap guanine nucleotide exchange factor 2) genes, respectively[13-18]. In addition to $(ATTTC)_n$ insertions, FAME1 was also associated with a $(ATTTG)_n$ insertion in *SAMD12,* in a large Chinese family[19].

Contrarily to normal alleles, that usually maintain their repeat size, pathogenic repeat expansions are characterized by repeat size variation between cells and tissues of the same individual (somatic instability) and when transmitted from parents to offspring (intergenerational instability)[11]. Myotonic dystrophy type 1 (DM1) is an example of a repeat expansion disease where somatic instability is well studied. DM1 is caused by a CTG repeat expansion ($(CTG)_{exp}$) located in the 3'-untranslated region (3'-UTR) of Dystrophia Myotonic Protein kinase (*DMPK)* gene[20, 21]. Unaffected individuals have 5-30 repeats and affected individuals have one allele with more than 50 to several thousand repeats[20, 21]. In this

disease, the $(CTG)_{exp}$ shows both inter- and intratissue instability. Interestingly, this somatic instability tends to increase with age, with older patients showing larger $(CTG)_{exp}$ size variation[22, 23]. Castel and colleagues compared $(CTG)_{exp}$ length in several tissues derived from DM1 affected fetuses and reported that the $(CTG)_{exp}$ was larger in the heart than in skeletal muscle, kidney, skin, brain and liver[22].

The expanded repeats usually expand when transmitted from an affected parent to the affected offspring[11]. Redman and colleagues analyzed the variation of $(CTG)_{exp}$ size in 110 DM1 parent-offspring transmissions and showed that 101 transmissions led to expansions, 5 to contractions and 4 did not show a significant change in $(CTG)_{exp}$ size[24]. Interestingly, the $(CTG)_{exp}$ is more unstable when the mother is the transmitting parent[24].

The repeat expansion size is often inversely correlated with the age of disease onset. An inverse correlation between the repeat expansion size and the age of disease onset has been identified in many repeat expansion diseases, including DM1, SCA1, FAME1 and FAME3[15, 17, 24, 25]. In DM1, parents with hundreds of CTG repeats and a mild, adult-onset form of the disease often have offspring with thousands of CTG repeat units that have a more severe form of the disease with childhood onset, that can be fatal in a few years[24].

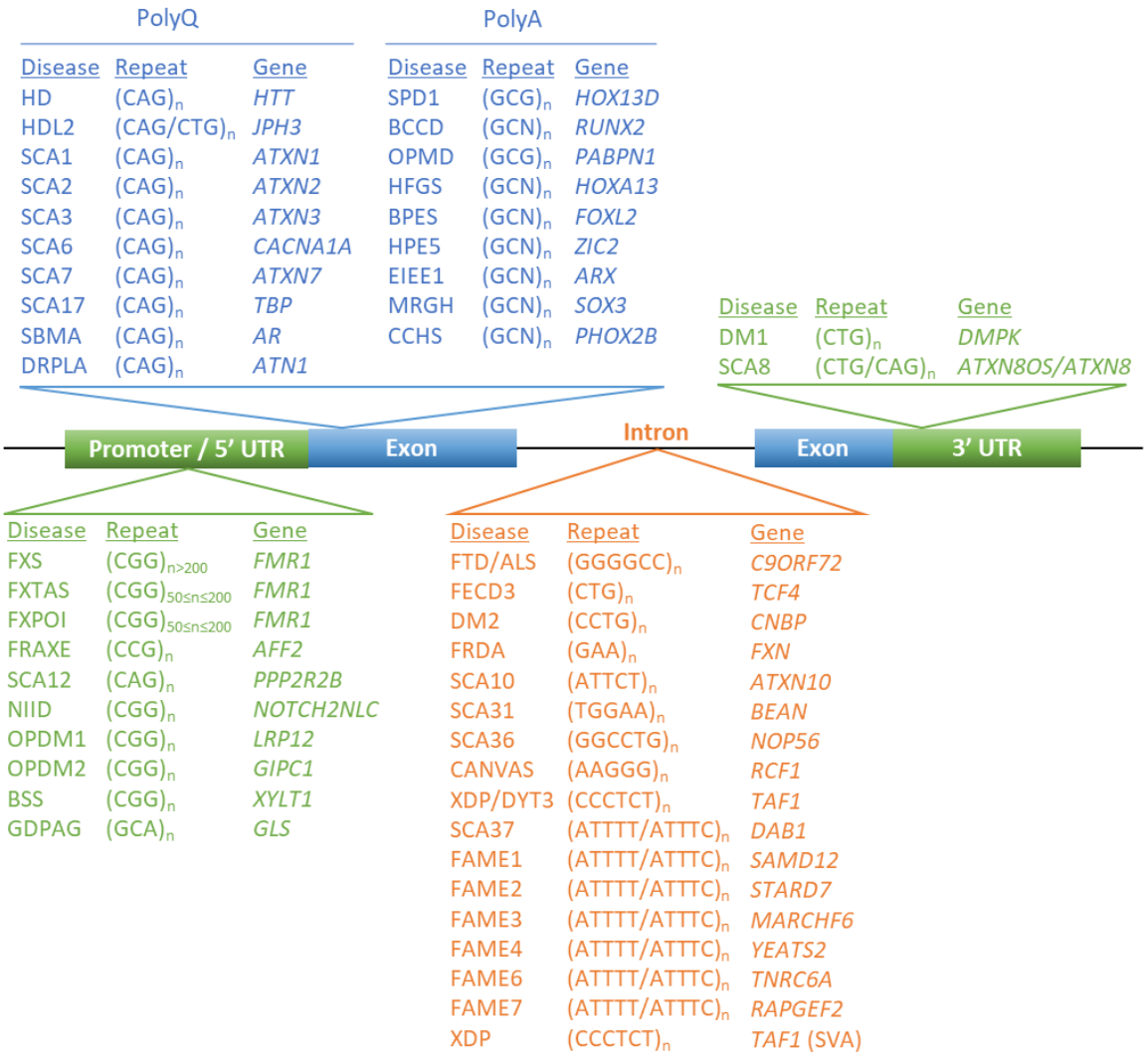


Figure 2. Location of pathogenic repeat expansions and insertions within genes. Repeat expansions may be found in gene coding regions, encoding polyglutamine (polyQ) or polyalanine (polyA) tracts, or in noncoding gene regions. Noncoding repeat expansions are located in promoters, 5'-untranslated regions (5'-UTRs), introns and 3'-UTRs. BCCD - brachydactyly and cleidocranial dysplasia; BSS - Baratela-Scott syndrome; CANVAS - cerebellar ataxia, neuropathy and vestibular areflexia syndrome; CCHS - congenital central hypoventilation syndrome; DM1 -myotonic dystrophy type 1; DM2 - myotonic dystrophy type 2; DRPLA - dentatorubral-pallidoluysian atrophy; EIEE1 - early infantile epileptic encephalopathy type 1; FAME - familial adult myoclonic epilepsy; FECD3 - Fuchs endothelial corneal dystrophy type 3; FRAXE - fragile XE syndrome; FRDA - Friedreich ataxia; FTD/ALS - frontotemporal dementia / amyotrophic lateral sclerosis; FXS - fragile X syndrome; FXTAS - fragile X-associated tremor ataxia syndrome; GDPAG - global developmental delay, progressive ataxia and elevated glutamine; HD – Huntington's disease; HDL2 - Huntington disease-like 2; HFGS - hand-foot-genital syndrome; HPE5 - holoprosencephaly type 5; MRGH - mental retardation with isolated growth hormone deficiency; OPMD - oculopharyngeal muscular dystrophy; NIID - neuronal intranuclear inclusion disease; OPDM1 - oculopharyngodistal myopathy type 1; OPDM2 - oculopharyngeal muscular dystrophy type 2; OPML1 - oculopharyngeal myopathy with leukoencephalopathy type 1; SBMA - spinal and bulbar muscular atrophy; SPD1 - synpolydactyly type 1; SCA - spinocerebellar ataxia; SVA - SINE-VNTR-Alu retrotransposon; XDP - X-linked dystonia parkinsonism. Adapted from *Depienne and Mandel, 2021*[11].

3. Pathogenic mechanisms associated with microsatellite repeat expansions and insertions

Depending on the repeat motif and location in a gene, there is evidence that repeat expansions may be associated with different pathogenic mechanisms[11]. In protein coding regions, only trinucleotide repeats were identified[11] (Figure 2). These repeats are translated in toxic repetitive peptides leading to protein gain-of-function[11]. Repeats located in noncoding regions may trigger disease by different mechanisms, for example, they can induce epigenetic changes that lead to gene silencing or gene loss-of-function, or they may be transcribed and initiate a cascade of RNA-mediated toxicity and/or repeat associated non-AUG dependent (RAN) translation[11] (Figure 3).

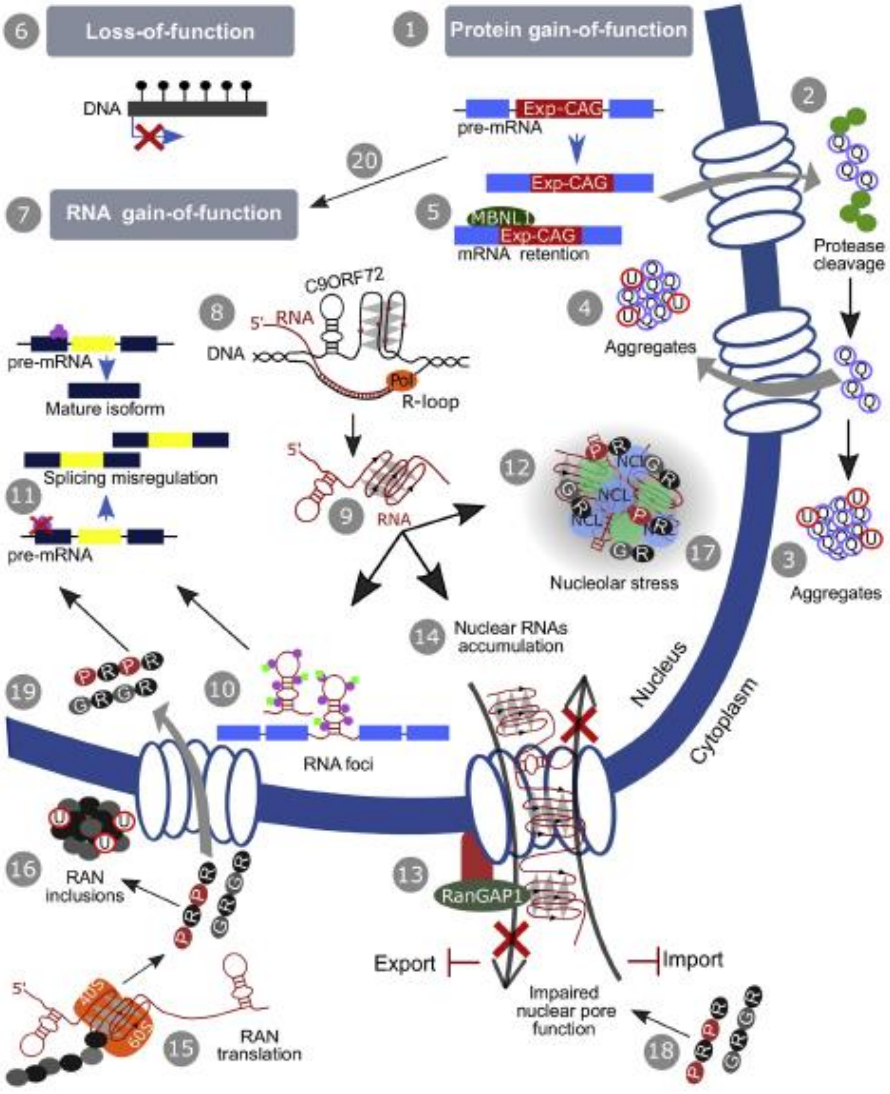


Figure 3. Pathogenic mechanisms associated with repeat expansion disorders. Three main disease mechanisms are associated with repeat expansions. (1) In coding regions, repeat expansions originate homopolymeric stretches of 1 amino acid, such as glutamine (Q), leading to protein gain-of-function. (2) The protein with the polyQ stretch can be cleaved and the polyQ tract (3) aggregates in the cytoplasm or is (4) imported to the nucleus. Cytoplasmic and nuclear aggregates are ubiquitin (U) positive and toxic to the neuronal cell. (5) On the other hand, the expanded CAG RNA has increased interaction with MBLN1, resulting in its nuclear retention and decreased protein from translation. In noncoding regions, repeat expansions lead to 2 exclusive pathogenic mechanisms; (6) gene loss-of-function due to hypermethylation and consequent gene silencing as in FRDA and FXS or (7) RNA gain-of-function, which includes several parallel mechanisms. (8) For the GGGGCC repeat expansion in *C9ORF72*, the hexanucleotide forms G-quadruplex structures in DNA and DNA:RNA hybrids (R-loop). (9) The G-quadruplexes are also detected in the repetitive expanded RNA. (10) In all noncoding RNA-mediated disorders, the expanded RNA forms RNA foci due to recruitment of RBPs (green squares and purple circles). (11) The consequent sequestration of RBPs to RNA foci leads to alternative splicing misregulation of other neuronal mRNAs. (12) In expanded *C9ORF72*, the repetitive RNA forms secondary structures able to recruit nucleolin (NCL), leading to nucleolar stress, that compromises rRNA biogenesis. (13) The *C9ORF72* hexanucleotide RNA can also sequester the RanGAP1 and interact with other proteins of the nuclear pore complex, disrupting nucleocytoplasmic transport and causing (14) accumulation of other mRNAs in the nucleus. (15) Repeat-containing transcripts can escape from the nucleus to the cytoplasm and be translated by RAN, leading to the synthesis of repetitive peptides, such as the dipeptide species of poly(PR) and poly(GR) from the *C9ORF72* repeat expansion. These RAN peptides are able to (16) aggregate in the cytoplasm, (17) cause nucleolar stress, (18) compromise nucleocytoplasmic transport and/or (19) impair mRNA splicing of other neuronal mRNAs. (20) For coding repeat expansion diseases such as SCA3/MJD and Huntington's disease, in addition to protein gain-of-function, RNA-mediated toxicity has also been detected. Adapted from Loureiro *et al*, 2016[26].

3.1. Protein gain-of-function

In coding regions, only CAG or GCN trinucleotide repeat expansions were found causing disease[11] (Figure 2). These trinucleotide repeats are translated in abnormal homopolymeric stretches of glutamines or alanines[11].

In polyglutamine (polyQ) diseases, like Huntington's disease (HD), spinocerebellar ataxia type 3/Machado-Joseph disease (SCA3/MJD) and SBMA, the abnormal polyQ tracts cause misfolding and mislocalization of mutant proteins, that adopt abnormal conformations more stable than their native conformation forming insoluble nuclear aggregates[27]. Such aggregates are usually marked with a polyubiquitin chain and targeted for proteasomal degradation, but the stability conferred by the polyQ tract inhibits their degradation[28]. Nuclear aggregates of polyQ-expanded proteins sequester other nuclear proteins, mostly transcription factors (TFs) involved in epigenetic regulation, preventing those TFs from performing their usual functions. This sequestration causes transcriptional and epigenetic abnormalities that induce cellular toxicity[29]. Interestingly, many TFs are found sequestered by polyQ aggregates in different diseases, as is the case of the CREB-binding protein (CBP) that is sequestered by polyQ aggregates in HD, SCA3/MJD and SBMA[30-32]. This indicates that this sequestration of TFs is specifically induced by the expanded polyQ tracts. CBP is

a transcriptional co-activator with histone acetyltransferase activity, when sequestered by polyQ-expanded proteins CBP can no longer perform its normal function, which leads to histone deacetylation and formation of repressive heterochromatin that disrupts transcription leading to cell death[30-32].

There are 9 diseases caused by polyA expansions, 8 of which are congenital[33]. PolyA diseases are associated with a smaller number of repeats when compared to polyQ diseases, in polyQ diseases the pathogenic thresholds range from 20 to 55 CAG repeats, whereas the pathogenic thresholds in polyA diseases range from 12 to 27 GCN repeats[11]. Eight out of the 9 known polyA repeat expansions occur in genes encoding developmental TFs[33]. In these diseases, the presence of an expanded polyA tract induces conformational changes in the TFs that lead to a decrease in their ability to bind DNA, with consequent complete or partial loss of transcription activity of the target genes during development[33]. Oculopharyngeal muscular dystrophy (OPMD) is unique in many ways, this disease is caused by a $(GCG)_{exp}$ of 12 to 17 repeat units in the Poly(A) binding protein nuclear 1 (*PABPN1*) gene, which encodes a polyadenylation factor that is involved in polyadenylated (poly(A)) RNA nucleocytoplasmic exportation and transcription regulation[11, 34, 35]. OPMD is the only polyA disease with adult onset, the only in which the polyA expansion does not occur in a TF and one of the few in which mutant protein aggregates have been observed *in vivo,* suggesting protein gain-of-function as the pathogenic mechanism[35]. Interestingly, *in vitro* experiments have shown that all 9 polyA expanded proteins can form cytoplasmic aggregates, and in brachydactyly and cleidocranial dysplasia (BCCD), hand-foot-genital syndrome (HFGS) and synpolydactyly type 1 (SPD1) cell models, these polyA aggregates sequester other polyA-containing proteins, including their cognate wild-type forms[36].

3.2. Gene loss-of-function

In diseases caused by repeat expansions in noncoding gene regions, the expanded repeats may lead to gene loss-of-function[11]. In this mechanism, the repeat expansion induces epigenetic alterations in CpGs and/or histones that silence the transcription of the mutant gene[37, 38]. This is the pathogenic mechanism that underlies FXS and Friedreich's ataxia (FRDA)[11].

FXS is an X-linked disease caused by a $(CGG)_{exp}$ over 200 repeat units in the 5'-UTR of the FMRP translational regulator 1 (*FMR1*) gene[39, 40]. This expansion leads to the methylation of CpG dinucleotides in the repeat that extends to the promoter of the gene, leading to the silencing of *FMR1* transcription and consequently to the absence of fragile X

mental retardation protein (FMRP)[41]. FMRP is an RNA-binding protein that regulates the transport and translation of several neuronal messenger RNAs (mRNAs), thus the loss of FMRP results in dysregulated neuronal pathways leading to cognitive impairment in fragile-X individuals[41, 42]. The CpG hypermethylation in *FMR1* is also accompanied by the hypermethylation of histones H3 and H4 promoting the formation of heterochromatin, a repressive chromatin conformation that impairs initiation and elongation of transcription by blocking the binding of TFs[41].

The most common cause of FRDA is a $(GAA)_{exp}$ in the first intron of the frataxin (*FXN*) gene[43]. FRDA is an autosomal recessive disease caused by the loss of frataxin protein[44]. Frataxin is a mitochondrial protein that regulates iron transport and respiration, thus the loss of the frataxin leads to oxidative stress, mitochondrial iron accumulation and neuronal death[44]. *FXN* nonpathogenic alleles have fewer than 36 GAA repeats whereas pathogenic alleles have 56-1345 GAA repeats[43, 45, 46]. The $(GAA)_{exp}$ triggers hypoacetylation and hypermethylation of the histone H3, which promotes the formation of heterochromatin leading to *FXN* transcription reduction[44, 46-49]. The *FXN* silencing is also triggered by the ability of the $(GAA)_{exp}$ tract to adopt abnormal secondary structures or form DNA/RNA hybrids that interfere with transcription by stalling RNA polymerase II[11, 50].

3.3. RNA-mediated toxicity

Diseases caused by transcribed noncoding repeat expansions are associated with an RNA-mediated toxicity mechanism[11]. There are many parallel cellular mechanisms through which noncoding transcribed repeats trigger toxicity, including 1) formation of nuclear RNA aggregates called RNA foci, 2) sequestration of RNA-binding proteins (RBPs) with important roles in mRNA metabolism, in RNA foci, with consequent misregulation of mRNA alternative splicing or polyadenylation and 3) impairment in nucleocytoplasmic transport[51].

DM1 is an example of a disease caused by RNA-mediated toxicity. The $(CUG)_{exp}$ RNA adopts non-canonical secondary structures that promote its aggregation in RNA foci[52]. To form RNA foci, the $(CUG)_{exp}$ RNA sequesters RBPs, including proteins from the muscleblind-like (MBNL) family that recognize the CUG motif[52, 53]. Furthermore, the $(CUG)_{exp}$ RNA also induces hyperphosphorylation of CUGBP Elav-like family member 1 (CELF1)[52].

The muscleblind-like protein 1 (MBNL1) regulates mRNA alternative splicing, promoting the skipping of fetal exons and expression of adult protein isoforms, the retention

of MBNL1 in RNA foci leads to the loss-of-function of this protein with consequent missplicing of the target mRNAS[37, 52, 54-57]. The CELF1 is an RBP that regulates mRNA alternative splicing antagonistically to MBNL proteins by promoting the inclusion of fetal exons in target mRNAs in embryonic and neonatal heart tissues[58]. The hyperphosphorylation of CELF1 increases nuclear CELF1 activity, which also results in the target mRNA missplicing[52, 57]. Target mRNAs of MBNL1 and CELF1 include chloride voltage-gated channel 1 (*CLCN1*), insulin receptor (*INSR*), bridging integrator 1 (*BIN1*) and troponin T2, cardiac type (*TNNT2*)[59]. The missplicing of these mRNAs leads to the myotonia, insulin insensitivity, muscle weakness and reduced myocardial function observed in DM1 affected individuals[52, 57, 60]. To date, MBNL1 sequestration and CELF1 upregulation are known to cause missplicing of over 30 mRNAs in the heart, brain and skeletal muscle from DM1 affected individuals[59].

In *C9ORF72* frontotemporal dementia/amyotrophic lateral sclerosis (*C9ORF72* FTD/ALS), the $(GGGGCC)_{exp}$ forms RNA foci and sequesters Ran GTPase-activating protein 1 (RanGAP) and nucleoporins, proteins required for efficient nucleocytoplasmic transport through the nuclear pore complex (NPC), thus impairing nucleocytoplasmic transport[61]. The impairment of nucleocytoplasmic transport leads to the mislocalization of nuclear and cytoplasmic proteins that are essential for cell survival inhibiting their function and causing cell death[61]. In *Drosophila*, overexpression of RanGAP suppresses $(GGGGCC)_{exp}$-mediated neurodegeneration whereas knockdown of RanGAP enhances neurodegeneration[61].

Some repeat expansions are bidirectionally transcribed and both sense and antisense repeats can contribute to the disease[62]. Bidirectional transcription was reported for the first time in DM1, in 2005[63]. Presently, bidirectional transcription was already detected in 14 diseases caused by microsatellite repeat expansions[62]. In *C9ORF72* FTD/ALS and fragile X-associated tremor ataxia syndrome (FXTAS), for example, both noncoding sense and antisense expanded transcripts aggregate in RNA foci that sequester essential neuronal RBPs leading to the misregulation of mRNA metabolism[64-66]. Interestingly, bidirectional transcription is also detected in diseases caused by coding repeat expansions. In SCA2, the sense expanded transcript is translated into a toxic polyQ tract that abnormally interacts with proteins in stress granules whereas the antisense expanded transcript forms RNA foci, sequesters splicing factors and leads to missplicing of other mRNAs[67].

In many polyQ diseases, such as HD, SCA1, SCA3/MJD, SCA7 and dentatorubral-pallidoluysian atrophy (DRPLA), the transcripts containing coding repeat expansions can aggregate in RNA foci as detected in HD, SCA1, SCA3/MJD, SCA7 and DRPLA fibroblasts,

HD lymphoblasts, induced pluripotent stem cells (iPSCs) and human and murine neuronal progenitors[68-71]. Furthermore, the number of foci often correlates positively with the length of the $(CAG)_{exp}$[72]. In these diseases, the $(CAG)_{exp}$ RNA itself seems to contribute to the pathogenic phenotype, as demonstrated by Li and colleagues in a *Drosophila* model of SCA3/MJD that by modifying the RNA sequence encoding the polyQ from $(CAG)_{exp}$ to a sequence that encodes polyQ but is unable to form a hairpin structure, a $(CAACAG)_{exp}$, the neurodegeneration is reduced[73].

3.4. Repeat associated non-AUG dependent (RAN) translation

The noncoding RNA containing repeat expansions is often able to escape to the cytoplasm and recruit the ribosomal subunits to initiate the translation of a repetitive peptide independently of the presence of an AUG start codon, a mechanism called repeat associated non-AUG dependent (RAN) translation[74].

3.4.1. *Canonical mRNA translation in eukaryotic organisms*

The proteins found in cells are synthesized from mRNAs transcribed using coding DNA as template[75]. The transcribed RNA, or precursor messenger RNA (pre-mRNA), is processed by proteins to originate a mature mRNA[75]. In a process named RNA capping, a N7-methylated guanosine (m7G) is added to the 5' end of the pre-mRNA, this 5'-cap 1) stabilizes and protects mRNA from ribonuclease degradation, 2) facilitates the transport of mRNA from the nucleus to the cytoplasm and 3) allows for mRNA recognition by ribosomes during translation initiation[75, 76]. The capped pre-mRNA is then spliced by the spliceosome, a large ribonucleoprotein complex, wherein introns are removed from the pre-mRNA[77]. The last step of pre-mRNA processing consists in polyadenylation. In this step, a multi-subunit protein complex recognizes the polyadenylation signal (PAS) at the 3' end of the pre-mRNA and recruits the enzyme poly(A) polymerase (PAP)[78]. The multi-subunit protein complex cleaves the pre-mRNA 10–30 nucleotides downstream of the PAS and the PAP adds a poly(A) tail of approximately 200 adenines to the cleaved 3' end of the pre-mRNA[78]. The poly(A) tail stabilizes the mRNA, protects it from phosphatase and ribonuclease degradation and facilitates its nucleocytoplasmic transport[75].

Before translation, the 5'- and 3'-ends of the mRNA are brought closer to each other by four specific interactions: the 5'-cap binds to the eukaryotic translation initiation factor 4E (eIF4E); eIF4E interacts with eIF4G; eIF4G binds to the poly(A) binding protein (PABP) and

PABP recognizes the 3' poly(A) tail. This ''closed-loop'' structure is thought to control mRNA translation[79].

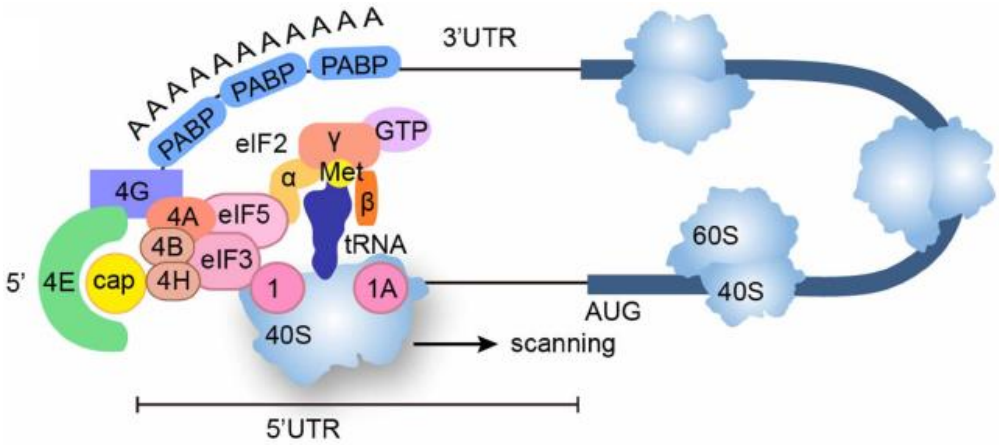


Figure 4. Canonical mRNA translation in eukaryotes. Translation initiation involves the eIF4F complex and the poly(A) tail binding protein PABP binding to the mRNA and subsequently interacting with the 43S complex (eIF5, eIF3, eIF2 and the 40S ribosome) to form the 48S complex. eIF4E and PABP both interact with eIF4G to create a 'closed loop complex'. eIF4A, with its cofactors eIF4B and eIF4H, interact with eIF4G and eIF4E to provide helicase activity to unwind secondary structures present in the 5'UTR. The 48S complex scans the mRNA for an AUG start codon, where the 60S ribosomal subunit is recruited through eIF5B and several of the initiator factors are displaced and recycled to initiate a new round of translation. Adapted from Castelli *et al*, 2021[80].

Once an mRNA has been transcribed, processed, transported into the cytoplasm and circularized, the polypeptide translation is initiated[75] (Figure 4). Translation requires the interaction of 3 types of RNA: mRNA, transfer RNA (tRNA) and ribosomal RNA (rRNA)[75]. Eukaryotes have small 40S and large 60S ribosomal subunits in the cytoplasm, which are usually dissociated[75]. The small 40S subunit comprises the 18S rRNA subunit and 33 proteins, and the large 60S subunit comprises the 5S, 5.8S and 28S rRNA subunits and 47 proteins[81]. Ribosomes contain four binding sites for RNA molecules, one for the mRNA and three (the A-site, P-site and E-site) for tRNAs[75] (Figure 5A). The tRNA is responsible for bringing the amino acids to the ribosome. One region of the tRNA, called anticodon, consists of three consecutive nucleotides that pair with the complementary codon in the mRNA molecule, another region at the 3'-end of the tRNA molecule binds to the amino acid that matches the codon[75] (Figure 5B).

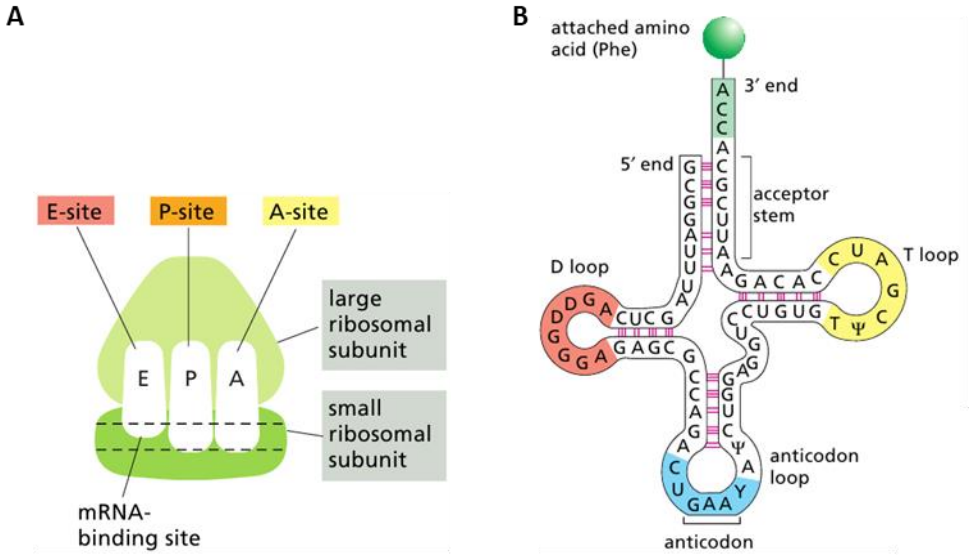


Figure 5. Schematic representation of a ribosome (A) **and a tRNA molecule** (B)**. (A)** Each ribosome has one binding site for mRNA and three binding sites for tRNA: the A-, P- and E-sites (short for aminoacyl-tRNA, peptidyl-tRNA and exit, respectively). **(B)** A tRNA specific for the amino acid phenylalanine (Phe) is depicted; all other tRNAs have similar structures. The complementary base-pairing (red lines) creates the double-helical regions of the molecule. The anticodon is the sequence of three nucleotides that base-pairs with a codon in mRNA. The amino acid matching the codon/anticodon pair is attached at the 3' end of the tRNA. tRNAs contain some unusual bases, which are produced by chemical modification after the tRNA has been synthesized. For example, the bases denoted Ψ (pseudouridine) and D (dihydrouridine) are derived from uracil. Adapted from Alberts *et al*, 2008[75].

Protein translation occurs in three phases: initiation, elongation and termination[75]. Initiation begins with binding of the initiator methionyl-tRNA complex (Met–tRNAi), along with additional eIFs to the P-site of the small ribosomal subunit[75]. Met–tRNAi differs in nucleotide sequence from the elongator methionyl-tRNA (Met-tRNA$_m$) that decodes internal AUG methionine codons; of all the aminoacyl-tRNAs in the cell, only Met-tRNAi is capable of tightly binding the small ribosome subunit without the complete ribosome being present[75]. Then, the small ribosomal subunit recognizes and binds to the 5'-end of an mRNA molecule and scans the mRNA by moving towards the 3'-end in consecutive groups of three nucleotides (codon by codon), until it finds an AUG codon[75]. Recognition of the AUG codon is highly dependent on its surrounding sequence. The Kozak consensus sequence is a nucleic acid motif ((GCC)GCC**R**CCAUG**G**) that functions as the translation initiation site in most eukaryotic mRNAs[82]. The contribution of the Kozak sequence nucleotides for AUG codon recognition is different in each position[82, 83]. Considering the adenine from the AUG codon as position +1, the most important elements for optimal sequence context in AUG codon recognition are the purine (R; adenine or guanine) at position -3 and the guanine (G) at position +4[82, 83]. However, the significance of the GCC motif in positions -9 to -7 remains

to be established[82]. In the presence of the complete Kozak sequence, with both the purine at position -3 and the guanine at position +4 (NNN**R**NNAUG**G**), translation is strongly efficient; in the presence of the purine at position -3 or the guanine at position +4 (NNN**R**NNAUG(**A/C/U**) or NNN**(C/U)**NNAUG**G**), translation is moderately efficient; in the absence of the purine at position -3 and the guanine at position +4 (NNN(**C/U**)NNAUG(**A/C/U**)), translation is poorly efficient[84]. Once the small ribosomal subunit finds the AUG codon, the initiation factors dissociate and the large ribosomal subunit joins the complex, forming an elongation-competent 80S ribosome[75].

Since Met–tRNAi binds to the P-site of the ribosome, when the elongation phase of translation begins, the A-site is vacant and ready to bind the aminoacyl-tRNA molecule that is complementary to the exposed codon[75]. Once that binding happens, the peptidyl transferase enzyme, which is integrated in the large ribosomal subunit, catalyzes the formation of a peptide bond between the methionine and the following encoded amino acid[75]. After the synthesis of this bond, the methionine unbinds from its corresponding tRNA[75]. Then, the mRNA is moved through the ribosome to the next codon leading to the release of the empty tRNA via the E (exit) site[75] (Figure 6). The A-site becomes vacant once again, available to bind to the next aminoacyl-tRNA complementary to the exposed mRNA codon[75].

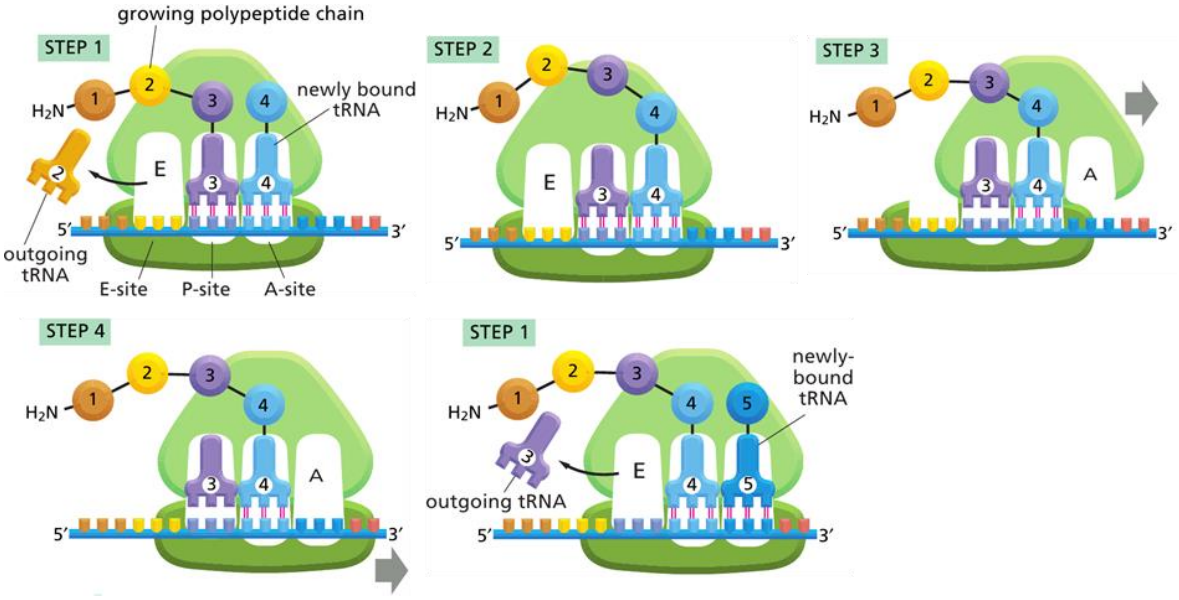


Figure 6. Elongation phase of mRNA translation in eukaryotes. Each amino acid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon on its attached tRNA molecule and the next codon on the mRNA chain. Because only one of the many types of tRNA molecules in a cell can base-pair with each codon, the codon determines the specific amino acid to be added to the growing polypeptide chain. The four-step cycle shown is repeated over and over during the synthesis of a protein. In step 1, an aminoacyl-tRNA molecule

binds to a vacant A-site on the ribosome and a spent tRNA molecule dissociates from the E-site. In step 2, a new peptide bond is formed. In step 3, the large subunit translocates relative to the small subunit, leaving the two tRNAs in hybrid sites: P on the large subunit and A on the small, for one; E on the large subunit and P on the small, for the other. In step 4, the small subunit translocates carrying its mRNA a distance of three nucleotides through the ribosome. This "resets" the ribosome with a fully empty A-site, ready for the next aminoacyl-tRNA molecule to bind. As indicated, the mRNA is translated in the 5'-to-3' direction and the N-terminal end of a protein is made first, with each cycle adding one amino acid to the C-terminus of the polypeptide chain. Adapted from Alberts *et al*, 2008[75].

The translation is terminated when the ribosome encounters one of the three stop codons: UAG, UAA or UGA[75]. Stop codons are not recognized by tRNA because there are no anticodons complementary to stop codons; instead, stop codons recruit proteins known as release factors, which lead to the addition of a water molecule to the growing polypeptide by the peptidyl transferase[75]. This reaction separates the polypeptide from the tRNA releasing the newly synthesized protein into the cytoplasm[75]. Then, the ribosome releases the mRNA and separates into the large and small subunits, which can be recycled and begin a new round of protein synthesis on the same, or another, mRNA molecule[75].

The mRNA molecules that are being translated are usually found in polyribosomes, composed by a single mRNA attached to several ribosomes that are translating the mRNA in the elongation phase[75]. This increases translational efficiency.

3.4.2. *RAN translation*

RAN translation was reported for the first time in 2011, in SCA8 and DM1[85]. In this work, Zu and colleagues demonstrated *in vitro,* that in the absence of an initiation (AUG) codon, the $(CAG)_{exp}$ RNA is translated in the 3 reading frames producing homopolymeric stretches of polyQ, polyserine (polyS) and polyA[85]. The authors also found RAN translated polyA aggregates in cerebellar Purkinje cells derived from SCA8 affected individuals and RAN translated polyQ aggregates in myoblasts and skeletal muscle derived from DM1 affected individuals[85]. After this discovery, abnormal translation of repeat expansions has been reported in 10 repeat expansion diseases, not only in diseases caused by trinucleotide repeat expansions but also by tetra-, penta- and hexanucleotide repeat expansions (Table 1).

Table 1. RAN peptides identified in repeat expansion diseases

Disease	Repeat motif	RAN polypeptides	Biological material							Ref
			Affected individuals		Animal models		Transfected/transduced cell lines			
			Cells/tissues	*Antibodies*	*Cells/tissues*	*Antibodies*	*Cells*	*Vector*	*Antibodies*	
DM1	CTG/CAG	PolyQ PolyA PolyS	Myoblasts, skeletal muscle, leukocytes	α-polyQ	Mouse cardio-myocytes, leukocytes	α-polyQ	HEK293[b)] N2a[b)]	pcDNA3.1 (6xSTOP-$(CAG)_n$-Myc+HA+Flag)	α-polyQ α-Myc α-HA α-Flag	20, 21, 85
DM2	CCTG/ CAGG	Poly(LPAC) Poly(QAGR)	Cortex, striatum, hippocampus	α-poly(LPAC) α-poly(QAGR)	-	-	HEK293T T98	pcDNA3.1 and pcDNA5/ FRT/TO (6xSTOP- $(CCTG)_n$-Flag+HA+Myc); pcDNA3.1 and pcDNA5/ FRT/TO (6xSTOP- $(CAGG)_n$-Flag+HA+Myc)	α-Myc α-Flag α-HA α-poly(LPAC) α-poly(QAGR)	86-88
FECD	CTG/CAG	PolyC PolyQ	Fibroblasts, corneal endothelium	α-polyQ α-polyC	-	-	HEK293 HCEnC21-T	pcDNA3.1 ($(CTG)_n$-3xFlag); pcDNA3.1 ($(CAG)_n$-3xFlag)	α-polyC α-polyQ α-Flag	89-91

FTD/ALS	GGGGCC/
	GGCCCC
		Poly(GA)
		Poly(GP)
		Poly(GR)
		Poly(PA)
		Poly(PR)
		Poly(GA:GP)
			Cerebellum, hippocampus, testes, neocortex, medial and lateral geniculate nuclei, iPSN	α-poly(GA)
				α-poly(GP)
				α-poly(GR)
					Mouse cerebellum, cortex, thalamus, striatum; Zebrafish brain, spinal cord ; *Drosophila* head	α-poly(GA)
						α-poly(GP)
						α-poly(GR)
						α-poly(PA)
						α-poly(PR)
						α-GFP
							HEK293T
							HeLa
								pAG3 $(CCCCGG)_n$;
								pcDNA3.1(6xSTOP-$(GGGGCC)_n$-Flag+HA+ Myc);
								pcDNA3.1(6xSTOP-$(GGCCCC)_n$-Flag+HA+ Myc);
								pEGFP-C1 ($(GGCCCC)_n$)
									α-poly(GA)
									α-poly(GP)
									α-poly(GR)
									α-HA
									α-Myc
									α-Flag
										64, 92-102

FXTAS	CGG/CCG	PolyG PolyA PolyP PolyR	Frontal cortex, cerebellum, hippocampus	α-polyG	*Drosophila*; Mouse hippocampus	α-polyG α-GFP	COS-7 SH-SY5Y	($(CGG)_n$-GFP); (FLAG-$(CGG)_n$)	α-polyG α-Flag α-GFP	66, 103

Disease	Repeat motif	RAN polypeptides	Biological material							Ref
			Affected individuals		Animal models		Transfected/transduced cell lines			
			Cells/tissues	*Antibodies*	*Cells/tissues*	*Antibodies*	*Cells*	*Vector*	*Antibodies*	
FXPOI	CGG	PolyG	Ovaries, pituitary gland	α-polyG α-ubiquitin	Mouse ovaries, pituitary gland	α-polyG α-ubiquitin	-	-	-	[104]
HD	CAG/CTG	PolyQ[a)] PolyA PolyS PolyL PolyC	Cortex, striatum, cerebellum	α-polyQ α-polyA α-polyS α-polyL α-polyC	Mouse cortex, cerebellum and striatum	α-polyA α-polyS	HEK293T T98 SH-SY5Y	pcDNA3.16xSTOP-$(CAG)_n$-Myc+Flag+HA)	α-polyQ α-Flag α-Myc α-HA	[68, 69, 85, 105, 106]
HDL2	CAG/CTG	PolyQ[a)] PolyA PolyS	-	-	Mouse cerebellum	α-polyQ	HEK293T Rabbit reticulocyte lysates	pcDNA3.1 (6xSTOP-$(CAG)_n$-Flag+HA+Myc/His-poly(A))	α-polyQ α-Flag α-HA α-His	[85, 107]
SCA2	CAG	PolyQ[a)] PolyA PolyS	-	-	-	-	HEK293T	pcDNA3.1 6xSTOP-CTG-$(CAG)_n$-HA+Myc+ Flag)	α-HA	[108, 109]
SCA3/ MJD	CAG	PolyQ[a)] PolyA PolyS	-	-	-	-	HEK293T HeLa SH-SY5Y	pcDNA3.1 (6xSTOP-$(CAG)_n$-Myc+HA+His-6xSTOP)	α-Myc α-HA α-His	[85, 110-113]
SCA8	CTG/CAG	PolyQ[a)] PolyA PolyS	Cerebellum, brainstem, hippocampus, cortex	α-polyA α-polyS	Mouse cerebellum, brainstem, hippocampus, cortex	α-polyQ α-polyA α-polyS	HEK293[b)] N2a [b)]	pcDNA3.1 (6xSTOP-$(CAG)_n$-Flag+HA+Myc/His-); pcDNA3.1 (6xSTOP-V5-$(CAG)_n$-Flag+HA+Myc/ His)	α-polyQ α-Myc α-HA α-Flag α-V5	[85, 114, 115]
SCA31	TGGAA	Poly(WNGME)	Purkinje cells	α-poly(WNGME)	*Drosophila* eye imaginal discs	α-poly(WNGME)	-	-	-	[116]

Disease	Repeat motif	RAN polypeptides	Biological material							Ref
			Affected individuals		Animal models		Transfected/transduced cell lines			
			Cells/tissues	*Antibodies*	*Cells/tissues*	*Antibodies*	*Cells*	*Vector*	*Antibodies*	
SCA36	GGCCTG/ CAGGCC	Poly(GP) Poly(WA) Poly(GL) Poly(PR) Poly(AQ)	iPSCs, fibroblasts, lymphoblastoid cell lines, brain organoid, cortex, cerebellum, spinal cord	α-poly(GP) α-poly(PR)	-	-	HEK293T	pAG3 (6xSTOP-$(GGCCTG)_n$-Flag+HA +Myc)	α-HA α-Myc α-Flag	[97, 98]

Abbreviations: DM1 - myotonic dystrophy type 1; DM2 - myotonic dystrophy type 2; FECD - Fuchs endothelial corneal dystrophy; FTD/ALS - frontotemporal dementia/amyotrophic lateral sclerosis; FXTAS - fragile X tremor/ataxia syndrome; FXPOI - fragile X-associated primary ovarian insufficiency; HD – Huntington's disease; HDL2 - Huntington disease-like 2; SCA2 - spinocerebellar ataxia type 2; SCA3/MJD - spinocerebellar ataxia type 3/Machado-Joseph disease; SCA8 - spinocerebellar ataxia type 8; SCA31 - spinocerebellar ataxia type 31; SCA36 - spinocerebellar ataxia type 36. [a)] This peptide can result from both canonical or RAN translation. [b)] Cells were transfected and transduced.

RAN translation consists in the translation initiation upstream or within the repeat tracts, at non-AUG codons, resulting in the production of potentially toxic repetitive peptides contributing to neurodegeneration[74, 117]. RAN translation occurs because expanded repeats induce the formation of secondary structures, such as hairpins, able to slow the movement of the 40S ribosomal subunit and to recruit eIFs promoting translation initiation at non-AUG codons[85, 118, 119] (Figure 8). Usually, RAN translation occurs in the 3 reading frames of the mRNA containing the repeat expansion. Interestingly, in many diseases as *C9ORF72* FTD/ALS, DM1, SCA8 and HD, RAN translation occurs in the three frames of both sense and antisense repeat transcripts[120] (Figure 7 and Table 1) .

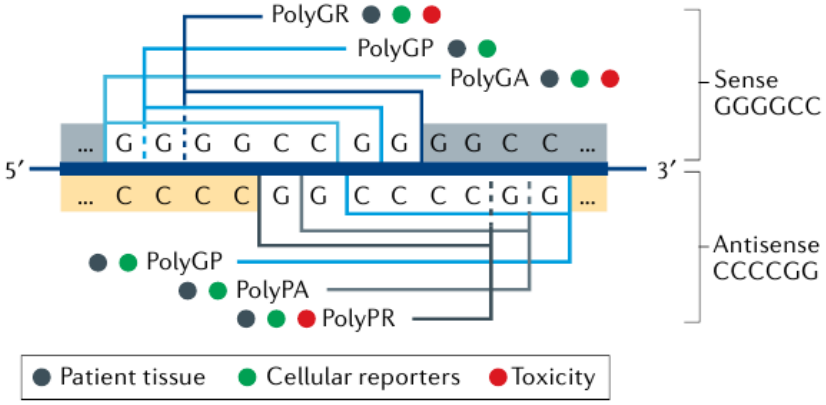
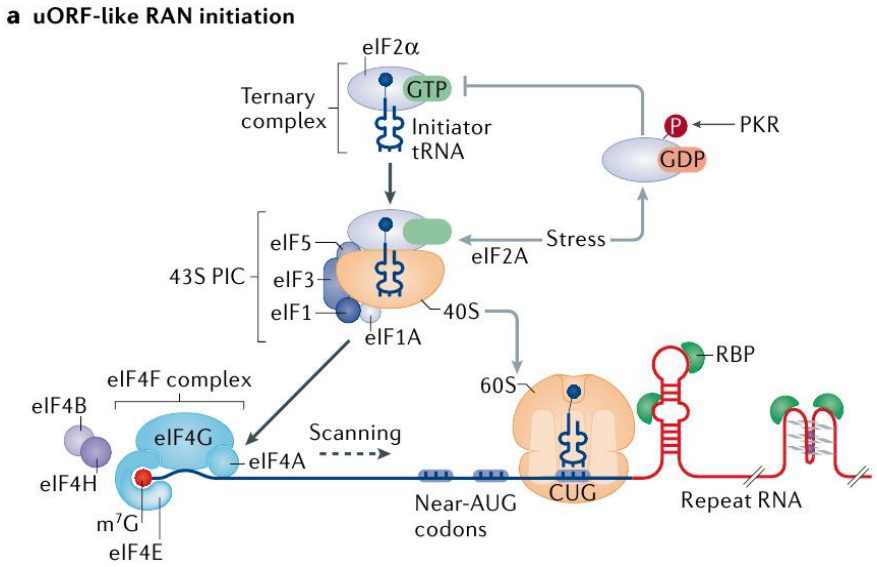


Figure 7. RAN translation from the *C9ORF72* GGGGCC sense and CCCCGG antisense transcripts generates multiple dipeptide repeats (DPRs). Although all DPRs are detected in tissues of individuals with *C9ORF72* FTD/ALS and generated by cellular reporters, arginine-containing DPRs have the highest intrinsic toxicity in model systems. Adapted from Malik *et al*, 2021[121].

Depending on the repetitive motif and gene context, RAN translation may be initiated by different mechanisms. In FXTAS, the $(CGG)_{exp}$ is bidirectionally transcribed and RAN translation occurs efficiently in 2 reading frames of the sense $(CGG)_{exp}$ generating polyglycine (polyG) and polyA, and also occurs in 3 reading frames of the antisense $(CCG)_{exp}$ generating polyproline (polyP), polyarginine (polyR) and polyA[66, 103]. Interestingly, mutagenesis *in vitro* assays showed that the RAN translation of polyG from the $(CGG)_{exp}$ RNA initiates at an upstream open reading frame (uORF) in ACG or GUG codons upstream of the repeat[122]. This mechanism of RAN translation initiation partially overlaps with canonical translation initiation since both mechanisms require a 5'-capped mRNA and eIFs[122] (Figures 4 and 8A).

In other diseases, as *C9ORF72* FTD/ALS, RAN translation of the expanded mRNA functionally overlaps with internal ribosomal entry site (IRES) translation initiation[123] (Figure 8B). IRES-mediated translation is an alternative translation mechanism, predominantly

described in viral mRNAs but also found in some cellular mRNAs, as p53 and c-myc[124, 125]. IRES are RNAs with highly structured sequences that promote translation initiation in the absence of an AUG codon and a 5'-cap, by directly recruiting initiation factors and ribosomal subunits[126]. In *C9ORF72* FTD/ALS, the $(GGGGCC)_{exp}$ is bidirectionally transcribed and RAN translation initiates in each frame of both sense and antisense repeat expansion transcripts originating 5 RAN polypeptides: poly(glycine-alanine) and poly(glycine-arginine) (poly(GA) and poly(GR)) from the sense $(GGGGCC)_{exp}$, poly(proline-alanine) and poly(proline-arginine) (poly(PA) and poly(PR)) from the antisense $(GGCCCC)_{exp}$, and poly(glycine-proline) (poly(GP)) from both sense and antisense hexanucleotide repeat expansions[64, 99] (Figure 7). The IRES-like structure leading to RAN translation initiation in *C9ORF72* FTD/ALS is supported by the ability of $(GGGGCC)_{exp}$ RNA to form complex structures similar to IRES and by the fact that depletion of ribosomal protein S25 (RPS25), an essential element for IRES-mediated translation initiation, reduces levels of RAN translated poly(GP), poly(GA) and poly(GR) in Hap1 cells transfected with the $(GGGGCC)_{exp}$, without affecting transcription levels[123, 127-129].



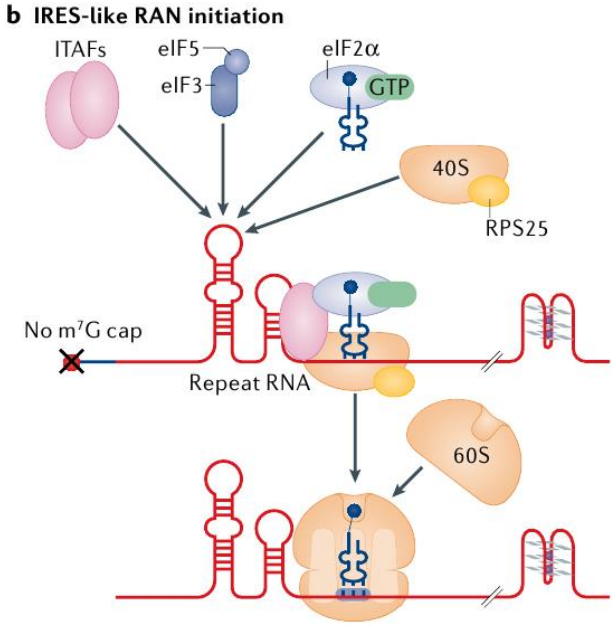


Figure 8. Mechanisms of RAN translation. (A) Canonical AUG-mediated translation initiation and some forms of RAN translation begin with binding of the eukaryotic initiation factor 4F (eIF4F) complex — consisting of eIF4E, eIF4G and eIF4A — to the mRNA m7G cap with eIF4B and/or eIF4H. This step is followed by recruitment of the 43S preinitiation complex (PIC), which comprises the 40S small ribosomal subunit and the ternary complex eIF2α–GTP–initiator tRNA (Met- tRNAi). The PIC scans the 5′ untranslated region for an AUG start codon. RAN translation initiation at near-AUG codons utilizes canonical translation machinery and resembles non-AUG upstream open reading frame (uORF) translation initiation. However, in RAN translation the repeat impedes PIC scanning, which lowers codon fidelity and boosts initiation at specific near-AUG codons. In stress conditions, eIF2α phosphorylation by protein kinase R (PKR) blocks the recycling of the ternary complex and inhibits canonical translation, but allows eIF2A-mediated RAN translation. RNA-binding proteins (RBPs) regulate RAN translation initiation by binding and altering repeat RNA structures. (**B)** RAN translation may also start through internal ribosome entry site (IRES)-like mechanisms in a cap-independent manner, which is supported by 40S ribosomal subunit S25 (RPS25) and other IRES trans-acting factors (ITAFs), such as eIF3 and eIF5. Adapted from Malik *et al*, 2021[121].

The number of polypeptide species produced by RAN translation is different for each disease. If the number of nucleotides in the pathogenic repeat motif is a multiple of 3, as in FXTAS and *C9ORF72* FTD/ALS, up to 3 polypeptides are translated from each mRNA (Table 1). If the number of nucleotides in the pathogenic repeat motif is not a multiple of 3, as in myotonic dystrophy type 2 (DM2) and SCA31, only one repeated polypeptide, with a number of amino acids equal to the number of nucleotides in the repeat motif, is translated from the 3 frames of the expanded transcript (Table 1). In DM2, the $(CCTG)_{exp}$ is bidirectionally transcribed and upon RAN translation, this expansion originates poly(leucine–proline–alanine–cysteine) (poly(LPAC)) from the 3 reading frames of the sense $(CCUG)_{exp}$ and poly(glutamine–alanine–glycine–arginine) (poly(QAGR)) from the 3 reading frames of the antisense $(CAGG)_{exp}$[87]. In SCA31, the $(TGGAA)_n$ insertion is translated in a pentapeptide poly(tryptophan-asparagine-glycine-methionine-glutamic acid)

(poly(WNGME)), encoded by the three reading frames of the repeat insertion[116]. However, since the $(UGGAA)_n$ transcript itself includes AUG codons upstream close to the repeat, it is unclear whether the polypeptide is generated through RAN translation or canonical translation[116]. In SCA31, the translation of the antisense $(TTCCA)_n$ was not investigated yet, but it can potentially be RAN translated in poly(phenylalanine-histidine-serine-isoleucine-proline) (poly(FHSIP)).

Interestingly, repeat expansions can originate polypeptides that are canonically translated, as is the case of the polyQ resulting from $(CAG)_{exp}$ in coding regions, or by RAN translation (Table 1). Independently of how they are translated, expanded polypeptides are potentially neurotoxic by forming aggregates that sequester proteins essential for cell survival[117].

The toxicity of RAN polypeptides has been extensively described in *C9ORF72* FTD/ALS. In this disease, poly(GA), poly(GR), poly(PA), poly(PR) and poly(GP) accumulate in the hippocampus, frontotemporal neocortex and cerebellum of *C9ORF72* FTD/ALS affected individuals, although poly(GA), poly(GR) and poly(GP) aggregates are more abundant than poly(PA) and poly(PR) aggregates[64, 65, 99, 130]. In numerous studies in yeast, transfected cell lines, *Drosophila*, zebrafish and mouse models, the individual expression of poly(GR), poly(PR) or poly(GA) was often sufficient to induce neurotoxicity[120, 121, 131]. Poly(GP) and poly(PA), however, seem to be nontoxic when individually expressed[95, 132, 133]. The highest levels of toxicity were consistently observed for poly(GR) and poly(PR)[95, 132, 133]. The arginine-rich domain in these dipeptides facilitates the binding to low complex domain (LCD)-containing proteins, which are present in the nuclear pore channel and in membraneless organelles as nucleoli, Cajal bodies and stress granules[132, 134-137]. The interaction of poly(GR) and/or poly(PR) with LCD-containing proteins, in transfected cell lines and *Drosophila,* inhibit the formation and the function of membraneless organelles and nuclear pore channels, which leads to mRNA missplicing, suppression of ribosomal RNA synthesis, impaired nucleocytoplasmic transport and cell death[132, 134-137]. In *C9ORF72* FTD/ALS induced pluripotent stem cell-derived neurons (iPSN), poly(GR) binds to mitochondrial ribosomal proteins and inhibits their functions, increasing oxidative stress and inducing DNA damage[138]. Although poly(GA) forms the most visible aggregates in the hippocampus, frontotemporal neocortex and cerebellum of *C9ORF72* FTD/ALS affected individuals, it is only moderately toxic compared to poly(GR) and poly(PR)[64, 65, 95, 99, 130, 133]. When overexpressed in primary neurons, poly(GA) sequesters proteins involved in axonal branching, proteasome proteins and activates caspase-3 leading to reduced dendritic branching, proteasomal inhibition, increased

endoplasmic reticulum stress and apoptosis[139, 140]. Interestingly, in *C9ORF72* FTD/ALS a chimeric poly(GA:GP) was also found showing that this peptide is likely a product of RAN translation that resulted from ribosomal frameshifting[98, 141] (Figure 9).

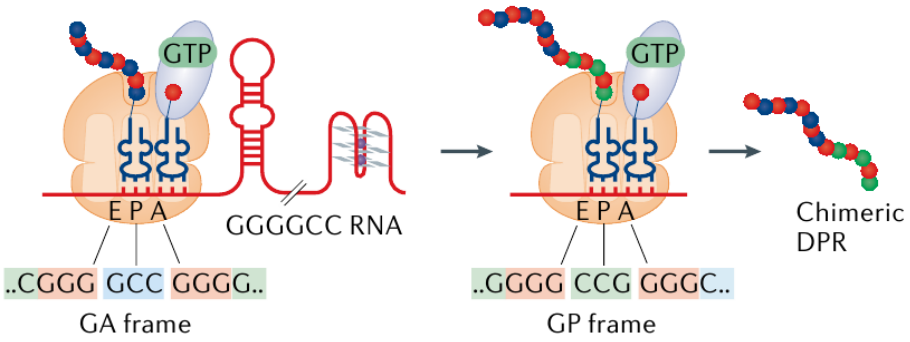


Figure 9. Ribosomal frameshifting during RAN translation. Stable RNA secondary structures formed by GGGGCC repeats induce ribosomal frameshifting during RAN translation, leading to the production of chimeric DPRs. Adapted from Malik *et al*, 2021[121].

The toxicity of RAN polypeptides was also demonstrated in other repeat expansion diseases, as FXTAS and DM2[4]. In FXTAS, polyG aggregates in the frontal cortex, cerebellum and hippocampus of affected individuals[103]. In FXTAS iPSN, overexpressed polyG sequesters proteins involved in the architecture of neuronal nuclear lamina increasing neuronal death[142]. When overexpressed in *Drosophila* models, FXTAS polyG impairs the ubiquitin proteasome system (UPS) inducing rough-eye phenotypes[103, 143]. In DM2, poly(LPAC) and poly(QAGR) aggregate in the cortex, hippocampus and striatum of affected individuals and when overexpressed in T98 cells, these RAN polypeptides induce cell death[87]. To confirm that T98 cell death is mediated by poly(LPAC) and poly(QAGR) peptides and not the $(CCUG)_{exp}$ or $(CAGG)_{exp}$ RNAs, Zu and colleagues transfected the T98 cells with vectors expressing alternative codons encoding poly(LPAC) and poly(QAGR) peptides and showed that cell death is still induced[87].

4. Spinocerebellar ataxias

Spinocerebellar ataxias (SCAs) are a group of clinically and genetically heterogeneous neurodegenerative diseases[144]. SCAs are characterized by cerebellar degeneration leading to gait imbalance, limb incoordination and dysarthria[144]. Other symptoms may be associated with specific SCA types, such as tremor in SCA12, SCA15 and SCA17, seizures in SCA10 and SCA19/22, or dementia in SCA2, SCA3, SCA10,

SCA14, SCA17, SCA19/22 and SCA21[144]. SCAs are usually autosomal dominant adult-onset diseases[144]. Their clinical manifestations are progressive and lead to permanent disability or premature death[144, 145]. There are 48 SCAs clinically described, but only 34 causative genes have been identified[146, 147]. Twenty-one SCAs are caused by classical mutations as point mutations, deletions or duplications whereas 13 SCAs are caused by unstable microsatellite repeats[147, 148]. SCA1, SCA2, SCA3, SCA6, SCA7, SCA17 and DRPLA are caused by polyQ-encoding $(CAG)_{exp}$, whereas SCA12 is caused by a noncoding $(CAG)_{exp}$ located in gene promoter, or 5'-UTR, depending on the transcripts (Figure 2). Regarding SCAs caused by noncoding transcribed repeats, SCA8 is caused by a $(CTG)_{exp}$ in 3'-UTR of *ATXN8OS* gene, and intronic $(ATTCT)_{exp}$ and $(GGCCTG)_{exp}$ cause SCA10 and SCA36 respectively (Figure 2). Two SCAs, SCA31 and SCA37, are caused by pentanucleotide repeat insertions in noncoding gene regions (Figure 2).

SCAs have an average prevalence of 2.7 in 100,000 individuals worldwide, being considered rare diseases[149]. SCA3/MJD is the most common SCA worldwide, followed by SCA2 and SCA6[149]. The regional prevalence of SCAs is highly variable, which reflects populational genetic diversity and founder effects[149]. In Portugal, SCA3/MJD is the most prevalent SCA[150, 151]. In Cuba, England and Venezuela, for example, the most prevalent SCA is SCA2, SCA6 and SCA7, respectively[152-154].

Although many advances have recently been made in human genetics, approximately a half of clinically diagnosed SCA families remain without gene assignment[151]. Consequently, these affected individuals do not have the molecular diagnosis required for the predictive diagnosis of the asymptomatic at-risk individuals and genetic counselling of the families. Understanding the molecular mechanisms underlying these diseases is also imperative for the identification of molecular targets required for the development of suitable treatments for these diseases, presently with no cure.

4.1. Spinocerebellar ataxia type 37

In 2017, the laboratory where I performed this work identified the genetic cause of SCA37. SCA37 is a pure cerebellar ataxia characterized by dysarthria as the first symptom and onset of the disease between late teens to sixties[13]. SCA37 is caused by an $(ATTTC)_n$ insertion in the middle of a normal polymorphic ATTTT repeat in the DAB adaptor protein 1 (*DAB1*) gene[13]. Unaffected individuals have two $(ATTTT)_n$ alleles ranging from 7-400 repeat units, while affected individuals carry an allele with the configuration $[(ATTTT)_{60-}$

$_{79}$(ATTTC)$_{31–75}$(ATTTT)$_{58–90}$] (Figure 10). SCA37 is a disease with hundreds of affected and at-risk individuals in Portugal and Spain[13, 155, 156].

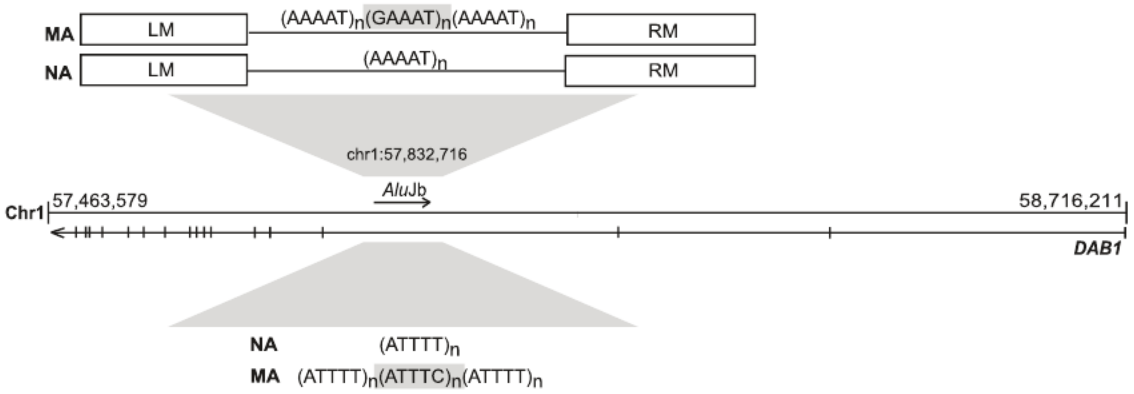


Figure 10. Schematic representation of the ATTTT/AAAAT simple repeat flanked by left (LM) and right (RM) monomers in the polymorphic middle A-rich region of an AluJb sequence in an intron of the *DAB1* 5'UTR; also depicted is the structure of normal alleles (NAs) with pure ATTTT/AAAAT repeats and mutant alleles (MAs) with the (ATTTC)$_n$ insertion. Adapted from Seixas *et al*, 2017[13].

DAB1 encodes a reelin signal transducer. DAB1 protein contains three main domains, an N-terminal protein interaction domain that binds to Reelin receptors, an internal tyrosine-rich region and a C-terminal serine/threonine-rich region[157, 158]. During brain development the binding of Reelin to its receptors induces DAB1 tyrosine phosphorylation, the tyrosine-phosphorylated DAB1 activates multiple signalling cascades resulting in the rearrangement of the cytoskeleton that guides the migration of cortical neurons to their proper location[157, 158]. In adult mice, the reelin-Dab1 pathway regulates the size and shape of dendritic spines, synaptic configuration and affects memory and learning processes[159, 160].

The (ATTTC)$_n$ insertion in *DAB1* is unstable upon parent to offspring transmission and prone to expansion especially when transmitted by the father[13]. The length of the (ATTTC)$_n$ insertion increased by 2-7 repeats in 67% of maternal transmissions and remained stable in the other 33%, whereas all paternal transmissions led to an expansion between 2-12 ATTTC repeats[13]. There is an inverse correlation between the (ATTTC)$_n$ insertion size and the age of disease onset (R=-0.68), meaning that affected individuals with larger repeat insertion alleles present earlier age of disease onset[13]. In SCA37 the size of the repeat insertion explains approximately 50% of the variation in age of onset (R^2=0.46)[13].

The *DAB1* gene encodes 4 mRNA variants (V1, V2, V3 and V4) that differ by alternative 5' and 3'-UTRs[13] (Figure 11A). V1, V2 and V4 encode the same 555 amino acid DAB1 protein whereas V3 encodes a 213 amino acid predicted protein[13]. *DAB1* V1, V3 and V4 are expressed preferentially in the cerebellum of human adult brains whereas V2 is expressed preferentially in the hippocampus[13] (Figure 11B). The level of expression of all

variants is much higher in human fetal central nervous system (CNS) than in adult brains (Figure 11), which is in line with the essential function of *DAB1* in guiding neurons migration during brain development[13].

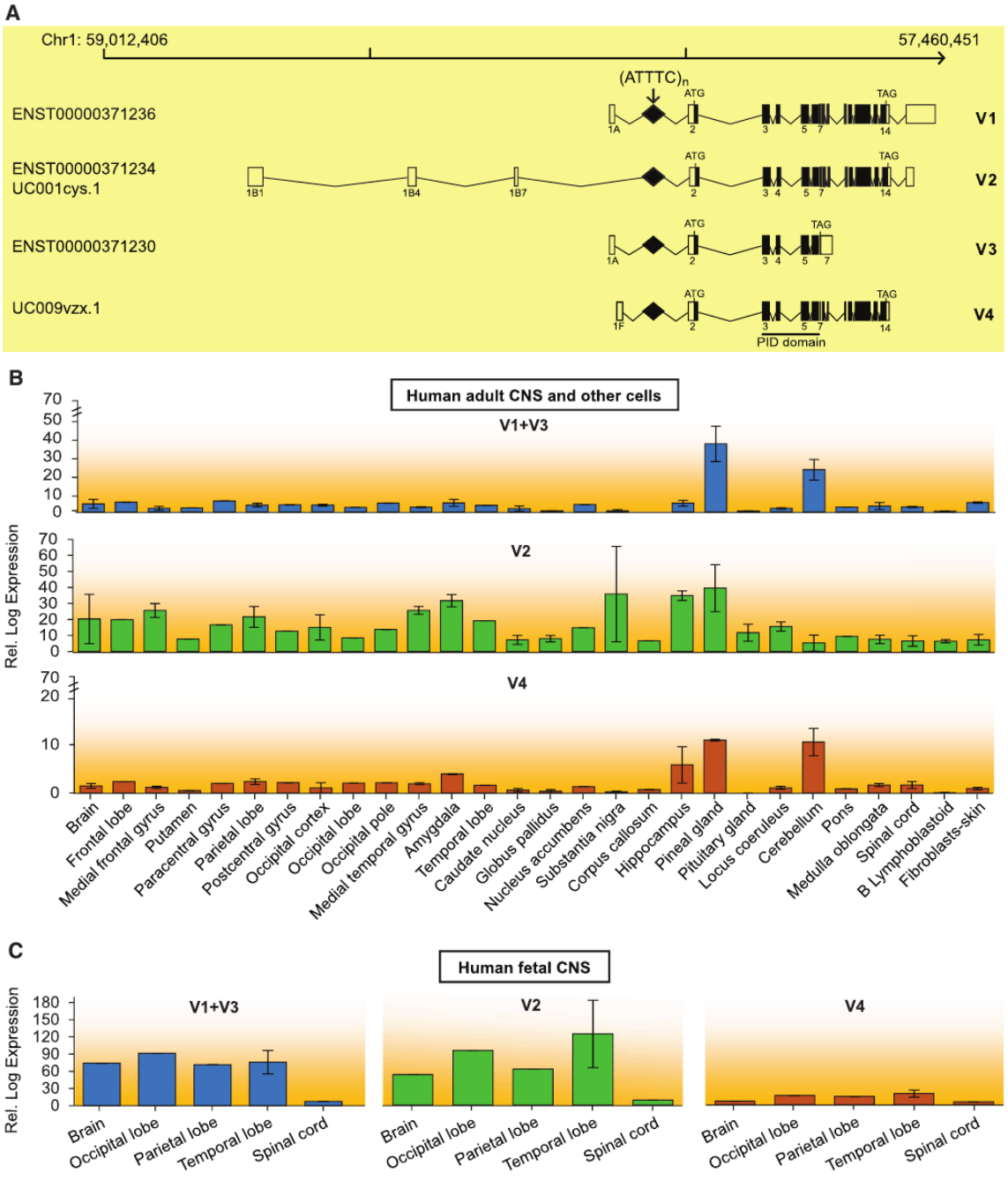


Figure 11. Expression of *DAB1* transcripts spanning the region containing the repeat insertion and *DAB1*. (A) Schematics of *DAB1* genomic position on chromosome 1 show transcripts identified by CAGE (FAMTOM5) to result from usage of alternative promoters. These transcripts are also annotated in Ensembl or the UCSC Genome Browser (hg19) and all have the repeat insertion region in the 5' UTR. Coding and noncoding exons in transcripts are represented by closed and open boxes, respectively. The location of the $(ATTTC)_n$ insertion region in transcript variants is represented by a diamond and indicated with an arrow and the locations of the ATG start codon and TAG stop

codon are shown. The following abbreviation is used: PID, phosphotyrosine interaction domain. **(B)** Mean expression levels of *DAB1* transcripts in different CNS regions, skin fibroblasts and B lymphoblastoid cells from human adults, as analyzed from CAGE data. **(C)** Mean expression levels of *DAB1* transcript variants in CNS regions of 20 to 29-week-old human fetuses. Samples available for expression analysis of each CNS region or cell type ranged from one to three in human adults and from one to two in human fetuses. Data represent the mean ± SD. Adapted from Seixas *et al*, 2017[13].

The aggregation of RNAs containing abnormal repeat expansions or insertions in cell nucleus is a hallmark of diseases caused by noncoding repeats[51]. When overexpressed in HEK293T cells, the *DAB1* $(ATTTC)_n$ insertion forms nuclear RNA aggregates contrarily to short nonpathogenic $(ATTTT)_7$ or large nonpathogenic $(ATTTT)_{139}$ alleles[13] (Figure 12).

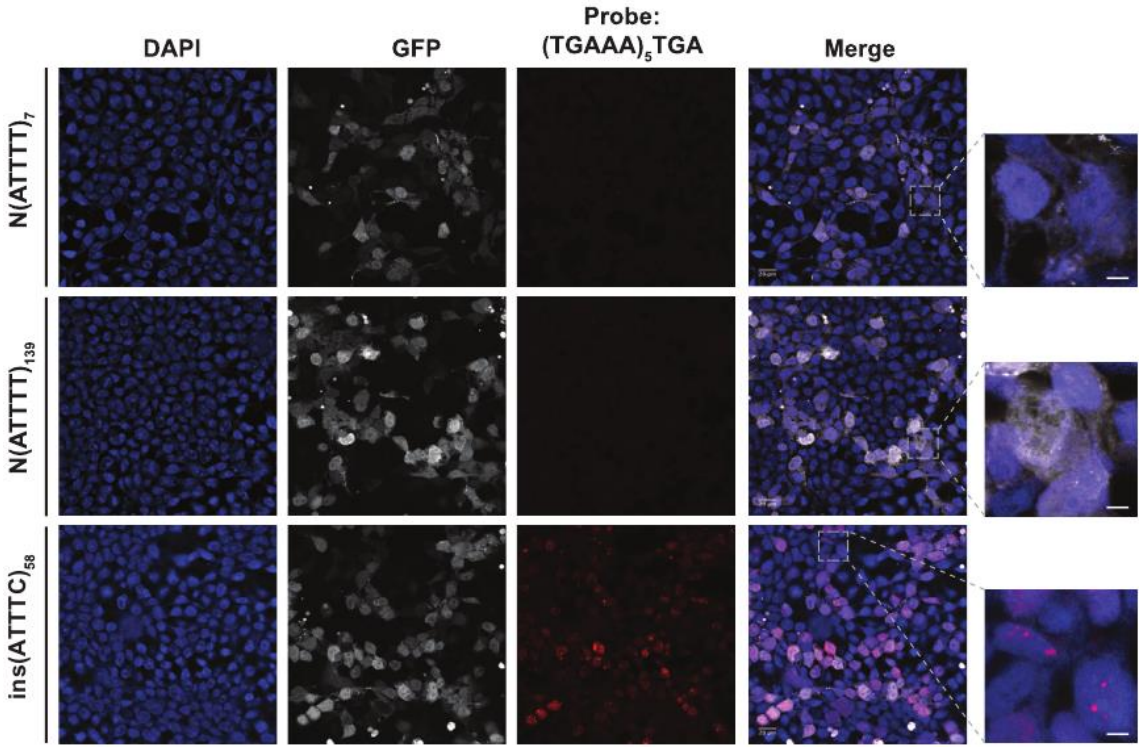


Figure 12. Formation of $(AUUUC)_n$ RNA aggregates in a human cell line. Transient overexpression of the pathogenic ATTTC repeat insertion, but not the normal ATTTT repeat of 7 or 139 units, in HEK293T cells leads to widespread formation of nuclear RNA aggregates visible after FISH staining with a probe, $(TGAAA)_5TGA$, predicted to hybridize to $(AUUUC)_n$. GFP expression was used as a marker for transfection. Represented are single-plane confocal images. Scale bars, 5 mm. Adapted from Seixas *et al*, 2017[13].

In addition to the ability to form RNA foci, the $(ATTTC)_n$ insertion is deleterious *in vivo*[13]. Zebrafish embryos injected with RNA containing the *DAB1* $(AUUUC)_n$ insertion show higher lethality rates and developmental defects when compared to zebrafish embryos

injected with a short nonpathogenic $(AUUUU)_7$ or a large nonpathogenic $(AUUUU)_{139}$ (Figure 13).

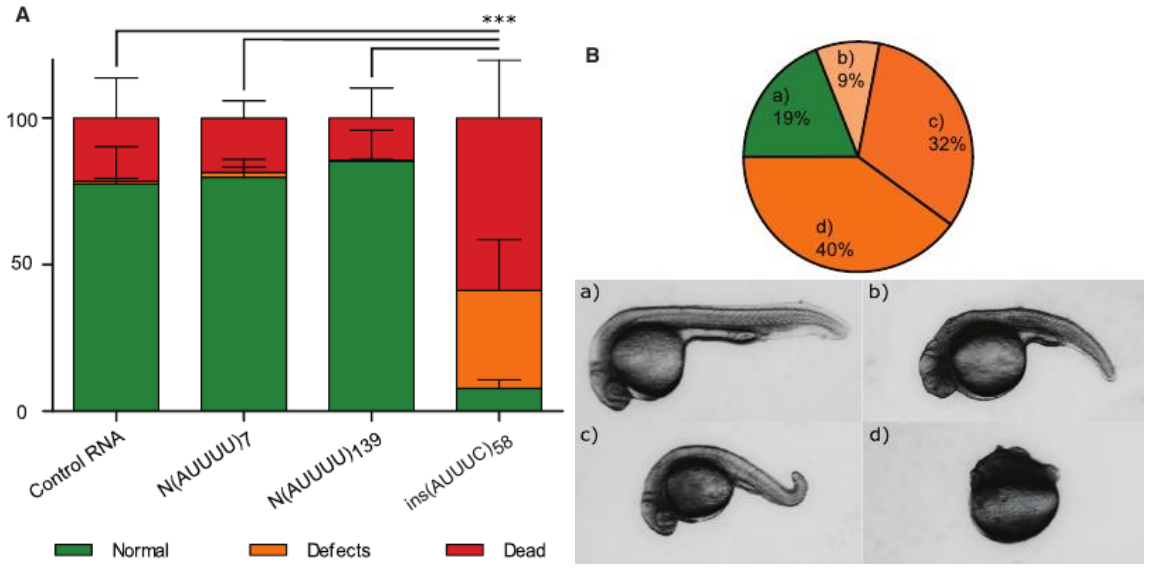


Figure 13. *In Vivo* deleterious effects of the $(ATTTC)_n$ insertion. (A) Percentage of embryos that presented with lethality (dead), developmental defects (defects), or a wild-type phenotype (normal) at 24 hpf after RNA injection of control Cas9RNA, $N(AUUUU)_7$, $N(AUUUU)_{139}$, or $ins(AUUUC)_{58}$ (average of three replicas with at least 200 embryos per replica; ***p % 0.001; c2 test for the lethality rate). Data represent the mean 5 SD. **(B)** Distribution of phenotypic classes (a–d) observed in $ins(AUUUC)_{58}$-injected embryos at 24 hpf and representative images of the observed phenotypic classes: (a) wild-type, (b) severe defects in the tail and head, (c) mild defect in the tail and (d) severe defects in the anterior-posterior axis. Adapted from Seixas *et al*, 2017[13].

SCA37 was the second described inherited disease associated with a noncoding microsatellite repeat insertion. Previously, in 2009, a $(TGGAA)_n$ insertion was found causing SCA31[12]. Similarly to SCA37, the SCA31 repeat insertion is also able to initiate a cascade of RNA-mediated toxic mechanisms, including RNA foci formation[12]. Interestingly, in SCA31, the $(TGGAA)_n$ insertion is abnormally translated in a pentapeptide poly(WNGME), which aggregates in cell bodies and dendrites of Purkinje cells from affected individuals and associates with the disease[116] (Figure 14). However, in SCA37 the translation of pentapeptides from the $(ATTTC)_n$ insertion was not assessed yet.

Recently, after the discovery of $(ATTTC)_n$ insertion as the molecular cause of SCA37, similar noncoding $(ATTTC)_n$ insertions were identified in six genes causing six types of familial adult myoclonic epilepsy (FAME) in families worldwide, however, the pathogenic mechanisms triggered by the $(ATTTC)_n$ insertion were also not assessed in these diseases[14-17].

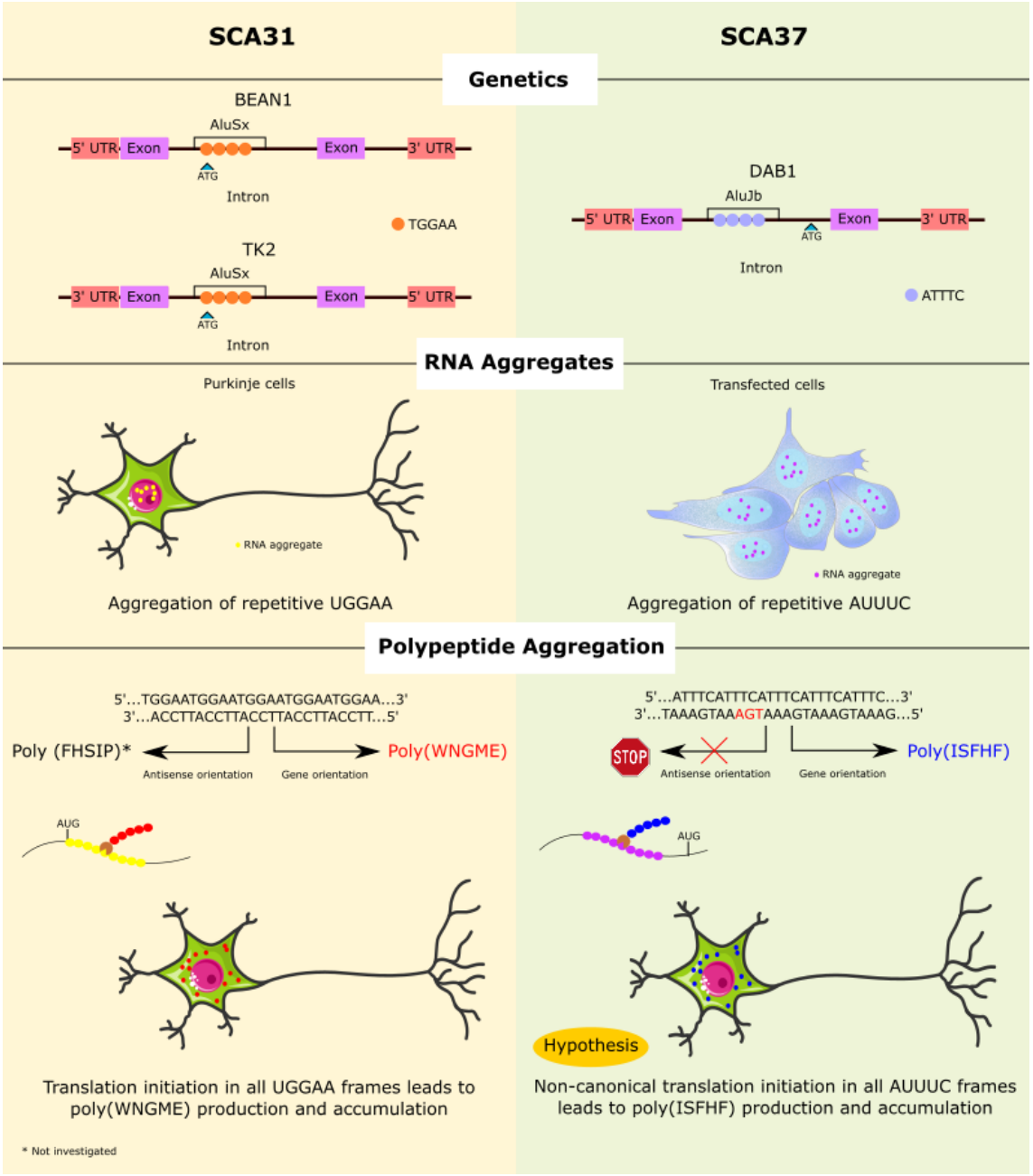


Figure 14. Diseases caused by pentanucleotide repeat insertions: SCA31 and SCA37. Both SCA31 and SCA37 are caused by pentanucleotide repeat insertions in intronic regions, SCA31 is caused by a $(TGGAA)_n$ insertion in an intron shared by the genes *BEAN1* and *TK2*, whereas SCA37 is caused by an $(ATTTC)_n$ insertion in an intron located in the *DAB1* gene. The transcript originated in SCA31, $(UGGAA)_n$, was found to form RNA foci in the nuclei of Purkinje cells from SCA31 affected individuals; the transcript originated in SCA37, $(AUUUC)_n$, was found to form RNA foci in the nuclei of transfected HEK293T cells. Transcription of the UGGAA repeat in the sense orientation produces an unexpected pentapeptide (poly(WNGME)), which aggregates in the cell bodies and dendrites of Purkinje cells from SCA31 affected individuals. Given that SCA31 and SCA37 share characteristics at the genetic and RNA level, it is hypothesized that transcription of the AUUUC repeat in the sense orientation produces an unexpected pentapeptide (poly(ISFHF)). Transcription of the UGGAA repeat in the antisense orientation might originate another unexpected and potentially aggregating pentapeptide (poly(FHSIP)); transcription of the AUUUC repeat in the antisense orientation contains several stop codons in all frames.

5. Aims

The noncoding $(ATTTC)_n$ insertion causing SCA37 is transcribed in a toxic $(AUUUC)_n$ RNA that abnormally initiates a cascade of pathogenic RNA-mediated mechanisms not yet fully understood. A common pathogenic mechanism associated with diseases caused by repeat expansions and insertions is the abnormal translation of the pathogenic repeat, by repeat associated non-AUG dependent (RAN) translation. In these diseases, the polypeptides originated by RAN translation are usually toxic contributing to the pathogenic phenotype. To give insights on the pathogenic mechanisms triggered by the $(AUUUC)_n$ RNA, the aim of this work is to investigate if the RAN translation of the $(AUUUC)_n$ is a pathogenic mechanism associated with SCA37, following the specific aims:

- Generation of constructs to investigate *in vitro* the occurrence of RAN translation in SCA37;
- Detection of poly(isoleucine-serine-phenylalanine-histidine-phenylalanine) (poly(ISFHF)) in human cell lines overexpressing the SCA37 pentanucleotide repeats.

MATERIALS AND METHODS

1. Generation of constructs to investigate pentanucleotide repeat translation in SCA37

1.1. Cloning of SCA37 pathogenic and nonpathogenic alleles

To determine whether the SCA37 $(ATTTC)_n$ insertion is abnormally translated in a poly(ISFHF), *DAB1* normal and mutant alleles were cloned in the mammalian expression vectors pCDH-CMV-MCS-EF1α-GreenPuro (for simplicity, this vector is referred as pCDH) and pCDNA3 (Appendix A). Normal $(ATTTT)_7$ and $(ATTTT)_{139}$ alleles, as well as the mutant $(ATTTT)_{57}(ATTTC)_{58}(ATTTT)_{73}$ $[(ATTTC)_{ins}]$ pentanucleotides were amplified from the previously described pCDH-CMV-MCS-EF1-GFP-T2A-Puro vectors harbouring these alleles[13]. These pentanucleotide repeats were amplified by Long Range PCR using 0.05 pg of each vector, 200 µM dNTPs, 1X PrimeSTAR GXL Buffer Mg2+ Plus, 1.25 U PrimeSTAR GXL DNA polymerase (Takara Bio), 0.2 µM primers STR24_GXL3_RAN_F and STR24_GXL3_RAN_R (Appendix B) and water up to 50 µL. After 1 minute of initial denaturation at 98 °C, DNA samples underwent 28 cycles of amplification (98 °C for 10 seconds and 68 °C for 1 minute and 50 seconds). To confirm the size of amplified alleles, 20 µL of PCR products were subjected to electrophoresis in 1% agarose gel, the remaining volume of PCR reactions was purified with DNA Clean & Concentrator-5 (Zymo Research) following the manufacturer instructions.

The STR24_GXL3_RAN_F and STR24_GXL3_RAN_R primers, used for amplification of *DAB1* pentanucleotide alleles, were designed with a 5'-tail compatible with *EcoRI* or *NotI* restriction enzymes. To ligate the pentanucleotide repeat alleles into pCDH and pCDNA3 vectors, 1.1 µg of each *DAB1* purified allele and 6 µg of each vector (pCDH or pCDNA3) were digested with 30 U of *EcoRI* (Anza), 30 U of *NotI* (Anza), 1X Anza Red Buffer and water up to 20 µL, overnight (O/N) at 37 °C. The restriction products were then separated in 1% agarose gel electrophoresis and purified with the Zymoclean Gel DNA Recovery Kit (Zymo Research), following the manufacturer instructions.

The digested nonpathogenic and pathogenic alleles were ligated into the linearized pCDH and pCDNA3 vectors with insert to vector ratio of 3:1. The appropriate mass of insert was determined with formula (1) for 300 ng of vector. Short normal $(ATTTT)_{7,}$ large normal $(ATTTT)_{139}$ and the pathogenic $(ATTTC)_{ins}$ were ligated into 300 ng of linearized pCDH and

pCDNA3 with 2 U T4 DNA Ligase (Roche), 1X T4 DNA Ligase buffer and water up to 20 µL, O/N at 4 °C.

$$ng\ insert = \frac{ng\ vector * kb\ insert}{kb\ vector} * ratio\frac{insert}{vector} \qquad (1)$$

Large normal alleles and repeat insertion alleles are prone to drastically contract during cloning. To reduce the repeat instability, different cloning protocols suitable for the amplification of vectors with different repeat size alleles were used. To clone the nonpathogenic $(ATTTT)_7$ allele, ligations were transformed into 100 µL of NZYStar competent cells (NZYTech) and incubated O/N at 37 °C in lysogeny broth (LB) agar plates with 100 µg/mL ampicillin. On the next day, colonies were picked to 5 mL of LB with 100 µg/mL ampicillin and incubated O/N at 37 °C with agitation at 200 rpm. To clone the $(ATTTT)_{139}$ allele, ligations were transformed into 100 µL of NZYStar competent cells (NZYTech) and incubated O/N at 37 °C in LB agar plates containing 100 µg/mL ampicillin. On the next day, colonies were picked to 5 mL of LB with 100 µg/mL ampicillin and incubated O/N at 30 °C, with agitation at 200 rpm. The reduction in the incubation temperature from 37 ºC to 30 ºC allows to decrease the bacteria growing rate and consequently the repeat contraction events[161]. To clone the SCA37 $(ATTTC)_{ins}$ allele, ligations were transformed into 100 µL of Stbl3 competent cells (Invitrogen). The Stbl3 is an *E. coli* strain that contains a genetically modified *recA* gene, a gene essential for DNA recombination. This strain has a reduced plasmid recombination ability, which is suitable for cloning repetitive sequences[162]. The transformed bacteria were incubated O/N at 30 °C in LB agar plates containing 100 µg/mL ampicillin. On the next day, colonies were picked to 5 mL of liquid LB with 100 µg/mL ampicillin and incubated O/N at 30 °C with agitation at 200 rpm. After cloning, plasmid DNA was isolated with NZYMiniprep (NZYTech) following manufacturer instructions.

Since $(ATTTT)_{139}$ and $(ATTTC)_{ins}$ were unstable and prone to contract, to confirm the repeat size, 500 ng of each plasmid were digested with 10 U of *EcoRI* (Anza), 10 U of *NotI* (Anza), 1X Anza Red Buffer and water up to 20 µL, for 3h, at 37 °C. Restriction products were separated by 1% agarose gel electrophoresis.

The resulting pCDH and pCDNA3 plasmids with $(ATTTT)_7$, $(ATTTT)_{139}$ and $(ATTTC)_{ins}$ alleles are thereafter referred as pCDH-Rep and pCDNA3-Rep.

1.2. Insertion of the stop codon cassette upstream the *DAB1* pentanucleotide alleles

To ensure that the repeat translation only initiates in the *DAB1* pentanucleotide repeat or in the flanking region, a stop codon cassette was cloned in pCDH-Rep and pCDNA3-Rep vectors, upstream the SCA37 pathogenic or nonpathogenic alleles. For that, two sets of oligonucleotides, with overhangs compatible with the multiple cloning sites of pCDH and pCDNA3, were designed encoding 6 TAG stop codons, 2 in each reading frame, as previously described[109].

For the insertion of the 6xSTOP cassette in pCDH plasmids, the oligonucleotides 6xSTOP_RAN_pCDH_F and 6xSTOP_RAN_pCDH_R (Appendix B), containing overhangs compatible to *XbaI* and *EcoRI* restriction sites respectively, were used. For the insertion of 6xSTOP cassette in pCDNA3 plasmids, the oligonucleotides 6xSTOP_RAN_pCDNA3_F and 6xSTOP_RAN_pCDNA3_R (Appendix B) containing overhangs compatible to *HindIII* and *EcoRI* restriction sites were used. The annealing of each pair of oligonucleotides was performed with 3 µM of each oligonucleotide, 1X annealing buffer (4.8 mM Tris pH 8, 24 mM NaCl and 0.48 mM EDTA) and water up to 50 µL. The annealing mixes were heated for 5 min at 95 °C followed by a temperature decrease to 85 °C at 2 °C/s and then to 25 °C at 0.1 °C/s.

To clone the annealed 6xSTOP cassette in pCDH-Rep vectors, 5.4 µg of each plasmid was digested O/N at 37 °C with 33.5 U of *XbaI* (Anza), 33 U of *EcoRI* (Anza), 1X Anza Red Buffer and water up to 50 µL. The restriction products were then purified with DNA Clean & Concentrator-5 (Zymo Research) and ligated with the annealed stop codon cassette, with an insert to vector ratio of 3:1 (formula (1)), for 300 ng of vector, using 2 U T4 DNA Ligase (Roche), 1X T4 DNA Ligase buffer and water up to 20 µL, O/N at 4 °C.

To clone the 6xSTOP cassette in pCDNA3-Rep vectors, 3 µg of each plasmid was digested O/N at 37 °C with 30 U of *HindIII* (Anza), 30 U of *EcoRI* (Anza), 1X Anza Red Buffer and water up to 50 µL. The restriction products were then purified with DNA Clean & Concentrator-5 (Zymo Research) and ligated with the annealed stop codon cassette. The ligation reactions were carried as described for the ligation of 6xSTOP cassette in pCDH-Rep vectors. Ligations containing short normal, large normal or SCA37 pathogenic allele were transformed as described in section 1.1.

Upon ligation of the 6xSTOP codon cassette, the *XbaI* restriction site was abolished in pCDH-6xSTOP-Rep and the *HindIII* recognition site was disrupted in pCDNA3-6xSTOP-Rep vectors. To confirm the insertion of the 6xSTOP cassette in pCDH-Rep and in pCDNA3-Rep, 500 ng of vectors were digested with 5 U of *XbaI* (Anza) or 10 U of *HindIII*

(Anza), 1X Anza Red Buffer and water up to 20 µL for 3h at 37 °C, using pCDH or pCDNA3 empty vector as positive control. Restriction products were subjected to electrophoresis in 1% agarose gel to confirm the ligation of the 6xSTOP codons into the vectors.

1.3. Insertion of 3 tags downstream the *DAB1* pentanucleotide repeats in pCDH-6xSTOP-Rep vectors

To detect the SCA37 RAN peptides in cell lines transfected with *DAB1* pathogenic allele and to disclose which repeat reading frames were subjected to RAN translation, three epitopes, HA, flag and myc were cloned into pCDH-6xSTOP-Rep downstream the pentanucleotide repeat, one in each reading frame.

A pair of oligonucleotides, Tags_RAN_pCDH_F and Tags_RAN_pCDH_R (Appendix B) encoding HA, flag and myc tags, were annealed as described in section 1.2. These oligonucleotides were previously described[109] and contained overhangs compatible with the *NotI* and *SfaAI* restriction sites, located in the multiple cloning site of the pCDH vector downstream the cloned *DAB1* pentanucleotide repeats.

To insert the annealed oligonucleotides encoding the tags in pCDH-6xSTOP-Rep, 3 µg of each plasmid were digested with 30 U *NotI* (Anza), 30 U *SfaAI* (Anza), 1X Anza Red Buffer and water up to 50 µL, O/N at 37 °C. The digested DNA was then purified with DNA Clean & Concentrator-5 (Zymo Research) and ligated with the annealed tags with insert to vector ratio of 3:1. Ligations were carried out using 300 ng of linearized pCDH-6xSTOP-Rep, the calculated amount of the 3Tags annealed oligonucleotides (formula (1)), 2 U T4 DNA Ligase (Roche), 1X T4 DNA Ligase buffer and water up to 20 µL, O/N at 4 °C. Ligations were transformed as described in section 1.1.

Upon ligation of the tags, both *NotI* and *SfaAI* restriction sites were abolished in pCDH-6xSTOP-Rep vectors. To confirm the ligation of the tags, 500 ng of pCDH-6xSTOP-Rep-3Tags vectors were digested with 10 U of *NotI* (Anza), 1X Anza Red Buffer and water up to 20 µL, for 3h at 37 °C, using pCDH empty vector as positive control of the restriction reaction. Restriction products were separated by 1% agarose gel electrophoresis to detect the size of the restriction fragments.

1.4. Insertion of 3 tags downstream the *DAB1* pentanucleotide repeats in pCDNA3-6xSTOP-Rep vectors

For insertion in pCDNA3 vectors, the 3 tags were amplified by conventional PCR using 0.4 ng of pCDH-6xSTOP-$(ATTTT)_7$-3Tags plasmid, 0.32 µM of primers

Tags_RAN_pCDNA3_F and Tags_RAN_pCDNA3_R (Appendix C), 1X NZYTaq II Green Master Mix (NZYTech) and water up to 25 µL. After 3 minutes of initial denaturation at 95 °C, DNA samples underwent 30 cycles of amplification (94 °C for 30 seconds, 62 °C for 30 seconds and 72 °C for 45 seconds) followed by a final extension at 72 °C for 10 minutes. The primer Tags_RAN_pCDNA_F contained an overhang compatible with the *NotI* restriction site and the primer Tags_RAN_pCDNA3_R contained an overhang compatible with the *XbaI* restriction site.

To insert the tags in pCDNA3-6xSTOP-Rep, 3 µg of each plasmid and the tags amplified by PCR were digested O/N, at 37 °C, with 30 U *NotI* (Anza), 30 U *XbaI* (Anza), 1X Anza Red Buffer and water up to 50 µL. The digested DNA was then purified with DNA Clean & Concentrator-5 (Zymo Research). The tags with *NotI* and *XbaI* overhangs were then ligated into the linearized vectors with insert to vector ratio of 3:1 (formula (1)) using 300 ng of each linearized vector. Ligations were performed with 2 U T4 DNA Ligase (Roche), 1X T4 DNA Ligase buffer and water up to 20 µL, O/N, at 4ºC and cloned as described in section 1.1.

The cloning of the tags in pCDNA3 vectors inserted a new *BpiI* restriction site in the constructs. Thus, to confirm the insertion of the tags, 500 ng of pCDNA3-6xSTOP-Rep-3Tags were digested with 5 U of *BpiI* (Anza), 1X Anza Red Buffer and water up to 20 µL, for 3h at 37 °C. Restriction products were analysed in 1% agarose gel electrophoresis.

1.5. Sanger sequencing

Sanger sequencing was performed to confirm the sequence of the 6xSTOP, *DAB1* pentanucleotide repeats and 3Tags in pCDH and pCDNA3 constructs generated in this work, using the primers in Appendix C.

The Sanger sequencing reaction is based on the random addition of fluorescently labelled terminators of the DNA chain, dideoxyribonucleotides (ddNTPs), during the polymerization of a new DNA strand. In Sanger sequencing reaction, the DNA polymerase adds deoxyribonucleotides (dNTPs) to a growing DNA strand by catalysing the formation of a phosphodiester bond between the free 3'-OH group of the last nucleotide and the 5'-phosphate of the next. When a ddNTP is incorporated, the strand polymerization is terminated due to the lack of a 3'-OH group in ddNTPs, which results in fragments of a DNA chain with random lengths[163]. These DNA sequences are then subjected to fluorescence capillary electrophoresis in which the fluorescence of each labelled ddNTP is detected and

the output is an electropherogram with different colours corresponding to each nucleotide in the template DNA sequence.

Sanger sequencing reactions were performed with 400 ng of plasmid DNA, 0.5 µM of primer, 0.5X Sequencing Buffer (Applied Biosystems), 1 µL BigDye™ Terminator v3.1 (Applied Biosystems) and water up to 10 µL. After 1 minute at 95 °C, DNA samples underwent 35 cycles of 95 °C for 30 seconds, 56 °C for 10 seconds and 60 °C for 2.5 minutes followed by a final extension step at 60 °C for 10 minutes.

Due to the large repetitive sequence of *DAB1* nonpathogenic $(ATTTT)_{139}$ and pathogenic $(ATTTC)_{ins}$ alleles, these repeats were sequenced with a different reaction mix and program. This reaction mix was prepared with 400 ng of plasmid DNA, 2.5 µM of primer, 0.5X Sequencing Buffer (Applied Biosystems), 4 µL BigDye™ Terminator v3.1 (Applied Biosystems) and water up to 20 µL. After 5 minutes at 95 °C, DNA samples underwent 50 cycles of 95 °C for 30 seconds, 56 °C for 10 seconds and 60 °C for 4 minutes, followed by a final extension step at 60 °C for 10 minutes.

Sanger sequencing products were subjected to capillary electrophoresis in an ABI3730xl DNA Analyzer, on a 36 cm capillary (Applied Biosystems), which only allows to sequence until 600-700 bp. Thus, in samples containing nonpathogenic alleles $(ATTTT)_7$ or $(ATTTT)_{139}$, the STR24_RAN_F primer could sequence the *DAB1* pentanucleotide repeat completely. In samples containing the pathogenic allele $(ATTTC)_{ins}$, the regions sequenced with STR24_RAN_F and STR24_RAN_R partially overlapped providing the sequence of the entire mutant *DAB1* pentanucleotide repeat.

1.6. Isolation of plasmid DNA for transfection of human cell lines - Midiprep

To obtain the amount of plasmid DNA required for the transfection of human cell lines, midiprep was performed for the 6 generated constructs. Thus, pCDH-6xSTOP-Rep-3Tags and pCDNA3-6xSTOP-Rep-3Tags vectors were transformed following the procedure described in section 1.1, however, to perform midiprep, colonies were picked to 100 mL of LB media containing 100 µg/mL of ampicillin. On the next day, 5 mL of liquid culture were used to perform plasmid DNA isolation by miniprep with NZYMiniprep (NZYTech), to confirm the sequence of the 6xSTOP, *DAB1* pentanucleotide repeats and 3Tags, in the constructs. The pentanucleotide repeat size was confirmed as described in section 1.1. The stop cassette, repeat and tags sequences were confirmed by Sanger sequencing as described in section 1.5. After confirming the sequence of the inserts cloned in the vectors,

the plasmid DNA from the remaining 95 mL of the liquid cultures was isolated by midiprep with NucleoBond® Xtra Midi (Takara Bio) following manufacturer instructions.

2. Cell culture

HEK293T cells were maintained in Dulbecco's modified eagle medium (DMEM) supplemented with GlutaMAX (Gibco), 10% fetal bovine serum (FBS) and 1X antibiotic/antimycotic (100 U/mL penicillin, 100 µg/mL streptomycin and 250 ng/mL Amphotericin B; Gibco), at 37 °C and 5% CO_2.

Human neural stem cells (hNSC) Cb192 were maintained at 37 °C and 5% CO_2, in DMEM:F12-GlutaMAX (Gibco) supplemented with 1X N-2 (Invitrogen), 0.05X B-27 (Invitrogen), 10 ng/mL EGF (Peprotech), 10 ng/mL FGF (Peprotech) and 100 U/mL penicillin/streptomycin (Gibco), as previously described[164].

The absence of mycoplasma contamination was confirmed every month in cultured cell lines, by PCR.

2.1. Transfection of HEK293T cells

HEK293T cells were transfected with the pCDH-6xSTOP-Rep-3Tags and pCDNA3-6xSTOP-Rep-3Tags constructs generated in this work, using FuGENE HD (Promega) in 6- and 24-well plates.

Prior to transfection, $1.5x10^5$ HEK293T cells were plated per well of 6-well plates and $3x10^4$ HEK293T cells were plated per well of 24-well plates. Transfection mixes were performed with a plasmid:transfection reagent ratio of 3:1, following the manufacturer instructions. Each transfection mix was added to the respective well of the plate and 6 hours after the addition of the transfection complexes, culture medium was replaced by fresh medium to stop transfection. Cells transfection was confirmed 72h post-transfection. For cells transfected with pCDH-6xSTOP-Rep-3Tags vectors, since this vector encodes green fluorescence protein (GFP), the transfection was confirmed by the detection of GFP positive cells with a ZOE Fluorescent Cell Imager (Bio-Rad). pCDNA3-6xSTOP-Rep-3Tags vectors do not encode GFP, thus the transfection efficiency of these constructs was estimated using a positive control in the transfection, a pCDH vector that was transfected in a parallel well.

2.2. Transfection of neural stem cells Cb192

Cb192 cells were transfected with pCDH-6xSTOP-Rep-3Tags using Lipofectamine 2000 (Invitrogen), as previously described[164]. Prior to cell plating, the wells of a 6-well plate were treated with 0.01% poly-L-lysine for 30 minutes at 37 °C and followed by 2 µg/mL laminin for 1h at 37 °C. After treatment, $7.5x10^5$ Cb192 cells were plated per well. Transfection mixes were performed with a ratio of plasmid to Lipofectamine 2000 of 5:4. Transfection was confirmed 72h post-transfection with ZOE Fluorescent Cell Imager (Bio-Rad) through the detection of GFP expressing cells.

2.3. Fluorescence *in situ* hybridization (FISH)

To confirm the expression of SCA37 repetitive RNA in transfected cells, RNA FISH was performed in HEK293T cells transfected with pCDH-6xSTOP-Rep-3Tags and pCDNA3-6xSTOP-Rep-3Tags constructs, using an OMe-$(GAAAT)_5$ probe labelled with Texas Red.

HEK293T cells were transfected as described in section 2.1. The cells transfected for RNA FISH were plated in glass coverslips previously treated with 0.01% poly-L-lysine in 24-well plates. Seventy-two hours post-transfection, HEK293T cells were fixed in 4% paraformaldehyde for 15 minutes and washed twice with 1X PBS. After fixation, cells were permeabilized with 2% acetone for 5 minutes followed by 3 washes with DEPC-treated 1X PBS. After permeabilization, coverslips were incubated in prehybridization solution (30% formamide and 2X SSC) for 10 minutes. After this, coverslips were incubated in hybridization solution (0.5 ng/µL OMe-$(GAAAT)_5$ probe, 30% formamide, 2X SSC, 0.02% BSA, 2 mM ribonucleoside vanadyl complex and 66 µg/mL yeast tRNA) in a dark wet chamber for 2h, at 37 °C. The coverslips were incubated in prehybridization solution, for 30 minutes at 37 °C and washed twice in 2X SSC. Coverslips were counterstained with DAPI (1 mg/mL) for 10 minutes, washed twice with 1X PBS and mounted in a glass slide on ibidi Mounting Medium (ibidi). The FISH staining signal was detected in a Leica DMI6000 FFW microscope with a 63X glycerol objective, images were analysed with Fiji[165].

2.4. Cell extracts and protein quantification

Transfected HEK293T and Cb192 cells were lysed 72 hours post-transfection. The cells plated in each well of a 6-well plate were washed twice with cold 1X PBS and lysed

with 200 µL RIPA buffer (150 mM NaCl, 1.0% IGEPAL® CA-630, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris, pH 8.0; Sigma Aldrich) containing 1X cOmplete protease inhibitor (Roche). The lysate was scraped from each well, with a cell lifter, transferred to a 1.5 mL tube and incubated on ice for 30 minutes. Cell lysates were sonicated twice for 10 seconds, at output 3 and duty cycle 30%. The samples were on ice during sonication to make sure that they did not overheat because of the vibrations generated by the sonication process. Then, cell lysates were centrifugated for 20 minutes, at 10 000 g, at 4 °C, to pellet the cell debris. Clarified protein extracts were aliquoted and stored at -80 °C.

Total protein quantification was performed using the Pierce BCA Protein Assay Kit (Thermo Scientific), following the manufacturer instructions (Appendix D). This quantification method relies on the biuret reaction that consists in the ability of proteins to reduce Cu^{+2} to Cu^{+1} in an alkaline medium. Cu^{+1} ions are chelated by bicinchoninic acid (BCA) molecules originating a purple compound that absorbs strongly at 562 nm. A series of bovine serum albumin (BSA) dilutions of known concentrations were used during the quantification of the protein of interest, to create a standard curve of absorbance vs concentration, according to formula (2).

$$Absorbance\ (nm) = m * Concentration\ (\mu g/mL) + b \qquad (2)$$

After establishing a standard curve, the slope (m) and y-intercept (b) of the standard curve were used to calculate the concentration of the protein extracts of interest with formula (3).

$$Concentration\ (\mu g/mL) = \frac{Absorbance\ (nm) - b}{m} \qquad (3)$$

2.5. Dot blot

To investigate whether the *DAB1* pentanucleotide repeat alleles are subjected to RAN translation, dot blot was performed. Thus, nitrocellulose membranes were prepared with 1 cm in height and 1.5 cm/sample in length. Then, 40 µg of protein extracts from transfected HEK293T or Cb192, as well as 5 µg of extracts from HEK293T cells transfected with plasmids encoding HA, flag and myc (positive controls) were dropped in a designated space of the membrane. After the membranes air-dried for 30 minutes, the membranes were

blocked for 1h with 5% powder milk in 1X TBS-T. Then, the membranes were incubated O/N, at 4 °C with primary monoclonal mouse α-HA, α-flag, α-myc, or our primary polyclonal rabbit antibodies generated against the putative RAN pentapeptide poly(ISFHF) (α-RAN RB7543 and α-RAN RB7544; Biomatik). The antibody references and dilutions used are detailed in Appendix E. On the next day, membranes were washed 4 times with 1X TBS-T, incubated for 1h with an α-mouse or α-rabbit secondary antibody coupled with horseradish peroxidase (HRP) (Appendix E) and then washed again 4 times with 1X TBS-T. After this, the HRP substrate was added to the membrane and chemiluminescence was detected on ChemiDoc XRS+ (Bio-Rad), after 5 minutes of exposure for membranes incubated with α-RAN RB7543 and α-RAN RB7544 antibodies and 25 minutes of exposure for membranes incubated with α-HA, α-flag and α-myc.

2.6. Western blot

To analyse the translation of RAN peptides from the SCA37 pentanucleotide repeats, 60 µg of protein extracts from HEK293T transfected cells and 10 µg of positive controls (section 2.5) were desaturated at 95 ºC and then resolved by SDS-PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis) in a 12% polyacrylamide gel (Appendix F). SDS is a detergent that disrupts the tertiary structures of proteins and coats them with a negative charge, allowing them to migrate through a polyacrylamide gel under an electrical field, from the negatively charged cathode toward the positively charged anode. After SDS-PAGE, the proteins were transferred onto a PVDF membrane using a Hoefer SemiPhor TE77 semi-dry transfer unit for 1h at 45 mA, followed by 1h at 85 mA. After proteins transfer to membranes, membranes were blocked for 1h with 5% powder milk in 1X TBS-T and then incubated O/N, at 4 °C, with primary monoclonal mouse α-HA, α-flag and α-myc antibodies (Appendix E). On the next day, membranes were washed 4 times with 1X TBS-T, incubated for 1h with an α-mouse secondary antibody coupled with HRP and then washed an additional 4 times with 1X TBS-T. Finally, the HRP substrate was added to membranes and chemiluminescence was detected on ChemiDoc XRS+ (Bio-Rad) after 25 minutes of exposure.

RESULTS

1. Generation of constructs for *in vitro* detection of RAN translation in SCA37

To investigate whether the noncoding $(ATTTC)_n$ insertion causing SCA37 may initiate RAN translation, I generated several constructs with the SCA37 $(ATTTC)_n$ insertion [$(ATTTC)_{ins}$], the short $(ATTTT)_7$ and the large $(ATTTT)_{\sim 120}$ nonpathogenic alleles. I developed these constructs based on the methodology used to investigate RAN translation, *in vitro*, in other diseases caused by repeat expansions, such as SCA8, DM1, SCA2 and SCA3[85, 109, 113]. I cloned the *DAB1* pentanucleotide repeats together with a flanking region with 66 bp upstream the repeat and 30 bp downstream the repeat, in pCDH-CMV-MCS-EF1α-GreenPuro (referred as pCDH for simplicity) and pCDNA3 backbones (Figure 15). To ensure that the repeat translation initiates in the repetitive region or in the flanking region, I cloned a stop codon cassette upstream the repeat. Since the ATTTC repeat encodes the same poly(ISFHF) in the three reading frames (Table 2), to disclose which reading frame(s) were subjected to RAN translation, I also inserted 3 epitope tags, HA, flag and myc, downstream the repeat, one in each reading frame (Figure 15).

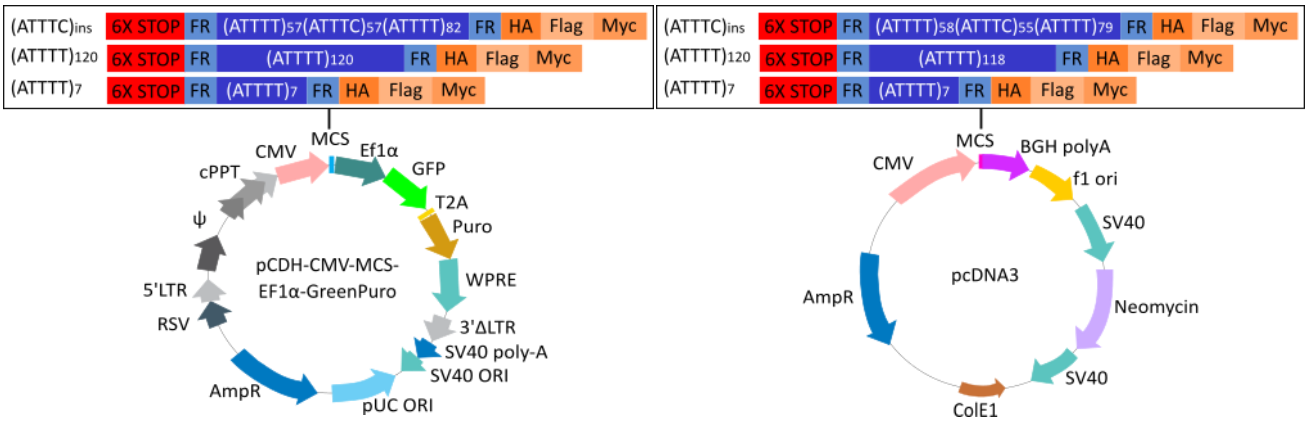


Figure 15. Schematic representation of the constructs developed to investigate RAN translation in SCA37. The short nonpathogenic $(ATTTT)_7$, large nonpathogenic $(ATTTT)_{\sim 120}$ and the mutant $(ATTTC)_{ins}$ and its flanking regions (dark and light blue boxes) were cloned downstream 6 stop codons (red box) and upstream HA, flag and myc epitopes (orange boxes) in pCDH-CMV-MCS-EF1α-GreenPuro and pCDNA3 backbones. Due to repeat size instability, the large nonpathogenic allele had 120 ATTTT repeat units in pCDH and 118 ATTTTs in pCDNA3 whereas the pathogenic allele had the configuration $(ATTTT)_{57}(ATTTC)_{57}(ATTTT)_{82}$ in pCDH and $(ATTTT)_{58}(ATTTC)_{55}(ATTTT)_{79}$ in pCDNA3.

The *DAB1* large $(ATTTT)_{139}$ nonpathogenic allele and the mutant $(ATTTT)_{57}(ATTTC)_{58}(ATTTT)_{73}$ allele were highly unstable during cloning. The number of repeats in the large nonpathogenic allele contracted from 139 ATTTTs to 120 ATTTTs in pCDH and to 118 ATTTTs in pcDNA3 backbone (Appendix G). Regarding the mutant $(ATTTT)_{57}(ATTTC)_{58}(ATTTT)_{73}$ allele I obtained different allele configurations in the different backbones, $(ATTTT)_{57}(ATTTC)_{57}(ATTTT)_{82}$ in pCDH and $(ATTTT)_{58}(ATTTC)_{55}(ATTTT)_{79}$ in pCDNA3 (Appendix G). Therefore, I obtained a total of 6 constructs: pCDH-6xSTOP-$(ATTTT)_{7}$-3Tags; pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags; pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags; pCDNA3-6xSTOP-$(ATTTT)_{7}$-3Tags; pCDNA3-6xSTOP-$(ATTTT)_{118}$-3Tags; and pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags.

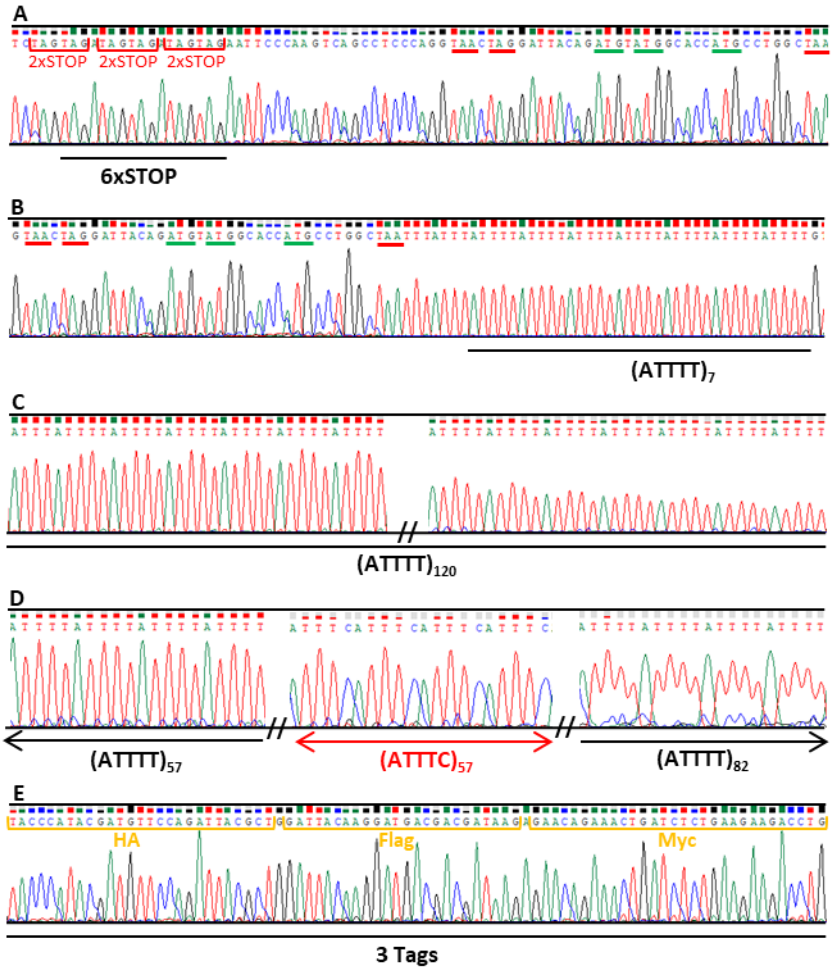


Figure 16. Representative Sanger sequencing electropherograms of the inserts cloned in pCDH and pCDNA3 vectors. The full electropherograms are available in Appendix G; **(A)** Stop codon cassette encoding six stop codons, two stop codons in each reading frame; **(B)** Short nonpathogenic $(ATTTT)_{7}$ allele; **(C)** Large nonpathogenic allele with 120 ATTTT repeats; **(D)** Pathogenic allele with the configuration $(ATTTT)_{57}(ATTTC)_{57}(ATTTT)_{82}$; **(E)** DNA sequence encoding

HA, flag and myc epitopes (3Tags), one tag in each reading frame. Start and stop codons located in the flanking region upstream the *DAB1* pentanucleotide repeat are underlined in green and red respectively in **(A)** and **(B)**.

After the generation of these constructs, I performed an *in silico* analysis with the ExPASy Translate software, thus I confirmed the putative pentapeptides translated from the three reading frames of the *DAB1* pentanucleotide repeats fused with one tag (Appendix H). The cloned flanking region upstream the *DAB1* pentanucleotide repeat contained start and stop codons in two reading frames (Figure 16 and Table 2). In one frame, the ATG start codon was followed by a TAA stop codon only 7 bp upstream of the repeat (Figure 16), therefore, as expected, according to the *in silico* analysis using ExPASy Translate software, no peptides were expected to be translated from this reading frame (Table 2 and Appendix H). In the other frame the ATG codon was not followed by any stop codon (Figure 16 and Table 2); however, this ATG codon was not located within a Kozak sequence ((GCC)GCCRCCAUGG) and the translation initiation from this codon would not be efficient (Figure 16). Interestingly, the presence of this ATG upstream the pentanucleotide repeat suggested that the ATTTT repeat may also be translated in the context of the $(ATTTC)_{ins}$ in poly-$(ILFYF)_n(ISFHF)_n(ILFYF)_n$ or even in the context of the large nonpathogenic repeat in poly$(ILFYF)_n$. The putative pentapeptides translated from the three frames of the inserts cloned in pCDH-6xSTOP-Rep-3Tags and pCDNA3-6xSTOP-Rep-3Tags vectors are presented in Table 2 and Appendix H.

Table 2. Nucleotide sequence and predicted peptide sequence from each frame of each construct.

Vector	Alelle	Reading frame	Nucleotide sequence ↳ Predicted peptide sequence
pCDH	$(ATTTT)_7$	1	TAGTAG-TAG-ATG-$(ATTTT)_7$-Myc ↳ M-$(ILFYF)_2$IL-Myc
		2	TAGTAG-$(ATTTT)_7$-HA ↳ FYF(ILFYF)ILF-HA
		3	TAGTAG-TAA-ATG-ATG-TAA-$(ATTTT)_7$-Flag ↳ M-M-STOP
	$(ATTTT)_{120}$	1	TAGTAG-TAG-ATG-$(ATTTT)_{120}$-HA ↳ M-$(ILFYF)_{40}$-HA
		2	TAGTAG-$(ATTTT)_{120}$-Flag ↳ FYF$(ILFYF)_{39}$IL-Flag
		3	TAGTAG-TAA-ATG-ATG-TAA-$(ATTTT)_{120}$-Myc ↳ M-M-STOP
	$(ATTTC)_{ins}$	1	TAGTAG-TAG-ATG-$(ATTTT)_{57}(ATTTC)_{57}(ATTTT)_{82}$-Flag ↳ M-$(ILFYF)_{19}(ISFHF)_{19}(ILFYF)_{26}$IL-Flag
		2	TAGTAG-$(ATTTT)_{57}(ATTTC)_{57}(ATTTT)_{82}$-Myc ↳ FYF$(ILFYF)_{18}$ILFHF$(ISFHF)_{18}$ISFYF $(ILFYF)_{26}$ILF-Myc

		3	TAGTAG-TAA-ATG-ATG-TAA-$(ATTTT)_{57}(ATTTC)_{57}(ATTTT)_{82}$-HA ↳ M-M-STOP
pCDNA3	$(ATTTT)_7$	1	TAGTAG-TAA-ATG-ATG-TAA-$(ATTTT)_7$-Myc ↳ M-M-STOP
		2	TAGTAG-TAG-ATG-$(ATTTT)_7$-HA ↳ M-$(ILFYF)_2$IL-HA
		3	TAGTAG-$(ATTTT)_7$-Flag ↳ LFYF(ILFYF)ILF-Flag
	$(ATTTT)_{118}$	1	TAGTAG-TAA-ATG-ATG-TAA-$(ATTTT)_{118}$-Myc ↳ M-M-STOP
		2	TAGTAG-TAG-ATG-$(ATTTT)_{118}$-HA ↳ M-$(ILFYF)_{39}$IL-HA
		3	TAGTAG-$(ATTTT)_{118}$-Flag ↳ LFYF$(ILFYF)_{38}$ILF-Flag
	$(ATTTC)_{ins}$	1	TAGTAG-TAA-ATG-ATG-TAA-$(ATTTT)_{58}(ATTTC)_{55}(ATTTT)_{79}$-Flag ↳ M-M-STOP
		2	TAGTAG-TAG-ATG-$(ATTTT)_{58}(ATTTC)_{55}(ATTTT)_{79}$-Myc ↳ M-$(ILFYF)_{19}$ILFHF$(ISFHF)_{18}(ILFYF)_{26}$-Myc
		3	TAGTAG-$(ATTTT)_{58}(ATTTC)_{55}(ATTTT)_{79}$-HA ↳ LFYF$(ILFYF)_{19}(ISFHF)_{18}(ILFYF)_{26}$IL-HA

Red- stop codons; green- start codons; blue- *DAB1* pentanucleotide repeat; orange- epitope tags.

2. Investigation of RAN translation in SCA37 using HEK293T cells

2.1. Transfection of HEK293T cells with constructs expressing *DAB1* pentanucleotide repeats

To investigate whether the SCA37 pentanucleotide repeat insertion was able to initiate RAN translation, I transfected HEK293T cells with pCDH-6xSTOP-$(ATTTT)_7$-3Tags; pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags; and pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags. I confirmed the success of transfection 72h post-transfection, at this time-point most of the cells were expressing GFP, indicating that the *DAB1* pentanucleotide repeats were being transcribed in these cells (Figure 17). I also transfected HEK293T cells with pCDNA3-6xSTOP-$(ATTTT)_7$-3Tags; pCDNA3-6xSTOP-$(ATTTT)_{118}$-3Tags; and pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags, however the pCDNA3 does not encode GFP, thus the success of the transfection was estimated using pCDH-6xSTOP-Rep-3Tags vectors as positive controls in the transfection. The high number of HEK293T cells expressing GFP suggests that these cells were overexpressing the *DAB1* pentanucleotide repeat alleles.

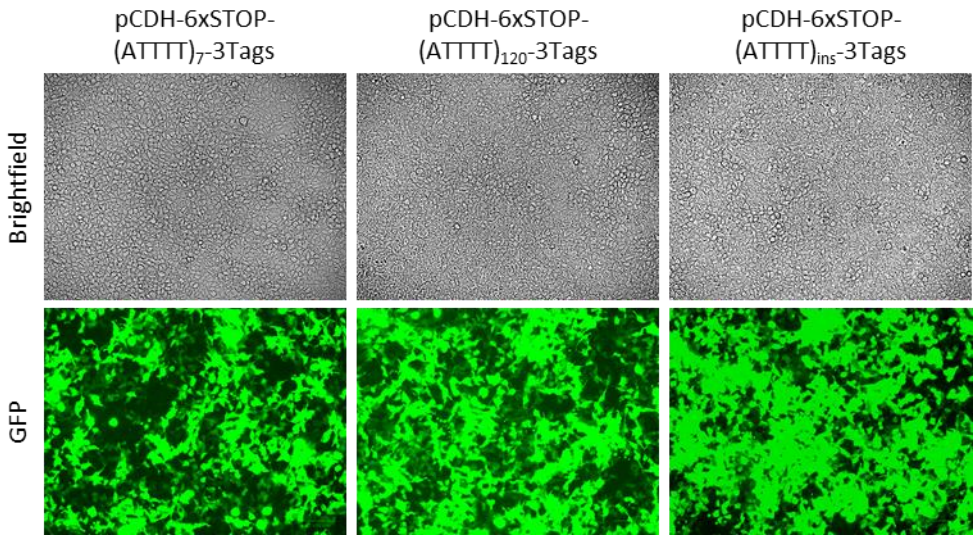


Figure 17. Transfection efficiency in HEK293T cells. Brightfield and fluorescence images at 72h post-transfection of HEK293T cells with pCDH-6xSTOP-Rep-3Tags vectors. Green cells were transfected cells expressing GFP encoded in the transfected vectors.

2.2. Expression of $(AUUUC)_n$ RNA in transfected HEK293T cells

To demonstrate that $(AUUUC)_n$ RNA was transcribed in HEK293T cells transfected with the pathogenic alleles in the different backbones (pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags and pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors), I performed RNA FISH with a OMe-$(GAAAU)_5$ probe specific to hybridize with the $(AUUUC)_n$ RNA, 72h post-transfection (Figure 18). HEK293T cells transfected with pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags and pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors showed $(AUUUC)_n$ RNA expression. In cells transfected with each of these vectors, the expressed $(AUUUC)_n$ RNA forms similar nuclear aggregates, which is in line with previous findings[13] (Figure 18). However, it was also possible to detect FISH signal spread over the cell nucleus, suggesting that the probe may be labelling free $(AUUUC)_n$ RNA in this compartment. In the cytoplasm of the transfected cells, I did not observe FISH staining with the OMe-$(GAAAU)_5$ probe. HEK293T cells transfected with pCDH-6xSTOP-$(ATTTT)_7$-3Tags; pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags; pCDNA3-6xSTOP-$(ATTTT)_7$-3Tags; and pCDNA3-6xSTOP-$(ATTTT)_{118}$-3Tags vectors did not show FISH signal (Figure 18).

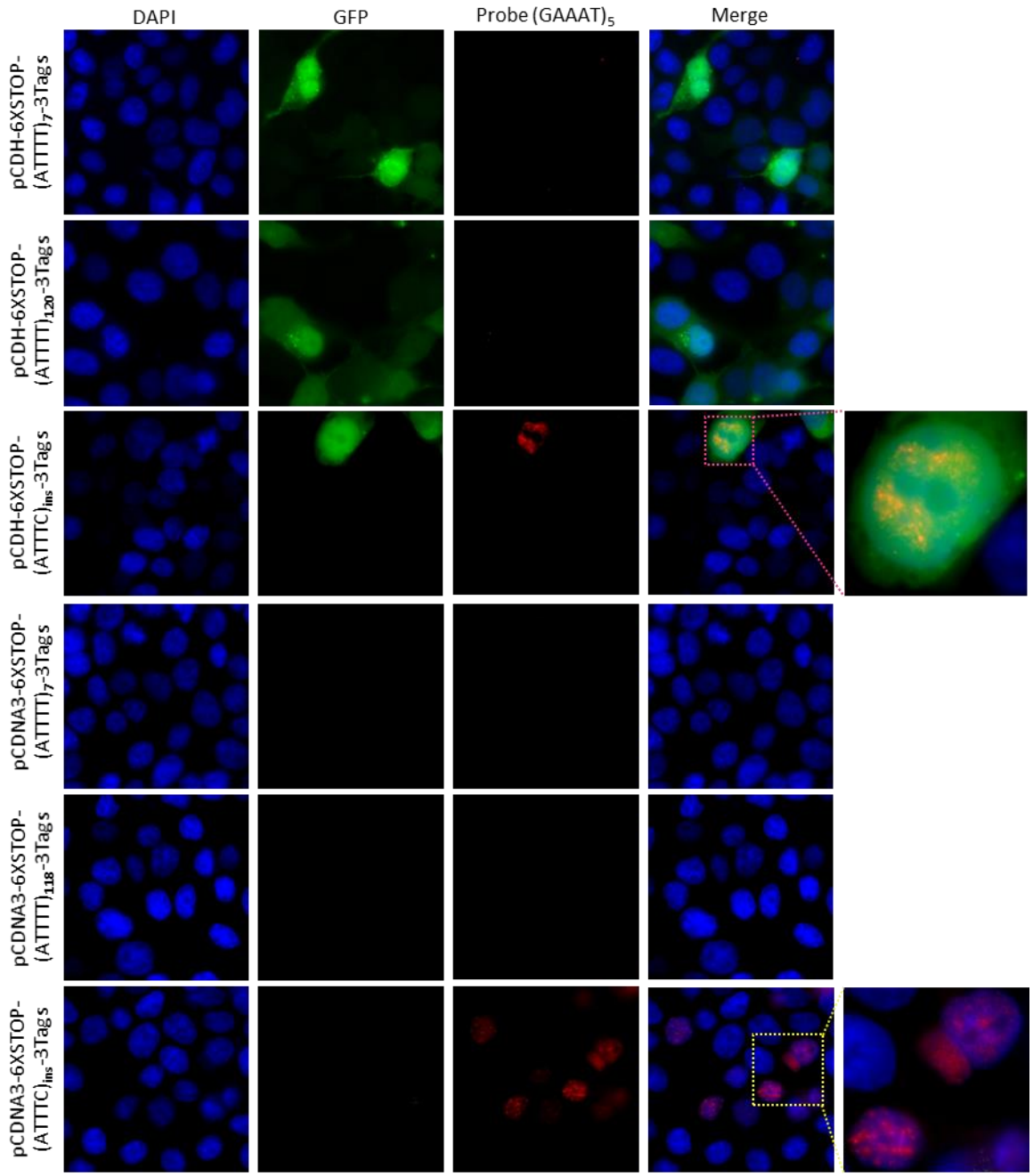


Figure 18. $(AUUUC)_n$ RNA expression in HEK293T transfected cells. Using a probe specific for $(AUUUC)_n$ RNA, I detected nuclear aggregates in HEK293T cells transfected with pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags and pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors. No $(AUUUC)_n$ RNA expression was detected in cells transfected with vectors containing short nonpathogenic $(ATTTT)_7$ or large nonpathogenic $(ATTTT)_{\sim 120}$ alleles.

2.3. Translation of SCA37 RAN pentapeptides in transfected HEK293T cells

To investigate if the *DAB1* pentanucleotide repeats were able to produce poly-pentapeptides *in vitro*, I performed dot blots and western blots using protein extracted from cells overexpressing pCDH-6xSTOP-$(ATTTT)_7$-3Tags, pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags

and pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors. I performed the dot blots and western blots in triplicated. In each replicate I used protein extracts obtained from a different batch of HEK293T cells transfections. To understand which reading frame(s) were subjected to pentanucleotide repeat translation, I used α-HA, α-flag and α-myc primary antibodies targeting the HA, flag and myc epitopes located at the C-terminus of the putative pentapeptide in each reading frame. Although the pentanucleotide repeats were predicted to produce a pentapeptide fused with an epitope (Appendix H) I did not detect RAN peptides by dot blot or western blot analysis in any frame of the pentanucleotide repeat in both SCA37 pathogenic or nonpathogenic alleles (Figure 19).

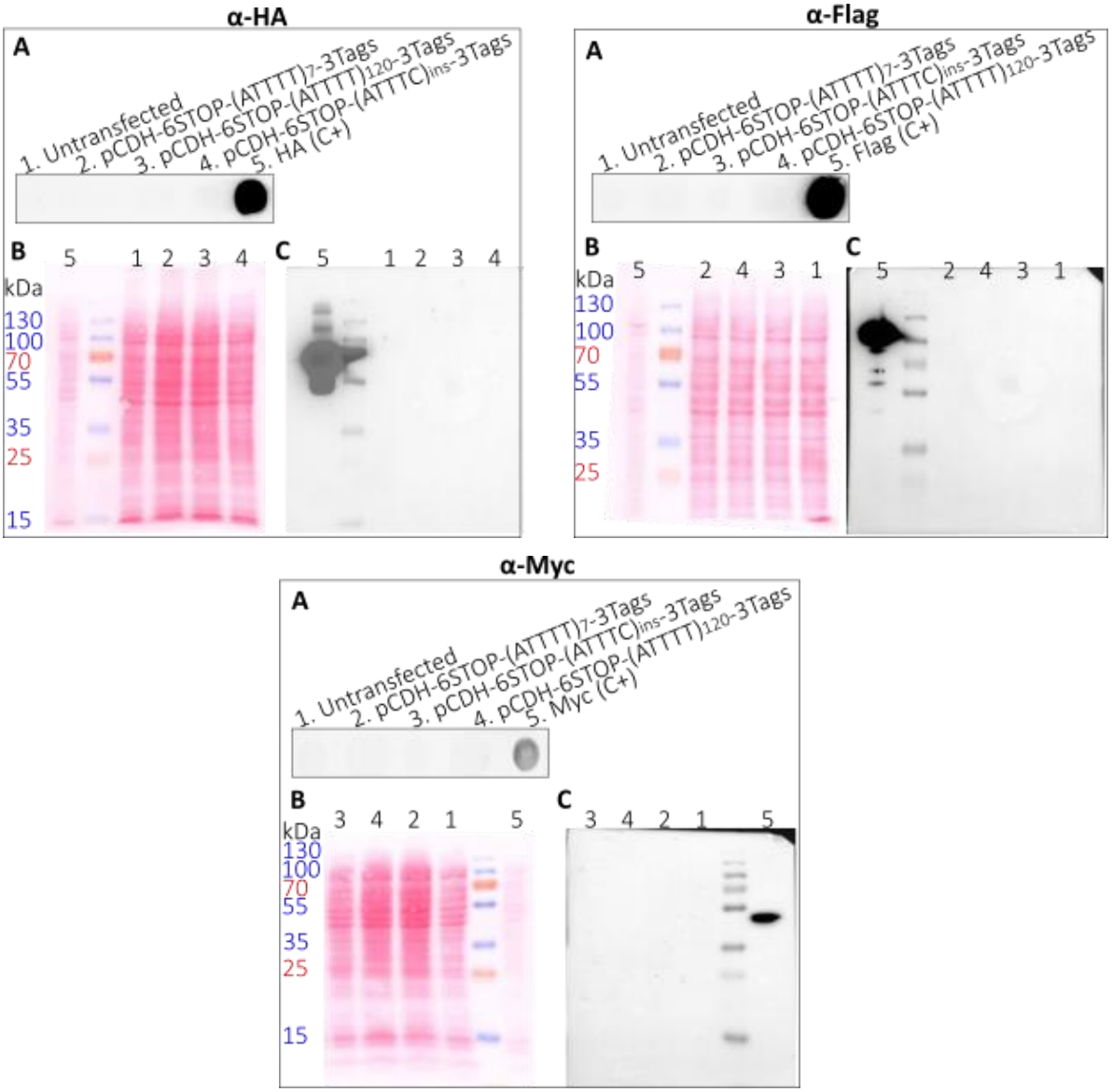


Figure 19. Representative dot blots (A)**, Ponceau stains** (B) **and western blots** (C) **performed in extracts of HEK293T cells transfected with pCDH-6xSTOP-Rep-3Tags vectors**. Dot blots and

western blots were performed with primary antibodies α-HA, α-flag and α-myc targeting the HA, flag and myc epitopes. Each dot blot and western blot was performed in triplicated using protein extracts from independent transfections of HEK293T cells. No RAN peptides were detected.

I also performed dot blots in protein extracts from HEK293T cells transfected with pCDH-6xSTOP-$(ATTTT)_7$-3Tags, pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags and pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors, using two primary antibodies raised against the putative RAN pentapeptide poly(ISFHF), encoded in the 3 frames of the ATTTC repeat insertion (α-RAN RB7543 and α-RAN RB7544). Both antibodies showed non-specific binding to protein extracts from HEK293T cells transfected with $(ATTTT)_7$ and $(ATTTT)_{120}$ alleles and to HEK293T cells untransfected (Figure 20).

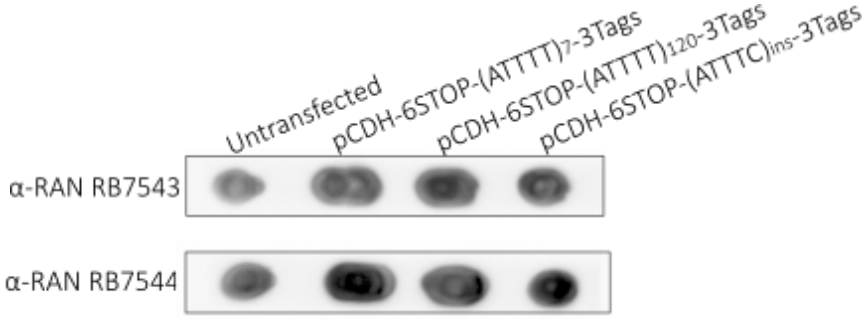


Figure 20. Dot blots performed in proteins extracted from HEK293T cells transfected with pCDH-6xSTOP-Rep-3Tags vectors, with primary antibodies α-RAN RB7543 and α-RAN RB7544 targeting the putative RAN pentapeptide poly(ISFHF).

Since I did not detect the poly(ISFHF) pentapeptide in extracts of HEK293T cells overexpressing pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags, I also performed dot blots in protein extracted from HEK293T cells transfected with pCDNA3-6xSTOP-$(ATTTT)_7$-3Tags, pCDNA3-6xSTOP-$(ATTTT)_{118}$-3Tags and pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags. Contrarily to the pCDH backbone, the pCDNA3 encodes a bgh-poly(A) signal at the end of the multiple cloning site. The efficient polyadenylation is required for the normal nucleocytoplasmic transport and translation of mRNA molecules. However, using pCDNA3 as backbone, I also did not detect RAN peptides translated from any frame of the mutant or normal pentanucleotide alleles using α-HA, α-flag and α-myc primary antibodies targeting the HA, flag and myc epitopes fused with the putative poly(ISFHF) (Figure 21). I performed the dot blots in triplicated using protein extracts derived from different batches of HEK293T cells transfections.

Figure 21. Representative dot blots performed in extracts of HEK293T cells transfected with pCDNA3-6xSTOP-Rep-3Tags vectors. Dot blots were performed with primary antibodies α-HA, α-flag and α-myc targeting the HA, flag and myc epitopes. Each dot blot was performed in triplicated using protein extracts from independent transfections of HEK293T cells. No RAN peptides were detected.

3. Investigation of RAN translation in SCA37 using a human Neural Stem Cell line (Cb192)

3.1. Transfection of hNSC with constructs expressing *DAB1* pentanucleotide repeats

The overexpression of *DAB1* pentanucleotide repeats in HEK293T cells did not allow to detect RAN poly-pentapeptides in SCA37. The cell line used to perform this investigation may not be permissive to the translation of these peptides. Since SCA37 is a disease leading to neuronal degeneration, I also investigated whether RAN translation was initiated in SCA37 locus using a cell line closer to the cells affected in the disease. I used a hNSC line derived from human neural tissue from a foetus with 50-55 days, named Cb192[166]. Cb192 cells were transfected with pCDH-6xSTOP-$(ATTTT)_7$-3Tags, pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags and pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors. At 72h post-transfection, most cells were expressing GFP (Figure 22), suggesting that the cells were expressing the *DAB1* pentanucleotide repeats fused with tags.

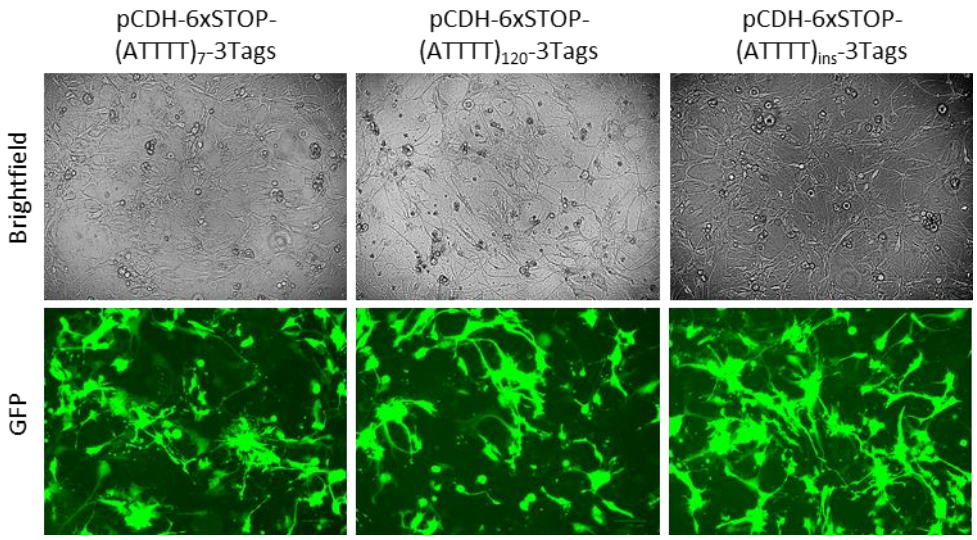


Figure 22. Transfection efficiency in hNSC. Brightfield and fluorescence images at 72h post-transfection of Cb192 cells with pCDH-6xSTOP-Rep-3Tags vectors.

3.2. Translation of SCA37 RAN pentapeptides in transfected hNSC

After transfecting hNSC with pCDH-6xSTOP-$(ATTTT)_7$-3Tags, pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags and pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags vectors, I performed dot blots using extracts from these transfected cells and primary antibodies α-HA, α-flag and α-myc targeting the epitopes fused with the predicted pentapeptides (Figure 23). The dot blot did not show signal corresponding to the detection of poly-pentapeptides, in any repeat reading frame for the mutant $(ATTTC)_{ins}$ or the normal $(ATTTT)_7$ or $(ATTTT)_{120}$ repeats.

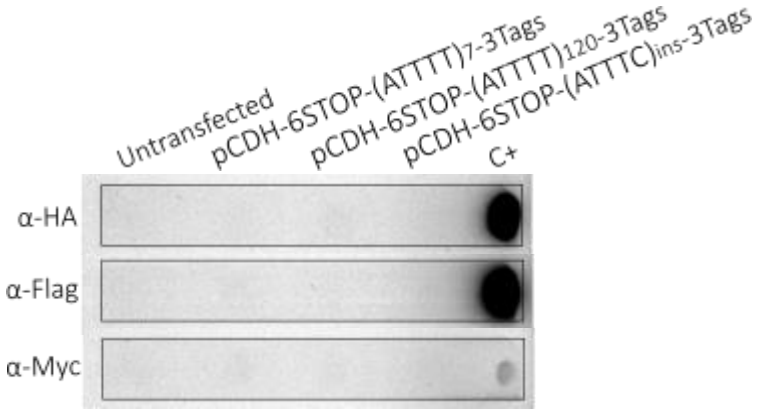


Figure 23. Dot blots performed in extracts of hNSC transfected with pCDH-6xSTOP-Rep-3Tags vectors. Dot blots were performed with primary antibodies α-HA, α-flag and α-myc targeting the HA, flag and myc epitopes. No RAN peptides were detected.

DISCUSSION AND FUTURE PERSPECTIVES

In 2017, the research group where I developed this master dissertation identified a noncoding $(ATTTC)_n$ insertion in an intronic region of *DAB1* as the genetic cause of SCA37[13]. The $(AUUUC)_n$ RNA forms abnormal nuclear aggregates in human cell lines and is toxic *in vivo,* leading to high lethality rates and developmental malformations when injected in zebrafish embryos[13]. Although our previous findings showed that the $(AUUUC)_n$ RNA initiates a cascade of RNA-mediated toxicity, the pathogenic mechanisms triggered by this toxic RNA remain elusive. The translation of the repeat expansion by repeat associated non-AUG dependent (RAN) translation is widely found in diseases caused by repeat expansions and insertions such as DM1, DM2, *C9ORF72* FTD/ALS, HD, Huntington disease-like 2 (HDL2), SCA2, SCA3, SCA8, SCA31 and SCA36[64, 65, 85, 87, 91, 97, 99, 106, 109, 113].

In this work I investigated, *in vitro*, the potential of the SCA37 $(ATTTC)_n$ insertion to be translated by RAN in a potentially toxic poly(ISFHF), encoded in the three reading frames of the $(ATTTC)_n$ insertion. To perform this work, I used a similar methodology previously used to find RAN translation of expanded repeats in DM1, DM2, *C9ORF72* FTD/ALS, HD, SCA2 and SCA8[64, 85, 87, 106, 109]. In these studies, researchers engineered constructs consisting in the repetitive pathogenic allele, 2 stop codons in each reading frame upstream the repeats and 3 C-terminal epitopes (HA, flag and myc) fused with the repeat expansion. This methodology allowed the successful detection of RAN peptides in HEK293T cells transfected with the generated constructs[64, 85, 87, 106, 109].

In SCA37, HEK293T cells transfected with pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags showed that the generated construct is able to drive $(AUUUC)_n$ RNA transcription and, as previously reported, this repetitive RNA forms nuclear foci[13]. However, I could not detect poly(ISFHF) RAN peptides translated from any repeat reading frame of the $(ATTTC)_n$, using this strategy. The absence of poly(ISFHF) translation in this *in vitro* system, may be explained by the fact that the pCDH backbone does not encode a poly(A) signal downstream the HA, flag and myc epitopes, which has a role in nucleocytoplasmic transport and efficient translation of mRNAs. To overcome this issue, I used a pCDNA3 vector, that contains a sequence encoding a poly(A) signal downstream the HA, flag and myc epitopes. In fact, the pCDNA3 and the derived pCDNA3.1 vectors were the backbones used in most studies to overexpress repeat expansions when RAN translation was investigated in other repeat diseases[65, 85, 87, 91, 106, 109, 113]. However, after transfecting HEK293T cells with pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags, I could not detect RAN peptides originated from any $(ATTTC)_n$ reading frame, as with the pCDH backbone vector. HEK293T cells are widely used in

fundamental research due to their easy culture, transfection and protein production, but they are derived from embryonic kidney cells and may not accurately mimic the cellular pathways in neurons. Therefore, I investigated the translation of poly-pentapeptides transfecting human derived neural stem cells (hNSC) Cb192[166], but I also did not detect SCA37 RAN peptides in these cells after transfecting them with pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags.

One possible explanation for the absence of RAN peptides in HEK293T cells overexpressing the SCA37 $(ATTTC)_n$ insertion is that the overexpressed RNA seems to be retained in nuclear aggregates when transfecting pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags or pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags. These large nuclear aggregates decrease the amount of free RNA that may escape to the cytoplasm and be subjected to RAN translation. The formation of RNA foci in brain cells of SCA37 affected individuals has not been investigated yet, but based on what is known for other repeat expansion diseases as FAME1, $(AUUUC)_n$ RNA most likely aggregates in nuclear RNA foci, which are usually much smaller than the RNA aggregates found in HEK293T cells overexpressing the $(ATTTC)_n$ insertion[17]. Therefore, it is possible that similarly to other diseases caused by intronic repeat expansions such as *C9ORF72* FTD/ALS, more $(AUUUC)_n$ RNA is able to escape to the cytoplasm and be translated in SCA37 affected individuals, compared to transfected cells[167]. However, RNA FISH needs to be performed in Cb192 transfected cell lines to understand if in this cell line the $(AUUUC)_n$ RNA is also prone to form large RNA aggregates or if the repetitive RNA may escape to the cytoplasm and be available to RAN translation.

Another explanation for the absence of detection of RAN pentapeptides *in vitro* may be the fact that SCA37 is a late-onset disease and RAN peptides may need decades of accumulation in cerebellar neurons to become detectable. In this work, I analysed the production of RAN polypeptides from the $(ATTTC)_n$ in proliferative cell lines, 72h post-transfection. This might be a short period for RAN peptides to accumulate if the translation in these cells is not efficient.

The detection of RAN translation, *in vitro*, may be also influenced by the region upstream the repeat cloned in the construct. During the investigation of RAN translation in SCA3/MJD, Jazurek-Ciesiolka and colleagues detected RAN translation when cloning the $(CAG)_{exp}$ together with the 66 bp upstream the CAG repeat, however, when truncating part of this flanking region, the authors detected that the RAN translation of the polyQ frame was abolished[113]. In this work, I cloned the $(ATTTC)_n$ flanking region 66 bp upstream the repeat, but other regions upstream may also be required for the RAN translation in SCA37.

Interestingly, it would not be the first time that RAN peptides are present in the brain of affected individuals but cannot be detected *in vitro*. Mori and colleagues when investigating RAN translation in *C9ORF72* FTD/ALS using antibodies α-$(GA)_{15}$, α-$(GR)_6$ and α-$(GP)_{15}$, against the $(GGGGCC)_{exp}$ associated RAN peptides, detected two RAN peptides, poly(GP) and poly(GA), but not poly(GR) in HEK293T cells overexpressing 75-145 GGGGCC repeats[99]. However, the poly(GR) is also found in brain tissue of affected individuals using the same α-(GR) antibody[99]. Gendron and colleagues, focusing on RAN dipeptides translated from the antisense $(CCCCGG)_{exp}$ in *C9ORF72* FTD/ALS, did not show poly(PA) translation in HEK293T cells transfected with a vector containing 66 CCCCGG repeats, contrarily to what is found in brain tissue of affected individuals[64]. In both studies the authors did not determine the repeat size of the pathogenic allele in the brain tissue of affected individuals, but perhaps the RAN translation was not detected in HEK293T because the size of the repeat expansion was too small compared to what is usually found in the brain of affected individuals[64, 99]. In this work, the size of the $(ATTTC)_n$ insertion used in transfection assays had 57 ATTTCs in pCDH and 55 ATTTCs in pCDNA3. Perhaps RAN translation may not be detected in these $(ATTTC)_n$ but it may be detected using larger repeat insertions, since in SCA37 the $(ATTTC)_n$ insertion ranges between 31 and 75 repeat units and there is an inverse correlation between the repeat size and the age of disease onset with larger repeats showing worse phenotypes[13].

Other methodologies should also be considered in the future to investigate RAN translation in SCA37. Although antibodies against epitopes as HA, myc and flag are more readily available and much more specific than antibodies against particular polypeptides, these antibodies do not allow to detect RAN polypeptides in cells and tissues derived from affected individuals. For instance, in this work, I also used polyclonal antibodies raised against the putative poly(ISFHF), however these antibodies were non-specific since they bound to other proteins extracted from untransfected cells and cells transfected with pCDH-6xSTOP-$(ATTTT)_7$-3Tags and pCDH-6xSTOP-$(ATTTT)_{120}$-3Tags. Therefore, these antibodies are not suitable for the detection of RAN peptides in cells and tissues derived from SCA37 affected individuals and new antibodies should be generated for this purpose. A specific antibody against poly(ISFHF) would allow to detect RAN pentapeptides in SCA37 in induced pluripotent stem cell-derived neurons (iPSN) and brain tissue derived from SCA37 affected individuals.

In this work, I did not find pentapeptides translated from the $(ATTTC)_n$ insertion. Nevertheless, understanding if RAN translation is a pathogenic mechanism contributing to SCA37 and other diseases caused by $(ATTTC)_n$ insertion remains an important step for the

development of effective therapeutic strategies to treat these diseases and provide a better quality of life to affected individuals.

REFERENCES

1. Strachan T and Read A. Human Molecular Genetics. 4th ed. 2010.
2. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001.
3. Konkel MK, Walker JA, Hotard AB, et al. Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project. Genome Biology and Evolution. 2015.
4. Banez-Coronel M and Ranum LPW. Repeat-associated non-AUG (RAN) translation: insights from pathology. Laboratory Investigation. 2019.
5. Tautz D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research. 1989.
6. Zhao J, Bacolla A, Wang G, et al. Non-B DNA structure-induced genetic instability and evolution. Cellular and Molecular Life Sciences. 2010.
7. Gadgil R, Barthelemy J, Lewis T, et al. Replication stalling and DNA microsatellite instability. Biophysical Chemistry. 2017.
8. Bagshaw A. Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. Genome Biology and Evolution. 2017.
9. Krüger J and Schleinitz D. Genetic Fingerprinting Using Microsatellite Markers in a Multiplex PCR Reaction: A Compilation of Methodological Approaches from Primer Design to Detection Systems. Methods in Molecular Biology. 2017.
10. Moxon ER and Wills C. DNA microsatellites: agents of evolution? Scientific American. 1999.
11. Depienne C and Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? American Journal of Human Genetics. 2021.
12. Sato N, Amino T, Kobayashi K, et al. Spinocerebellar Ataxia Type 31 Is Associated with "Inserted" Penta-Nucleotide Repeats Containing (TGGAA)n. American Journal of Human Genetics. 2009.
13. Seixas AI, Loureiro JR, Costa C, et al. A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. American Journal of Human Genetics. 2017.
14. Corbett MA, Kroes T, Veneziano L, et al. Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. Nature Communications. 2019.
15. Florian RT, Kraft F, Leitão E, et al. Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. Nature Communications. 2019.
16. Yeetong P, Pongpanich M, Srichomthong C, et al. TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4. Brain. 2019.
17. Ishiura H, Doi K, Mitsui J, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nature Genetics. 2018.
18. Cen Z, Jiang Z, Chen Y, et al. Intronic pentanucleotide TTTCA repeat insertion in the SAMD12 gene causes familial cortical myoclonic tremor with epilepsy type 1. Brain. 2018.
19. Cen Z, Chen Y, Yang D, et al. Intronic (TTTGA) n insertion in SAMD12 also causes familial cortical myoclonic tremor with epilepsy. Movement Disorders. 2019.
20. Fu YH, Pizzuti A, Fenwick RG, et al. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science. 1992.
21. Brook JD, McCurrach ME, Harley HG, et al. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell. 1992.

22. López Castel A, Nakamori M, Tomé S, et al. Expanded CTG repeat demarcates a boundary for abnormal CpG methylation in myotonic dystrophy patient tissues. Human Molecular Genetics. 2011.
23. Wong LJ, Ashizawa T, Monckton DG, et al. Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. American Journal of Human Genetics. 1995.
24. Redman JB, Fenwick RG, Fu YH, et al. Relationship between parental trinucleotide GCT repeat length and severity of myotonic dystrophy in offspring. The Journal of the American Medical Association. 1993.
25. Zoghbi HY and Orr HT. Spinocerebellar ataxia type 1. Seminars in Cell Biology. 1995.
26. Loureiro JR, Oliveira CL and Silveira I. Unstable repeat expansions in neurodegenerative diseases: Nucleocytoplasmic transport emerges on the scene. Neurobiology of Aging. 2016.
27. Lee D, Lee Y-I, Lee Y-S, et al. The Mechanisms of Nuclear Proteotoxicity in Polyglutamine Spinocerebellar Ataxias. Frontiers in Neuroscience. 2020.
28. Katsuno M, Watanabe H, Yamamoto M, et al. Potential therapeutic targets in polyglutamine-mediated diseases. Expert Review of Neurotherapeutics. 2014.
29. Chung CG, Lee H and Lee SB. Mechanisms of protein toxicity in neurodegenerative diseases. Cellular and Molecular Life Sciences. 2018.
30. Nucifora FC, Sasaki M, Peters MF, et al. Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity. Science. 2001.
31. McCampbell A, Taylor JP, Taye AA, et al. CREB-binding protein sequestration by expanded polyglutamine. Human Molecular Genetics. 2000.
32. Li F, Macfarlan T, Pittman RN, et al. Ataxin-3 is a histone-binding protein with two independent transcriptional corepressor activities. The Journal of Biological Chemistry. 2002.
33. Shoubridge C and Gecz J. Polyalanine tract disorders and neurocognitive phenotypes. Advances in Experimental Medicine and Biology. 2012.
34. Brais B, Bouchard JP, Xie YG, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nature Genetics. 1998.
35. Messaed C and Rouleau GA. Molecular mechanisms underlying polyalanine diseases. Neurobiology of Disease. 2009.
36. Hughes JN and Thomas PQ. Molecular Pathology of Polyalanine Expansion Disorders: New Perspectives from Mouse Models. Methods in Molecular Biology. 2013.
37. Nelson DL, Orr HT and Warren ST. The unstable repeats-Three evolving faces of neurological disease. Neuron. 2013.
38. Rodriguez CM and Todd PK. New pathologic mechanisms in nucleotide repeat expansion disorders. Neurobiology of Disease. 2019.
39. Oberlé I, Rousseau F, Heitz D, et al. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science. 1991.
40. Verkerk AJMH, Pieretti M, Sutcliffe JS, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991.
41. Kumari D, Gazy I and Usdin K. Pharmacological Reactivation of the Silenced FMR1 Gene as a Targeted Therapeutic Approach for Fragile X Syndrome. Brain Sciences. 2019.
42. Serrano M. Epigenetic cerebellar diseases. Handbook Of Clinical Neurology. 2018.
43. Campuzano V, Montermini L, Moltò MD, et al. Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science. 1996.
44. Al-Mahdawi S, Pinto RM, Ismail O, et al. The Friedreich ataxia GAA repeat expansion mutation induces comparable epigenetic changes in human and transgenic mouse brain and heart tissues. Human Molecular Genetics. 2008.
45. Delatycki MB, Paris DB, Gardner RJ, et al. Clinical and genetic study of Friedreich ataxia in an Australian population. American Journal of Medical Genetics. 1999.

46. Delatycki MB and Bidichandani SI. Friedreich ataxia- pathogenesis and implications for therapies. Neurobiology of Disease. 2019.
47. Greene E, Mahishi L, Entezam A, et al. Repeat-induced epigenetic changes in intron 1 of the frataxin gene and its consequences in Friedreich ataxia. Nucleic Acids Research. 2007.
48. Wells RD. DNA triplexes and Friedreich ataxia. The FASEB Journal. 2008.
49. Kim E, Napierala M and Dent SYR. Hyperexpansion of GAA repeats affects post-initiation steps of FXN transcription in Friedreich's ataxia. Nucleic Acids Research. 2011.
50. Groh M, Lufino MMP, Wade-Martins R, et al. R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. PLoS Genetics. 2014.
51. Swinnen B, Robberecht W and Van Den Bosch L. RNA toxicity in non-coding repeat expansion disorders. The EMBO Journal. 2020.
52. Ozimski LL, Sabater-Arcis M, Bargiela A, et al. The hallmarks of myotonic dystrophy type 1 muscle dysfunction. Biological reviews of the Cambridge Philosophical Society. 2021.
53. Fardaei M, Rogers MT, Thorpe HM, et al. Three proteins, MBNL, MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. Human Molecular Genetics. 2002.
54. Tian B, White RJ, Xia T, et al. Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. RNA. 2000.
55. Ho TH, Charlet-B N, Poulos MG, et al. Muscleblind proteins regulate alternative splicing. The EMBO Journal. 2004.
56. Renoux AJ and Todd PK. Neurodegeneration the RNA way. Progress in Neurobiology. 2012.
57. Echeverria GV and Cooper TA. RNA-binding proteins in microsatellite expansion disorders: Mediators of RNA toxicity. Brain Research. 2012.
58. Wang ET, Ward AJ, Cherone JM, et al. Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. Genome Research. 2015.
59. López-Martínez A, Soblechero-Martín P, de-la-Puente-Ovejero L, et al. An Overview of Alternative Splicing Defects Implicated in Myotonic Dystrophy Type I. Genes. 2020.
60. Botta A, Vallo L, Rinaldi F, et al. Gene expression analysis in myotonic dystrophy: Indications for a common molecular pathogenic pathway in DM1 and DM2. Gene Expression. 2007.
61. Zhang K, Donnelly CJ, Haeusler AR, et al. The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. Nature. 2015.
62. Castro AF, Loureiro JR, Bessa J, et al. Antisense Transcription across Nucleotide Repeat Expansions in Neurodegenerative and Neuromuscular Diseases: Progress and Mysteries. Genes. 2020.
63. DH C, CP T, SE M, et al. Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. Molecular Cell. 2005.
64. Gendron TF, Bieniek KF, Zhang YJ, et al. Antisense transcripts of the expanded C9ORF72 hexanucleotide repeat form nuclear RNA foci and undergo repeat-associated non-ATG translation in c9FTD/ALS. Acta Neuropathologica. 2013.
65. Zu T, Liu Y, Bañez-Coronel M, et al. RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. Proceedings of the National Academy of Sciences of the United States of America. 2013
66. Krans A, Kearse MG and Todd PK. Repeat-associated non-AUG translation from antisense CCG repeats in fragile X tremor/ataxia syndrome. Annals of Neurology. 2016.
67. Li PP, Sun X, Xia G, et al. ATXN2-AS, a gene antisense to ATXN2, is associated with spinocerebellar ataxia type 2 and amyotrophic lateral sclerosis. Annals of neurology. 2016.
68. De Mezer M, Wojciechowska M, Napierala M, et al. Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. Nucleic Acids Research. 2011.

69. Sun X, Li PP, Zhu S, et al. Nuclear retention of full-length HTT RNA is mediated by splicing factors MBNL1 and U2AF65. Scientific Reports. 2015.
70. Khristich AN and Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. The Journal of Biological Chemistry. 2020.
71. Tsoi H, Lau TCK, Tsang SY, et al. CAG expansion induces nucleolar stress in polyglutamine diseases. Proceedings of the National Academy of Sciences of the United States of America. 2012.
72. Urbanek MO, Jazurek M, Switonski PM, et al. Nuclear speckles are detention centers for transcripts containing expanded CAG repeats. Biochimica et Biophysica Acta - Molecular Basis of Disease. 2016.
73. Li LB, Yu Z, Teng X, et al. RNA toxicity is a component of ataxin-3 degeneration in Drosophila. Nature. 2008.
74. Cleary JD, Pattamatta A and Ranum LPW. Repeat-associated non-ATG (RAN) translation. The Journal of Biological Chemistry. 2018.
75. Alberts B. Molecular Biology of the Cell. 5th ed. 2008.
76. Ramanathan A, Robb GB and Chan SH. mRNA capping: biological functions and applications. Nucleic Acids Research. 2016.
77. Wan R, Bai R and Shi Y. Molecular choreography of pre-mRNA splicing by the spliceosome. Current Opinion in Structural Biology. 2019.
78. Sun Y, Hamilton K and Tong L. Recent molecular insights into canonical pre-mRNA 3'-end processing. Transcription. 2020.
79. Vicens Q, Kieft JS and Rissland OS. Revisiting the Closed-Loop Model and the Nature of mRNA 5'-3' Communication. Molecular Cell. 2018.
80. Castelli LM, Huang W-P, Lin Y-H, et al. Mechanisms of repeat-associated non-AUG translation in neurological microsatellite expansion disorders. Biochemical Society Transactions. 2021.
81. Kressler D, Hurt E and Baßler J. A Puzzle of Life: Crafting Ribosomal Subunits. Trends in Biochemical Sciences. 2017.
82. Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Research. 1987.
83. Pisarev AV, Kolupaeva VG, Pisareva VP, et al. Specific functional interactions of nucleotides at key-3 and+4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. Genes and Development. 2006.
84. Dhar AK, Robles-Sikisaka R, Saksmerprome V, et al. Biology, Genome Organization, and Evolution of Parvoviruses in Marine Shrimp. Advances in Virus Research. 2014.
85. Zu T, Gibbens B, Doty NS, et al. Non-ATG-initiated translation directed by microsatellite expansions. Proceedings of the National Academy of Sciences of the United States of America. 2011.
86. Liquori CL, Ricker K, Moseley ML, et al. Myotonic dystrophy type 2 caused by a CCTG expansion in intron I of ZNF9. Science. 2001.
87. Zu T, Cleary JD, Liu Y, et al. RAN Translation Regulated by Muscleblind Proteins in Myotonic Dystrophy Type 2. Neuron. 2017.
88. Tusi SK, Nguyen L, Thangaraju K, et al. The alternative initiation factor eIF2A plays key role in RAN translation of myotonic dystrophy type 2 CCUG•CAGG repeats. Human Molecular Genetics. 2021.
89. Wieben ED, Aleff RA, Tosakulwong N, et al. A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. PLoS ONE. 2012.
90. Mootha VV, Gong X, Ku HC, et al. Association and familial segregation of CTG18.1 trinucleotide repeat expansion of TCF4 gene in fuchs' endothelial corneal dystrophy. Investigative Ophthalmology and Visual Science. 2014.

91. Soragni E, Petrosyan L, Rinkoski TA, et al. Repeat-associated non-ATG (RAN) translation in fuchs' endothelial corneal dystrophy. Investigative Ophthalmology and Visual Science. 2018.
92. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. Neuron. 2011.
93. Renton AE, Majounie E, Waite A, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011.
94. Ash PEA, Bieniek KF, Gendron TF, et al. Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. Neuron. 2013.
95. Wen X, Tan W, Westergard T, et al. Antisense proline-arginine RAN dipeptides linked to C9ORF72-ALS/FTD form toxic nuclear aggregates that initiate invitro and invivo neuronal death. Neuron. 2014.
96. Swaminathan A, Bouffard M, Liao M, et al. Expression of C9orf72-related dipeptides impairs motor function in a vertebrate model. Human Molecular Genetics. 2018.
97. Todd TW, McEachin ZT, Chew J, et al. Hexanucleotide Repeat Expansions in c9FTD/ALS and SCA36 Confer Selective Patterns of Neurodegeneration In Vivo. Cell Reports. 2020.
98. McEachin ZT, Gendron TF, Raj N, et al. Chimeric Peptide Species Contribute to Divergent Dipeptide Repeat Pathology in c9ALS/FTD and SCA36. Neuron. 2020.
99. Mori K, Weng SM, Arzberger T, et al. The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. Science. 2013.
100. Almeida S, Gascon E, Tran H, et al. Modeling key pathological features of frontotemporal dementia with C9ORF72 repeat expansion in iPSC-derived human neurons. Acta Neuropathologica. 2013.
101. Batra R and Lee CW. Mouse Models of C9orf72 Hexanucleotide Repeat Expansion in Amyotrophic Lateral Sclerosis/ Frontotemporal Dementia. Frontiers In Cellular Neuroscience. 2017.
102. Shaw MP, Higginbottom A, McGown A, et al. Stable transgenic C9orf72 zebrafish model key aspects of the ALS/FTD phenotype and reveal novel pathological features. Acta Neuropathologica Communications. 2018.
103. Todd PK, Oh SY, Krans A, et al. CGG repeat-associated translation mediates neurodegeneration in fragile X tremor ataxia syndrome. Neuron. 2013.
104. Buijsen RAM, Visser JA, Kramer P, et al. Presence of inclusions positive for polyglycine containing protein, FMRpolyG, indicates that repeat-associated non-AUG translation plays a role in fragile X-associated primary ovarian insufficiency. Human Reproduction. 2016.
105. MacDonald ME, Ambrose CM, Duyao MP, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993.
106. Bañez-Coronel M, Ayhan F, Tarabochia AD, et al. RAN Translation in Huntington Disease. Neuron. 2015.
107. Margolis RL, O'Hearn E, Rosenblatt A, et al. A disorder similar to Huntington's disease is associated with a novel CAG repeat expansion. Annals of Neurology. 2001.
108. Pulst SM, Nechiporuk A, Nechiporuk T, et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinooerebellar ataxia type. Nature Genetics. 1996.
109. Scoles DR, Ho MHT, Dansithong W, et al. Repeat Associated non-AUG translation (RAN Translation) dependent on sequence downstream of the ATXN2 CAG repeat. PLoS ONE. 2015.
110. Kawaguchi Y, Okamoto T, Taniwaki M, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nature Genetics. 1994.
111. Schöls L, Vieira-saecker AMM, Schöls S, et al. Trinucleotide expansion within the MJD1 gene presents clinically as spinocerebellar ataxia and occurs most frequently in german SCA patients. Human Molecular Genetics. 1995.
112. Stochmanski SJ, Therrien M, Laganière J, et al. Expanded ATXN3 frameshifting events are toxic in Drosophila and mammalian neuron models. Human Molecular Genetics. 2012.

113. Jazurek-Ciesiolka M, Ciesiolka A, Komur AA, et al. RAN translation of the expanded CAG repeats in the SCA3 disease context. Journal of Molecular Biology. 2020.
114. Koob MD, Moseley ML, Schut LJ, et al. An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nature Genetics. 1999.
115. Ayhan F, Perez BA, Shorrock HK, et al. SCA 8 RAN polySer protein preferentially accumulates in white matter regions and is regulated by eIF 3F. The EMBO Journal. 2018.
116. Ishiguro T, Sato N, Ueyama M, et al. Regulatory Role of RNA Chaperone TDP-43 for RNA Misfolding and Repeat-Associated Translation in SCA31. Neuron. 2017.
117. Zu T, Pattamatta A and Ranum LPW. Repeat-associated non-ATG translation in neurological diseases. Cold Spring Harbor Perspectives in Biology. 2018.
118. Kozak M. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proceedings of the National Academy of Sciences of the United States of America. 1990.
119. Pelletier J and Sonenberg N. The Organizing Principles of Eukaryotic Ribosome Recruitment. Annual Review of Biochemistry. 2019.
120. Nguyen L, Cleary JD and Ranum LPWW. Repeat-Associated Non-ATG Translation: Molecular Mechanisms and Contribution to Neurological Disease. Annual Review of Neuroscience. 2019.
121. Malik I, Kelley CP, Wang ET, et al. Molecular mechanisms underlying nucleotide repeat expansion disorders. Nature Reviews Molecular Cell Biology. 2021.
122. Kearse MG, Green KM, Krans A, et al. CGG Repeat-Associated Non-AUG Translation Utilizes a Cap-Dependent Scanning Mechanism of Initiation to Produce Toxic Proteins. Molecular Cell. 2016.
123. Cheng W, Wang S, Mestre AA, et al. C9ORF72 GGGGCC repeat-associated non-AUG translation is upregulated by stress through eIF2α phosphorylation. Nature Communications. 2018.
124. Stoneley M, Paulin FE, Le Quesne JP, et al. C-Myc 5' untranslated region contains an internal ribosome entry segment. Oncogene. 1998.
125. Ray PS, Grover R and Das S. Two internal ribosome entry sites mediate the translation of p53 isoforms. EMBO Reports. 2006.
126. Johnson AG, Grosely R, Petrov AN, et al. Dynamics of IRES-mediated translation. Philosophical Transactions of the Royal Society of London Series B: Biological sciences. 2017.
127. Landry DM, Hertz MI and Thompson SR. RPS25 is essential for translation initiation by the Dicistroviridae and hepatitis C viral IRESs. Genes and Development. 2009.
128. Yamada SB, Gendron TF, Niccoli T, et al. RPS25 is required for efficient RAN translation of C9orf72 and other neurodegenerative disease-associated nucleotide repeats. Nature Neuroscience. 2019.
129. Hertz MI, Landry DM, Willis AE, et al. Ribosomal protein S25 dependency reveals a common mechanism for diverse internal ribosome entry sites and ribosome shunting. Molecular and Cellular Biology. 2013.
130. Mori K, Arzberger T, Grässer FA, et al. Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. Acta Neuropathologica. 2013.
131. Balendra R and Isaacs AM. C9orf72-mediated ALS and FTD: multiple pathways to disease. Nature Reviews Neurology. 2018.
132. Tao Z, Wang H, Xia Q, et al. Nucleolar stress and impaired stress granule formation contribute to C9orf72 RAN translation-induced cytotoxicity. Human Molecular Genetics. 2015.
133. Mizielinska S, Grönke S, Niccoli T, et al. C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. Science. 2014.
134. Shi KY, Mori E, Nizami ZF, et al. Toxic PR n poly-dipeptides encoded by the C9orf72 repeat expansion block nuclear import and export. Proceedings of the National Academy of Sciences of the United States of America. 2017.

135. Kwon I, Xiang S, Kato M, et al. Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. Science. 2014.
136. Lin Y, Mori E, Kato M, et al. Toxic PR Poly-Dipeptides Encoded by the C9orf72 Repeat Expansion Target LC Domain Polymers. Cell. 2016.
137. Lee KH, Zhang P, Kim HJ, et al. C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. Cell. 2016.
138. Lopez-Gonzalez R, Lu Y, Gendron TF, et al. Poly(GR) in C9ORF72-Related ALS/FTD Compromises Mitochondrial Function and Increases Oxidative Stress and DNA Damage in iPSC-Derived Motor Neurons. Neuron. 2016.
139. Zhang YJ, Jansen-West K, Xu YF, et al. Aggregation-prone c9FTD/ALS poly(GA) RAN-translated proteins cause neurotoxicity by inducing ER stress. Acta Neuropathologica. 2014.
140. May S, Hornburg D, Schludi MH, et al. C9orf72 FTLD/ALS-associated Gly-Ala dipeptide repeat proteins cause neuronal toxicity and Unc119 sequestration. Acta Neuropathologica. 2014.
141. Tabet R, Schaeffer L, Freyermuth F, et al. CUG initiation and frameshifting enable production of dipeptide repeat proteins from ALS/FTD C9ORF72 transcripts. Nature Communications. 2018.
142. Sellier C, Buijsen RAM, He F, et al. Translation of Expanded CGG Repeats into FMRpolyG Is Pathogenic and May Contribute to Fragile X Tremor Ataxia Syndrome. Neuron. 2017.
143. Oh SY, He F, Krans A, et al. RAN translation at CGG repeats induces ubiquitin proteasome system impairment in models of fragile X-associated tremor ataxia syndrome. Human Molecular Genetics. 2015.
144. Sullivan R, Yau WY, O'Connor E, et al. Spinocerebellar ataxia: an update. Journal of Neurology. 2019.
145. Carlson KM, Andresen JM and Orr HT. Emerging pathogenic pathways in the spinocerebellar ataxias. Current Opinion in Genetics and Development. 2009.
146. Scott SSO, Pedroso JL, Barsottini OGP, et al. Natural history and epidemiology of the spinocerebellar ataxias: Insights from the first description to nowadays. Journal of the Neurological Sciences. 2020.
147. Afonso-Reis R, Afonso IT and Nóbrega C. Current Status of Gene Therapy Research in Polyglutamine Spinocerebellar Ataxias. International Journal of Molecular Sciences. 2021.
148. Müller U. Spinocerebellar ataxias (SCAs) caused by common mutations. Neurogenetics. 2021.
149. Ruano L, Melo C, Silva MC, et al. The global epidemiology of hereditary ataxia and spastic paraplegia: A systematic review of prevalence studies. Neuroepidemiology. 2014.
150. Vale J, Bugalho P, Silveira I, et al. Autosomal dominant cerebellar ataxia: frequency analysis and clinical characterization of 45 families from Portugal. European Journal of Neurology. 2010.
151. Coutinho P, Ruano L, Loureiro JL, et al. Hereditary ataxia and spastic paraplegia in Portugal: a population-based prevalence study. JAMA Neurology. 2013.
152. González-Zaldívar Y, Vázquez-Mojena Y, Laffita-Mesa JM, et al. Epidemiological, clinical, and molecular characterization of Cuban families with spinocerebellar ataxia type 3/Machado-Joseph disease. Cerebellum and Ataxias. 2015.
153. Bargiela D, Yu-Wai-Man P, Keogh M, et al. Prevalence of neurogenetic disorders in the North of England. Neurology. 2015.
154. Paradisi I, Ikonomu V and Arias S. Spinocerebellar ataxias in Venezuela: Genetic epidemiology and their most likely ethnic descent. Journal of Human Genetics. 2016.
155. Serrano-Munuera C, Corral-Juan M, Stevanin G, et al. New subtype of spinocerebellar ataxia with altered vertical eye movements mapping to chromosome 1p32. JAMA Neurology. 2013.
156. Corral-Juan M, Serrano-Munuera C, Rábano A, et al. Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. Brain. 2018.

157. Gao Z and Godbout R. Reelin-Disabled-1 signaling in neuronal migration: splicing takes the stage. Cellular and Molecular Life Sciences. 2013.
158. Howell BW, Herrick TM and Cooper JA. Reelin-induced tyrosine phosphorylation of disabled 1 during neuronal positioning. Genes and Development. 1999.
159. Bosch C, Masachs N, Exposito-Alonso D, et al. Reelin Regulates the Maturation of Dendritic Spines, Synaptogenesis and Glial Ensheathment of Newborn Granule Cells. Cerebral Cortex. 2016.
160. Trotter J, Lee GH, Kazdoba TM, et al. Dab1 is required for synaptic plasticity and associative learning. The Journal of Neuroscience. 2013.
161. Johnson FH and Lewin I. The growth rate of E. coli in relation to temperature, quinine and coenzyme. Journal of Cellular and Comparative Physiology. 1946.
162. Ferenc M. Plasmids 101: Common Lab E. coli Strains. Addgene Blog. 2014. https://blog.addgene.org/plasmids-101-common-lab-e-coli-strains/
163. Sanger F, Nicklen S and Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977.
164. Rosmaninho P, Mükusch S, Piscopo V, et al. Zeb1 potentiates genome-wide gene transcription with Lef1 to promote glioblastoma cell invasion. The EMBO Journal. 2018.
165. Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. Nature Methods. 2012.
166. Sun Y, Pollard S, Conti L, et al. Long-term tripotent differentiation capacity of human neural stem (NS) cells in adherent culture. Molecular and Cellular Neurosciences. 2008.
167. Freibaum BD and Taylor JP. The role of dipeptide repeats in C9ORF72-related ALS-FTD. Frontiers in Molecular Neuroscience. 2017.

APPENDIXES

Appendix A. Vector maps

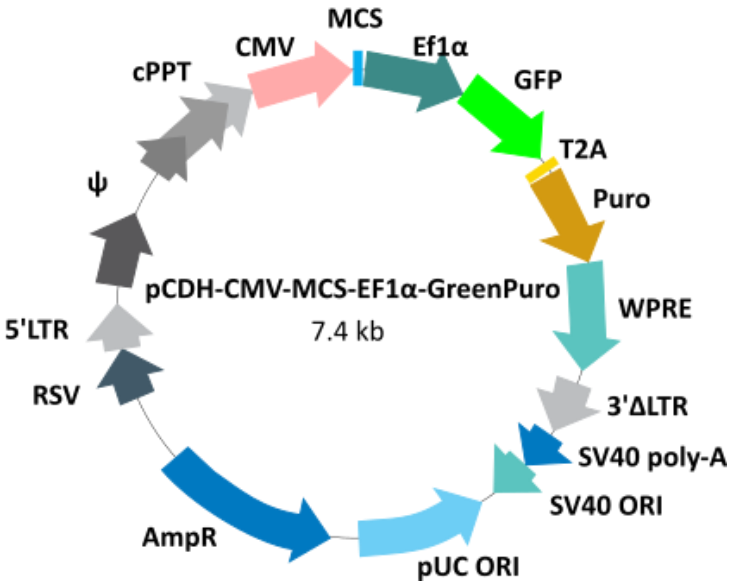


Figure A1. pCDH-CMV-MCS-EF1α-GreenPuro (System Biosciences).

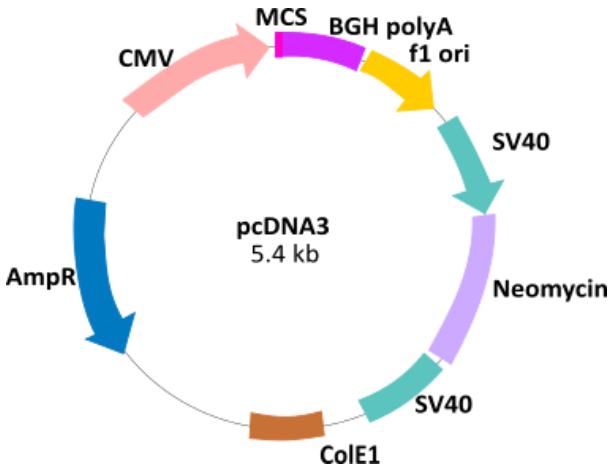


Figure A2. pCDNA3 (Invitrogen).

Appendix B. Oligonucleotides for cloning

Oligonucleotide ID	Oligonucleotide sequence (5'-3')
STR24_GXL3_RAN_F	GCGCGCGGAATTCCCAAGTCAGCCTCCCAGGTAAC
STR24_GXL3_RAN_R	ATTAAGCGGCCGCTGGGCAACACAGTGAGACCCTGTC
6xSTOP_RAN_pCDH_F	CTAGTAGATAGTAGATAGTAG
6xSTOP_RAN_pCDH_R	AATTCTACTATCTACTATCTA
6xSTOP_RAN_pCDNA3_F	AGCTGTAGTAGATAGTAGATAGTAG
6xSTOP_RAN_pCDNA3_R	AATTCTACTATCTACTATCTACTAC
Tags_RAN_pCDH_F	GGCCGTACCCATACGATGTTCCAGATTACGCTGGATTACAAGG ATGACGACGATAAGAGAACAGAAACTGATCTCTGAAGAAGACC TGTAAGAT
Tags_RAN_pCDH_R	CTTACAGGTCTTCTTCAGAGATCAGTTTCTGTCCTCTTATCGTC GTCATCCTTGTAATCCAGCGTAATCTGGAACATCGTATGGGTAC

Appendix C. Primers for PCR amplification and Sanger sequencing

Primer ID	Primer sequence (5'-3')	Used to sequence
STR24_RAN_F	TCCCAAGTCAGCCTCCCAGGTAAC	$(ATTTT)_n$
STR24_RAN_R	GACAGGGTCTCACTGTGTTGCCCAG	$(ATTTT)_n$
Tags_RAN_pCDNA3_F	TATAAGCGGCCGCTACCCATACGATGTTCCA	3Tags in pCDNA3
Tags_RAN_pCDNA3_R	CGGGCGTCTAGATCTTACAGGTCTTCTTCAGAG	3Tags in pCDNA3
CMV_F	CAAATGGGCGGTAGGCGTG	Stop codon cassette
Ef1a_R	TCTCTAGGCACCCGTTCAAT	3Tags in pCDH
SP6_F	ATTTAGGTGACACTATAG	3Tags in pCDNA3

Figure A3. Schematic representation of the annealing sites of the primers used for PCR amplification and Sanger sequencing. The black lines represent the backbones of the pCDH and pCDNA3 vectors.

Appendix D. Protein quantification

1. Preparation of albumin (BSA) standards

   c (stock BSA) = 50 mg/mL = 50000 µg/mL

   c (BSA standards): 1000 µg/mL, 700 µg/mL, 600 µg/mL, 500 µg/mL, 400 µg/mL, 300 µg/mL, 200 µg/mL, 100 µg/mL, 0 µg/mL (blank).

2. Preparation of BCA working reagent (WR)

   a. Use the following formula to determine the total volume of WR required:
      (# standards + # samples) × (# replicates) × (volume of WR per sample) = total volume WR required
      Note: 200 µL of the WR is required for each sample in the microplate procedure. I did 2 replicas of each standard and sample.
   b. Prepare WR by mixing 50 parts of BCA Reagent A with 1 part of BCA Reagent B.

3. Microplate procedure

   a. Pipette 10 µL of each standard and each sample in a 96-well plate.
   b. Add 200 µL of WR to each well.
   c. Put the plate in a shaker at 75 rpm, for 30 seconds.
   d. Incubate at 37 °C for 30 minutes, without agitation.
   e. Measure the absorbance at 562 nm.
   f. Calculate the average absorbance for each standard and sample.
   g. Subtract the blank average from the average of each standard and sample.
   h. Create a calibration curve with absorbance of standards vs concentration of standards, according to formula (2).

$$Absorbance\ (nm) = m * Concentration\ (\mu g/mL) + b \qquad (2)$$

   i. Determine the slope (m) and y-intercept (b) of the standard curve and calculate the concentration of the protein extracts of interest with formula (3).

$$Concentration\ (\mu g/mL) = \frac{Absorbance\ (nm) - b}{m} \qquad (3)$$

Appendix E. Antibodies

Primary antibody	Dilution for dot blot	Dilution for western blot
Monoclonal mouse α-HA-Tag (F-7) (sc-7392, Santa Cruz Biotechnology)	1:1000	1:1000
Monoclonal mouse α-c-myc (11667149001, Roche)	1:1000	1:1000
Monoclonal mouse α-FLAG M2 (F1804-200UG, Sigma-Aldrich)	1:1000	1:1000
Polyclonal rabbit α-RAN RB7543 (0.30 mg/mL, Biomatik)	1:200	-
Polyclonal rabbit α-RAN RB7544 (0.37 mg/mL, Biomatik)	1:200	-

Secondary antibody	Dilution for dot blot	Dilution for western blot
Mouse-IgGκ BP-HRP (sc-516102, Santa Cruz Biotechnology)	1:10 000	1:10 000
Polyclonal goat α-Rabbit IgG, H & L Chain Specific Peroxidase Conjugate (401393-2ML, EMD Millipore – Calbiochem)	1:10 000	-

Appendix F. Western blot

1. Preparation of polyacrylamide gels

a. Running gel - 12% polyacrylamide; 1.5 mm thick:

Reagent	Volume
H2O type I	4,4 mL
Tris 1.5M pH8.8	2,5 mL
40% Acrylamide/Bis Solution 29:1 (Bio-Rad)	3,0 mL
SDS 20%	50 µL
APS 25%*	40 µL
TEMED*	10 µL

b. Stacking gel - 4% polyacrylamide; 1.5 mm thick:

Reagent	Volume
H2O type I	1,6 mL
Tris 1.0M pH6.8	312,5 µL
40% Acrylamide/Bis Solution 29:1 (Bio-Rad)	250 µL
SDS 20%	12,5 µL
APS 25%*	10 µL
TEMED*	5 µL

* APS is a polymerization initiator and TEMED is a polymerization catalyser. Add these reagents in the fume hood, immediately before pipetting the mix into the casting.

2. Preparation of protein extracts to be resolved by SDS-PAGE

a. Prepare mix in 1.5 mL tubes, on ice:

Reagent	Quantity
Protein extract	60 µg
Laemmli 6x	1/6 of the protein extract volume
H_2O type I	Up to the highest volume of protein extract + Laemmli. Every well must be loaded with equal sample volumes.

Appendix G. Sequences of the generated constructs

pCDH-6xSTOP-$(ATTTT)_7$-3Tags

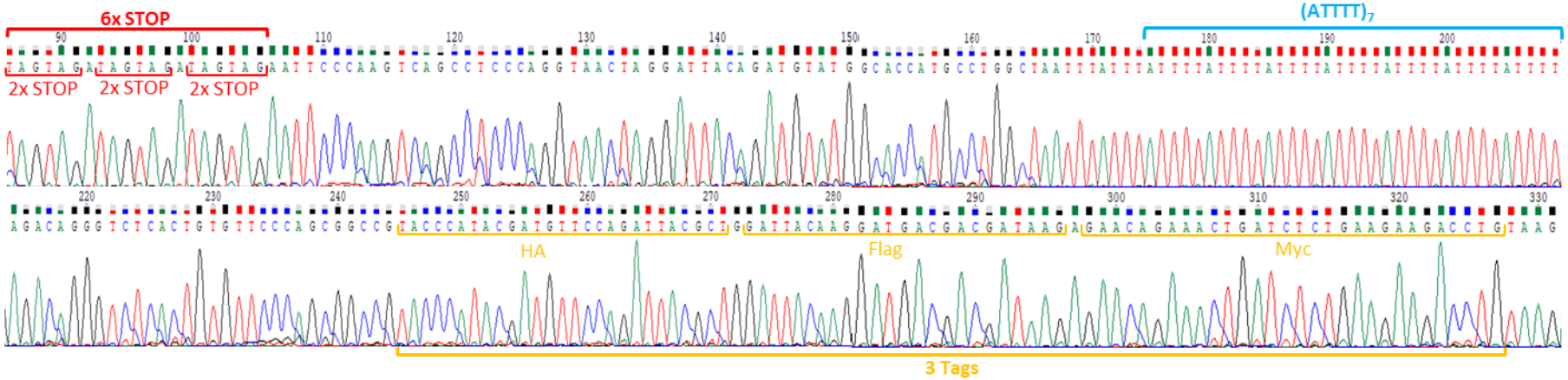


Figure A4. Sequencing of pCDH-6xSTOP-$(ATTTT)_7$-3Tags using primer CMV_F. The DNA sequence encoding 6 stop codons, two in each reading frame, is underlined in red. The DNA sequence encoding the short nonpathogenic allele is underlined in blue. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow.

pCDNA3-6xSTOP-(ATTTT)$_7$-3Tags

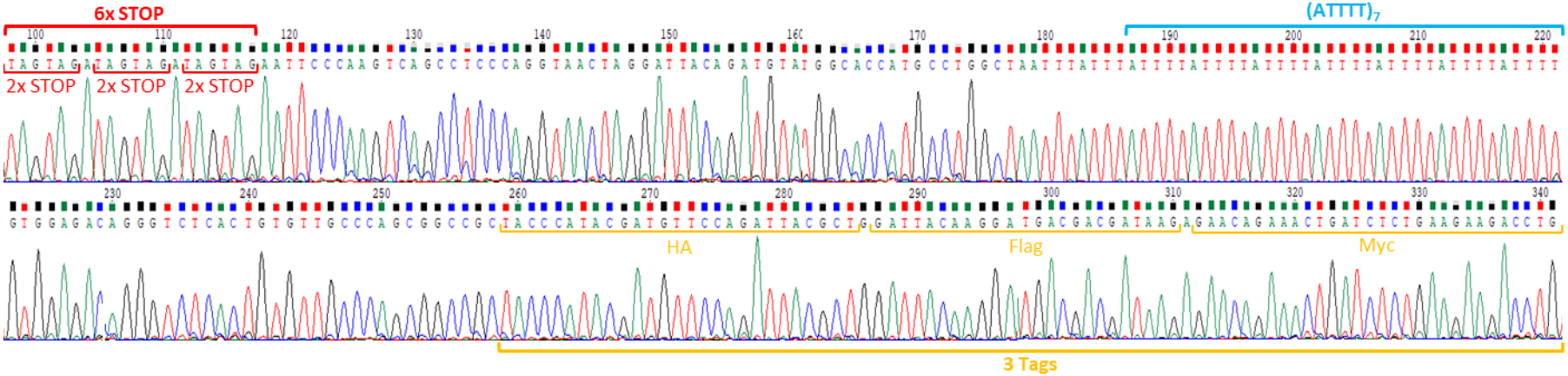


Figure A5. Sequencing of pCDNA3-6xSTOP-(ATTTT)$_7$-3Tags using primer CMV_F. The DNA sequence encoding 6 stop codons, two in each reading frame, is underlined in red. The DNA sequence encoding the short nonpathogenic allele is underlined in blue. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow.

pCDH-6xSTOP-(ATTTT)$_{120}$-3Tags

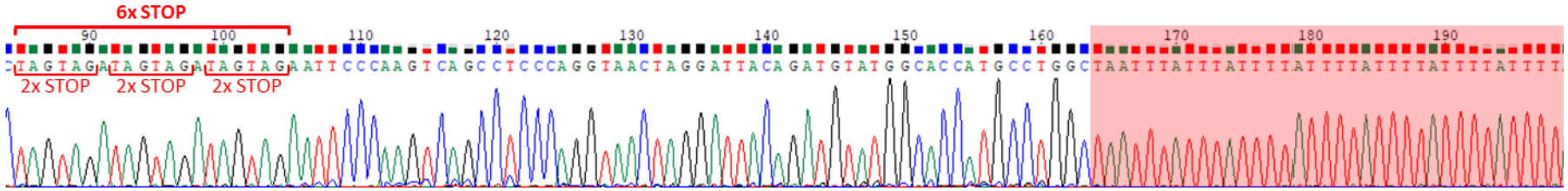


Figure A6. Sequencing of pCDH-6xSTOP-(ATTTT)$_{120}$-3Tags using primer CMV_F. The DNA sequence encoding 6 stop codons, two in each reading frame, is underlined in red. The region where the electropherograms in figures A6 and A7 overlap is highlighted in red.

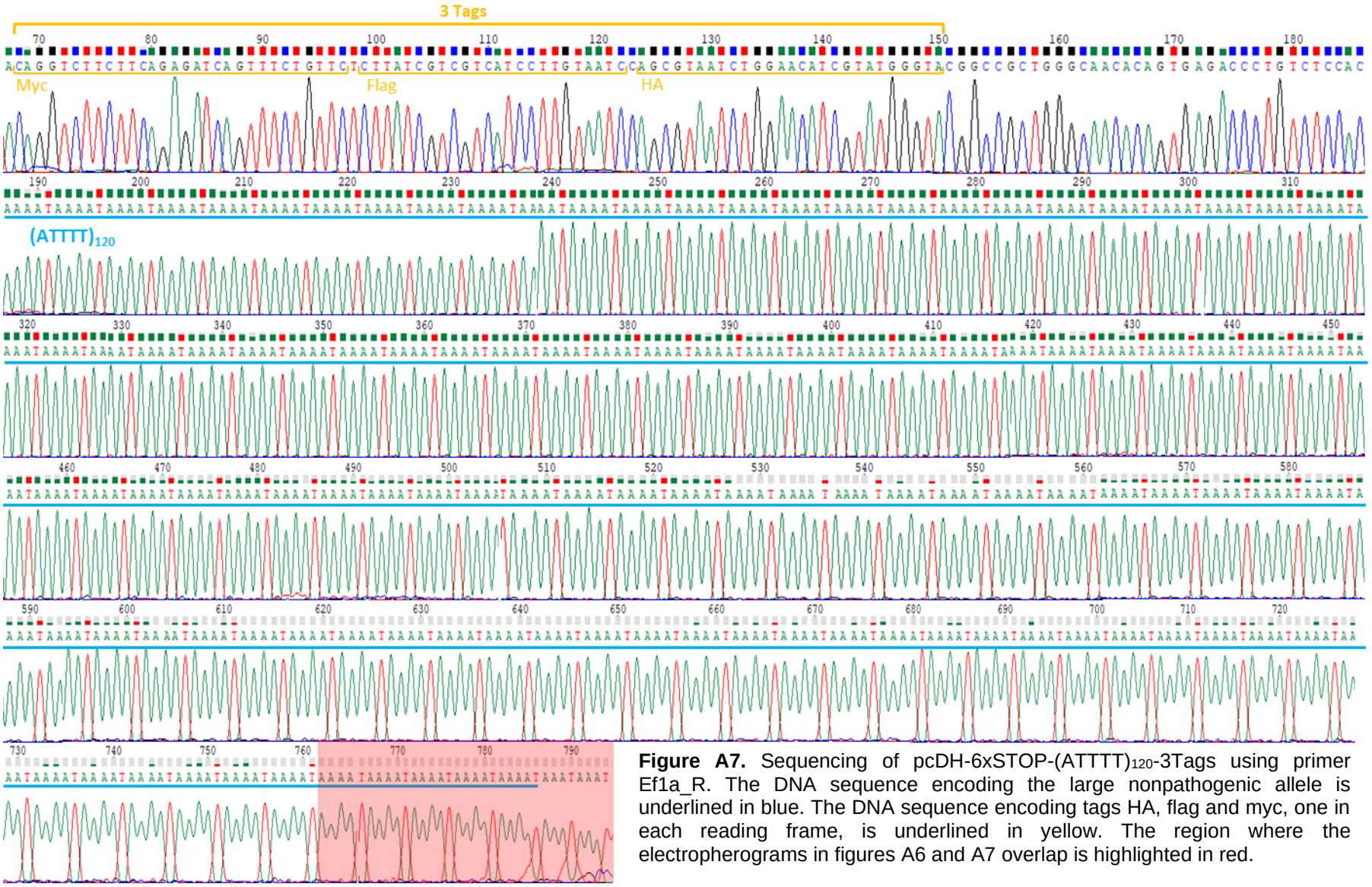


Figure A7. Sequencing of pcDH-6xSTOP-$(ATTTT)_{120}$-3Tags using primer Ef1a_R. The DNA sequence encoding the large nonpathogenic allele is underlined in blue. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow. The region where the electropherograms in figures A6 and A7 overlap is highlighted in red.

pCDNA3-6xSTOP-(ATTTT)$_{118}$-3Tags

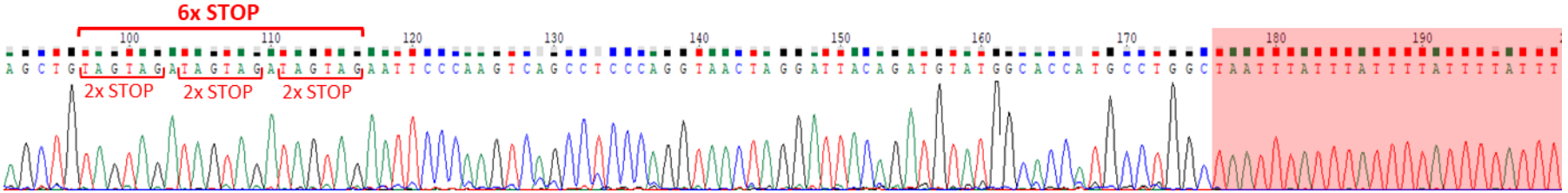


Figure A8. Sequencing of pCDNA3-6xSTOP-(ATTTT)$_{118}$-3Tags using primer CMV_F. The DNA sequence encoding 6 stop codons, two in each reading frame, is underlined in red. The region where the electropherograms in figures A8 and A10 overlap is highlighted in red.

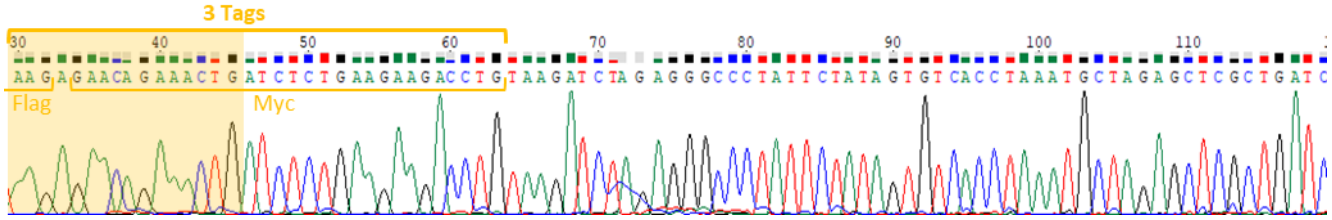


Figure A9. Sequencing of pCDNA3-6xSTOP-(ATTTT)$_{118}$-3Tags using primer Tags_RAN_pCDNA3_F. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow. The region where the electropherograms in figures A9 and A10 overlap is highlighted in yellow.

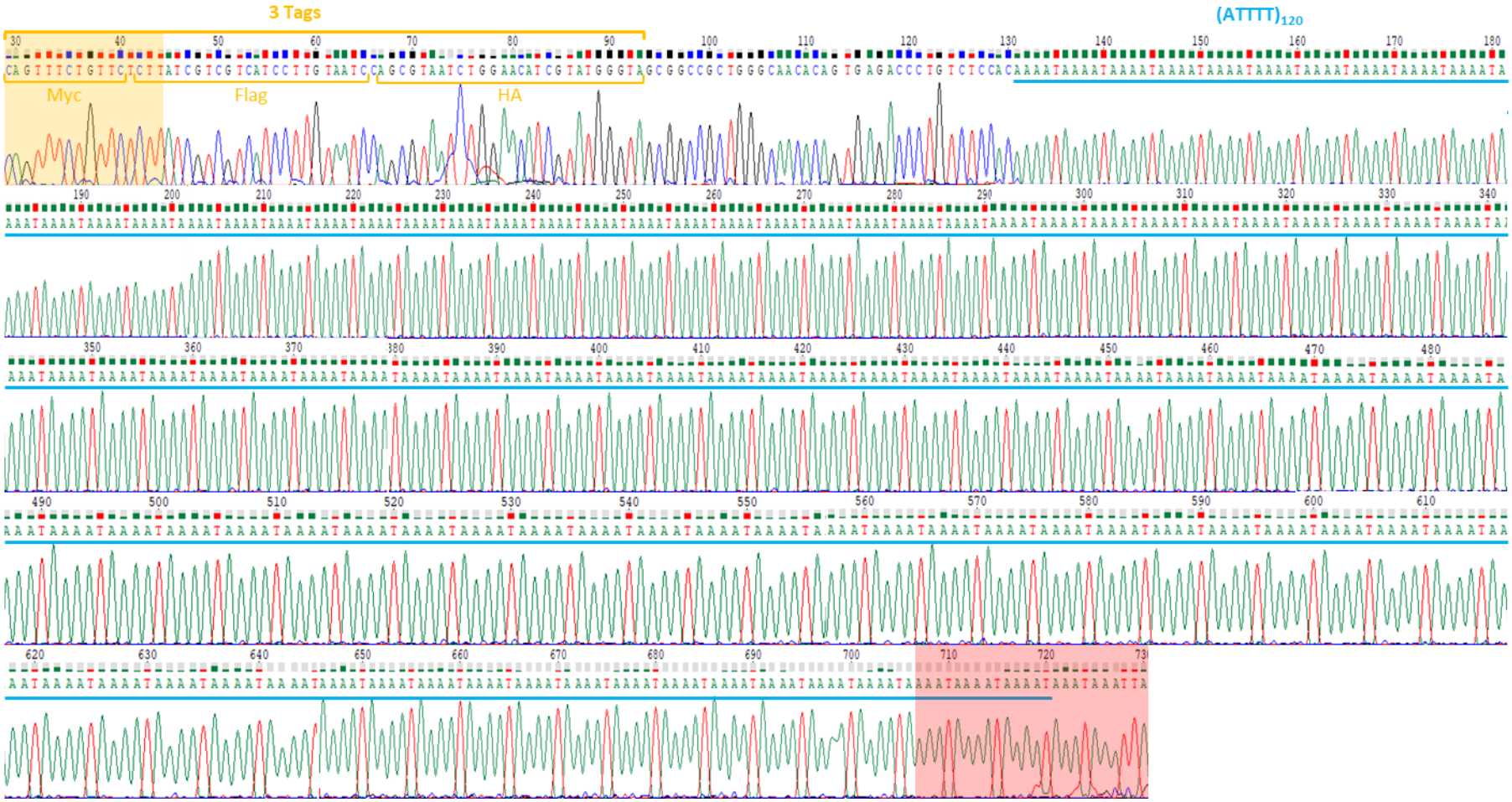


Figure A10. Sequencing of pCDNA3-6xSTOP-$(ATTTT)_{118}$-3Tags using primer SP6_F. The DNA sequence encoding the large nonpathogenic allele is underlined in blue. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow. The region where the electropherograms in figures A8 and A10 overlap is highlighted in red; the region where the electropherograms in figures A9 and A10 overlap is highlighted in yellow.

pCDH-6xSTOP-(ATTTT)$_{57}$(ATTTC)$_{57}$(ATTTT)$_{82}$-3Tags

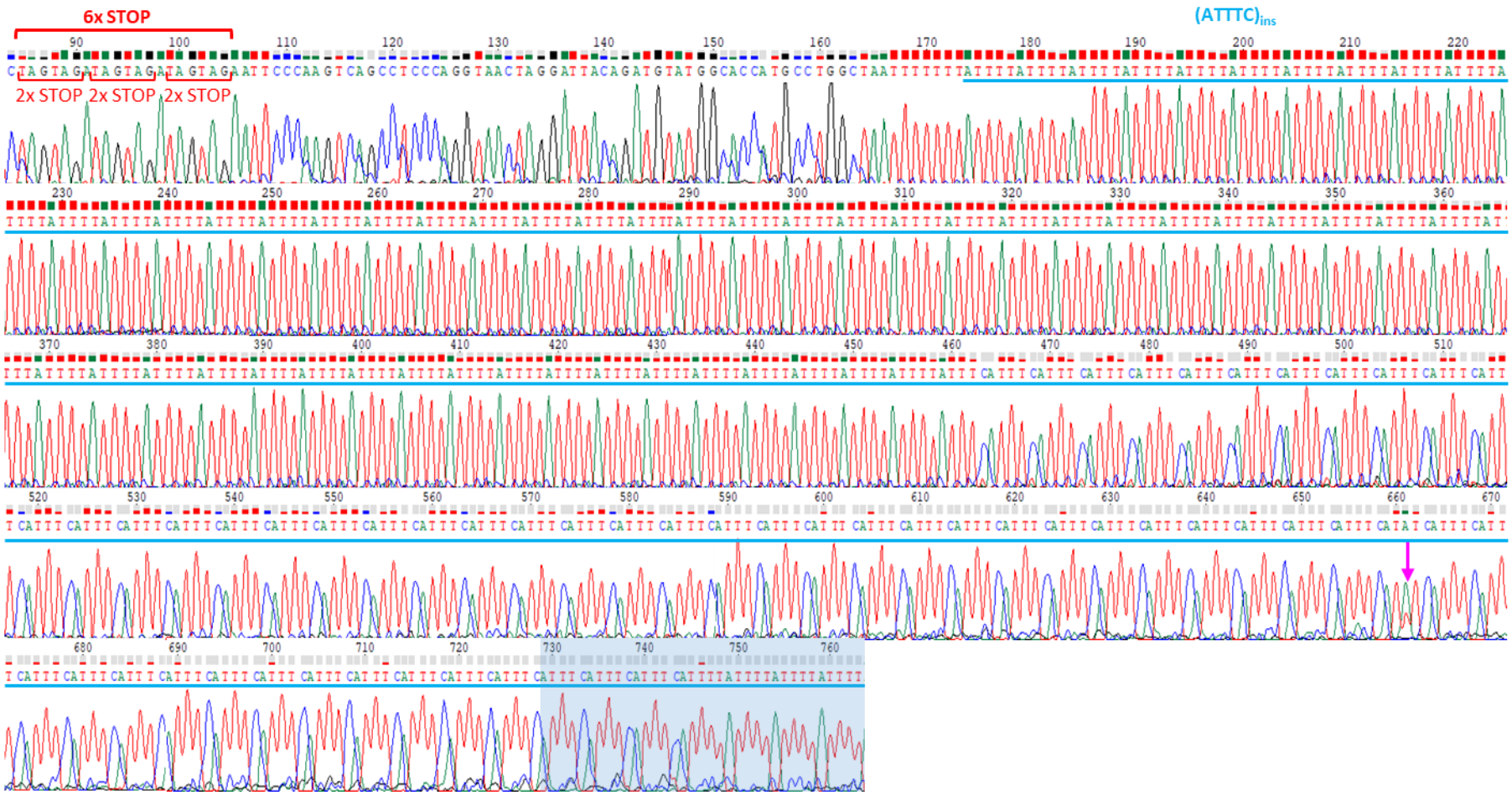


Figure A11. Sequencing of pCDH-6xSTOP-(ATTTC)$_{ins}$-3Tags using primer CMV_F. The DNA sequence encoding 6 stop codons, two in each reading frame, is underlined in red. The DNA sequence encoding the SCA37 pathogenic allele is underlined in blue. The region where the electropherograms in figures A11 and A12 overlap is highlighted in blue. Point mutations are signalled with a pink arrow.

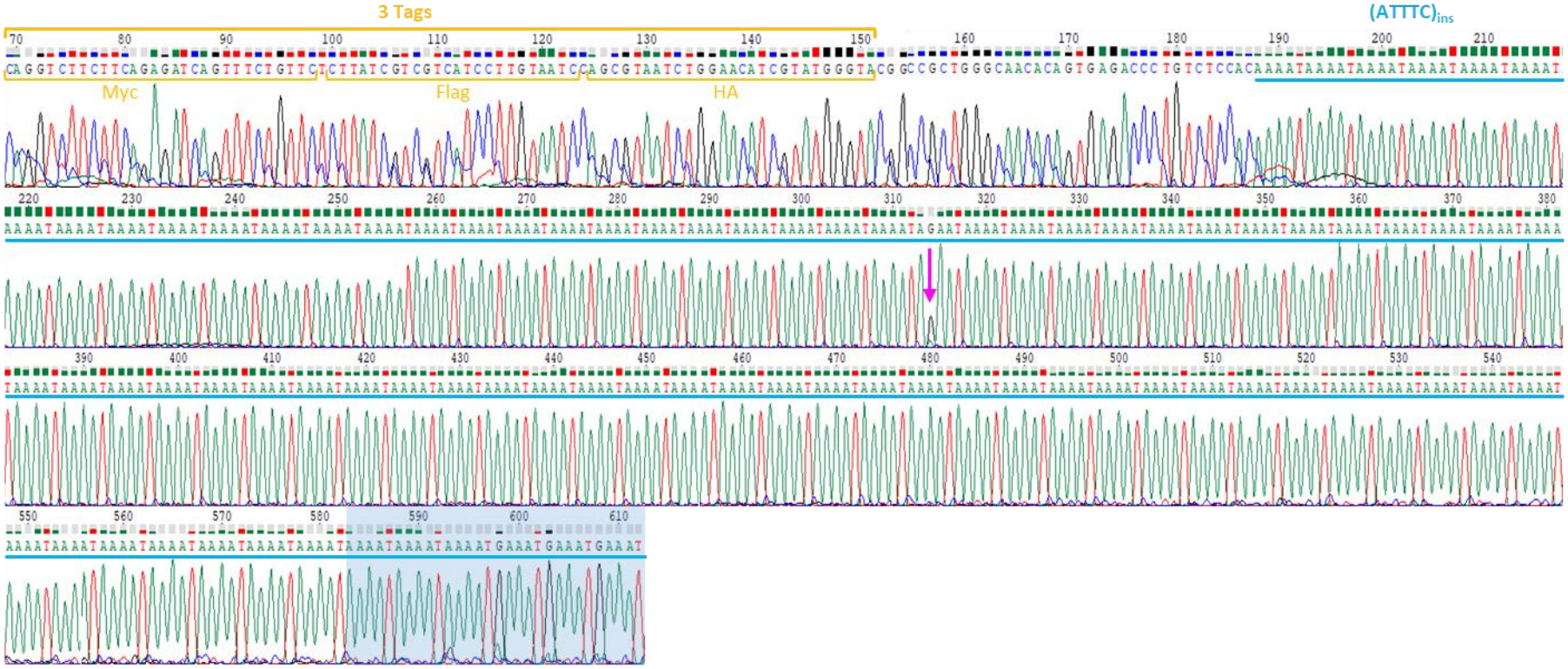


Figure A12. Sequencing of pCDH-6xSTOP-$(ATTTC)_{ins}$-3Tags using primer Ef1a_R. The DNA sequence encoding the SCA37 pathogenic allele is underlined in blue. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow. The region where the electropherograms in figures A11 and A12 overlap is highlighted in blue. Point mutations are signalled with a pink arrow.

pCDNA3-6xSTOP-$(ATTTT)_{58}(ATTTC)_{55}(ATTTT)_{79}$-3Tags

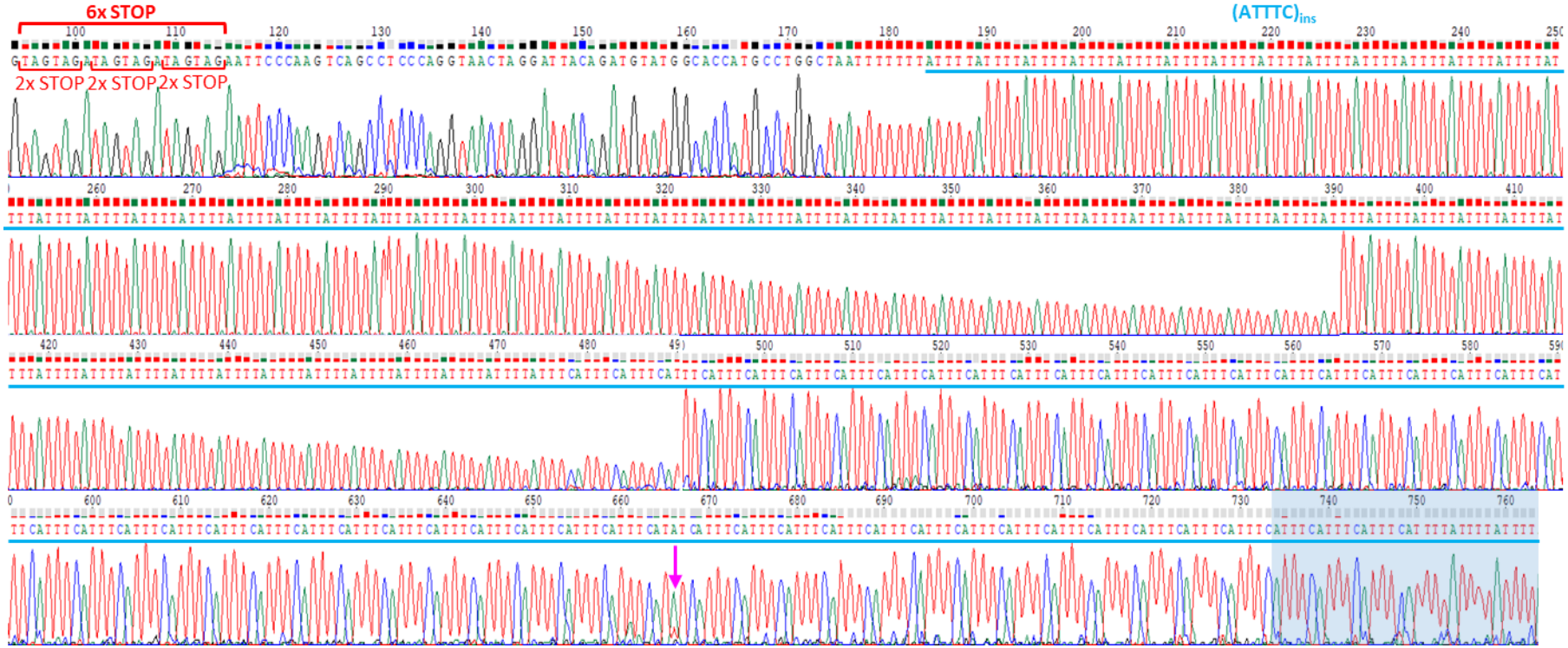


Figure A13. Sequencing of pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags using primer CMV_F. The DNA sequence encoding 6 stop codons, two in each reading frame, is underlined in red. The DNA sequence encoding the SCA37 pathogenic allele is underlined in blue. The region where the electropherograms in figures A13 and A14 overlap is highlighted in blue. Point mutations are signalled with a pink arrow.

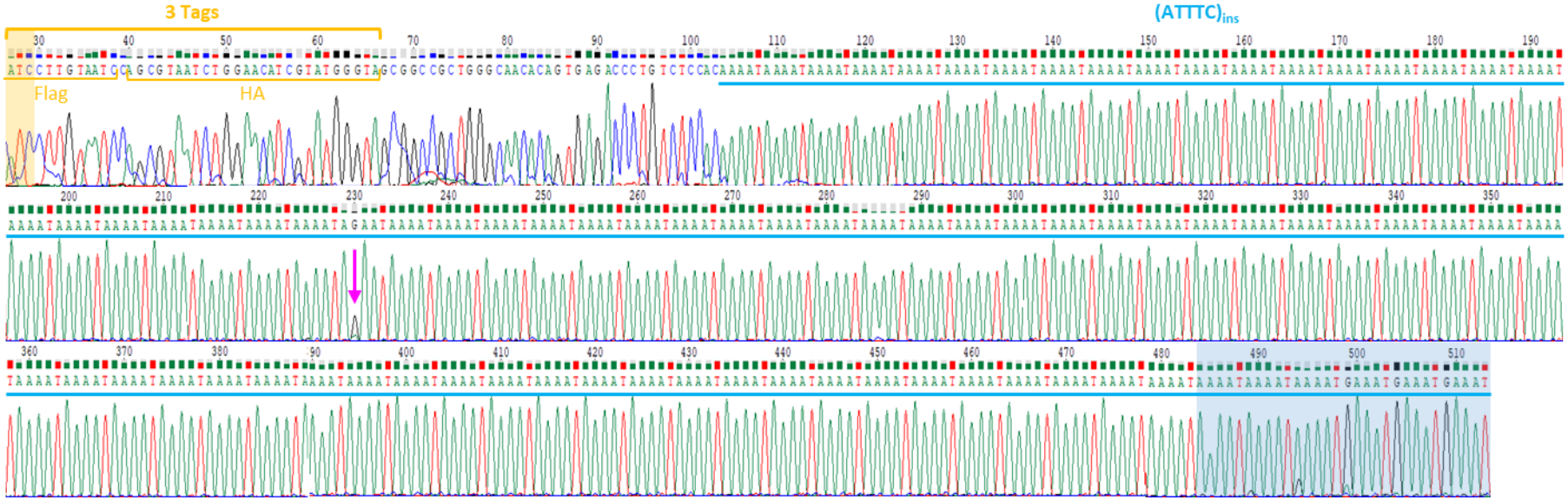


Figure A14. Sequencing of pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags using primer Tags_RAN_pCDNA3_R. The DNA sequence encoding the SCA37 pathogenic allele is underlined in blue. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow. The region where the electropherograms in figures A13 and A14 overlap is highlighted in blue; the region where the electropherograms in figures A14 and A15 overlap is highlighted in yellow. Point mutations are signalled with a pink arrow.

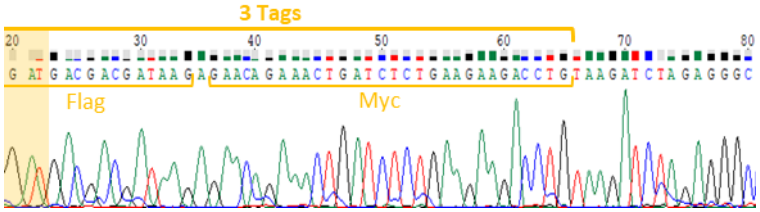
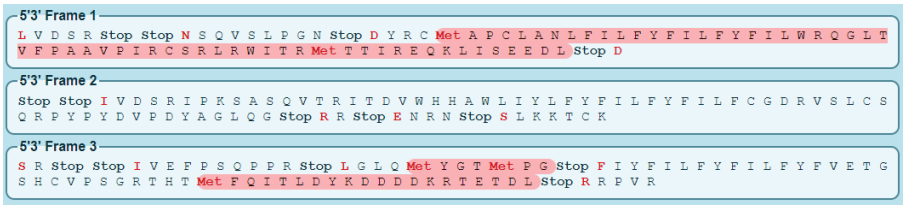


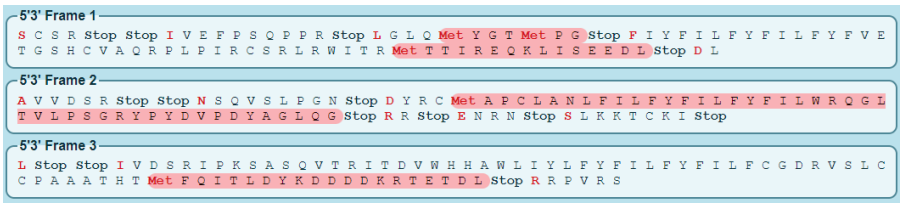
Figure A15. Sequencing of pCDNA3-6xSTOP-$(ATTTC)_{ins}$-3Tags using primer Tags_RAN_pCDNA3_F. The DNA sequence encoding tags HA, flag and myc, one in each reading frame, is underlined in yellow. The region where the electropherograms in figures A14 and A15 overlap is highlighted in yellow.

Appendix H. Predicted peptides

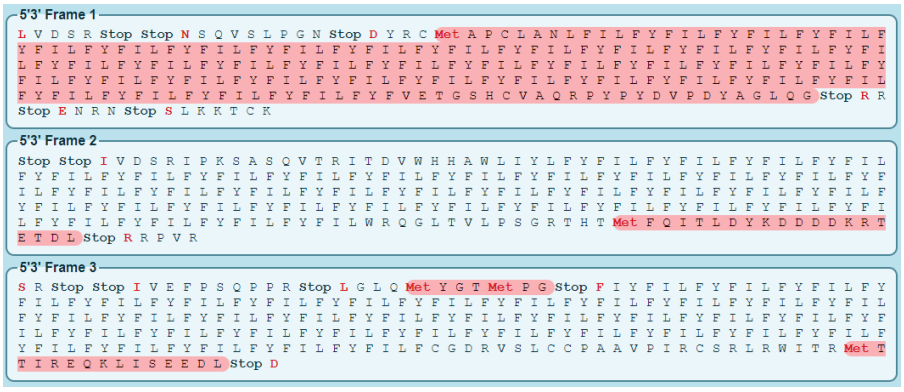
pCDH-6xSTOP-(ATTTT)$_7$-3Tags



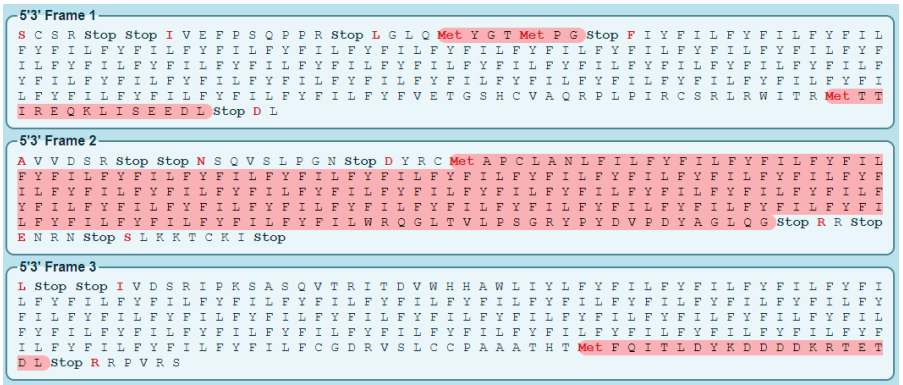
pCDNA3-6xSTOP-(ATTTT)$_7$-3Tags



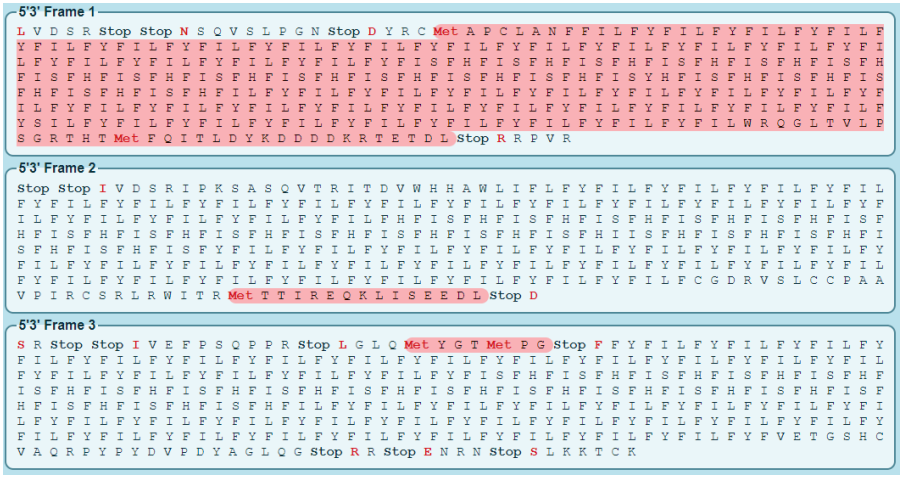
pCDH-6xSTOP-(ATTTT)$_{120}$-3Tags



pCDNA3-6xSTOP-(ATTTT)$_{118}$-3Tags



pCDH-6xSTOP-(ATTTT)$_{57}$(ATTTC)$_{57}$(ATTTT)$_{82}$-3Tags



pCDNA3-6xSTOP-(ATTTT)$_{58}$(ATTTC)$_{55}$(ATTTT)$_{79}$-3Tags

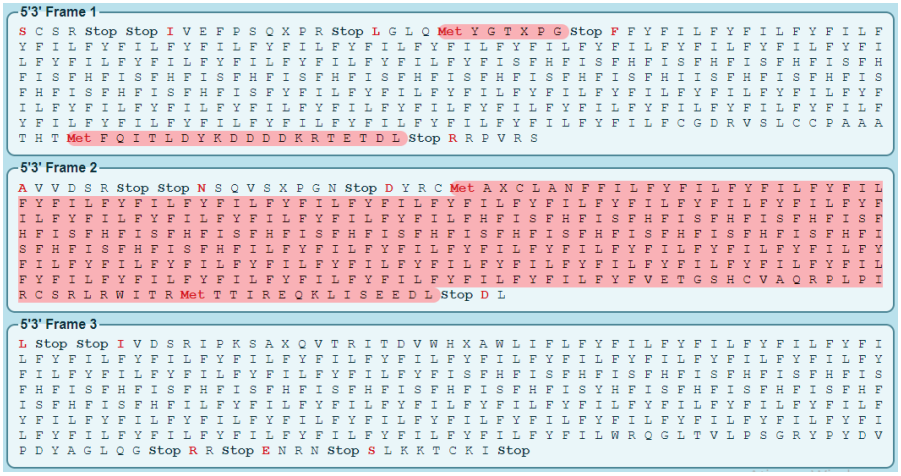


Figure A16. Predicted peptide sequences translated from the 6 generated constructs. Sequences were predicted from the nucleotide sequence of each construct, consisting of 6xSTOP codons, *DAB1* pentanucleotide repeats and 3 Tags, with ExPASy Translate software. YPYDVPDYA = peptide sequence of HA epitope; DYKDDDDK = peptide sequence of flag epitope; EQKLISEEDL = peptide sequence of myc epitope.